Received: 25 July 2016 | Revised: 13 March 2017 | Accepted: 15 May 2017 | Published on: 8 June 2017

DOI: 10.1002/sam.11350

ORIGINAL ARTICLE

WILEY

A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints

Qiwei Li[1] | Michele Guindani[2] | Brian J. Reich[3] | Howard D. Bondell[3] | Marina Vannucci[1]*

[1]Department of Statistics, Rice University, Houston, Texas
[2]Department of Statistics, University of California, Irvine, California
[3]Department of Statistics, North Carolina State University, Raleigh, North Carolina

Correspondence
Marina Vannucci, Department of Statistics, Rice University, 6100 Main Street, Houston, TX 77005.
Email: marina@rice.edu

In this paper, we consider the problem of modeling a matrix of count data, where multiple features are observed as counts over a number of samples. Due to the nature of the data generating mechanism, such data are often characterized by a high number of zeros and overdispersion. In order to take into account the skewness and heterogeneity of the data, some type of normalization and regularization is necessary for conducting inference on the occurrences of features across samples. We propose a zero-inflated Poisson mixture modeling framework that incorporates a model-based normalization through prior distributions with mean constraints, as well as a feature selection mechanism, which allows us to identify a parsimonious set of discriminatory features, and simultaneously cluster the samples into homogenous groups. We show how our approach improves on the accuracy of the clustering with respect to more standard approaches for the analysis of count data, by means of a simulation study and an application to a *bag-of-words* benchmark data set, where the features are represented by the frequencies of occurrence of each word.

KEYWORDS

Bayesian nonparametrics, count data, feature selection, Poisson mixture, text analysis

1 | INTRODUCTION

Modern data science often involves data sets where the features of interest are measured as counts. For example, text documents are typically summarized by their word frequencies, with the size of the dictionary determining the number of features (see eg, [2, 26]. In those applications, clustering and feature selection techniques are often employed to provide low-dimensional summaries of the data and investigate the topical content of a text [1, 3]. Similarly, in ecological survey data (see eg [22]), species counts are observed on a relatively large number of sites, with the objective of characterizing the different ecosystems across the sites. In biology, sequence data—for example, SAGE data, RNA-Seq and microbiome data [8, 17, 40]—are represented as a matrix of short sequence tags (eg taxonomic units in a microbiome experiment) and corresponding observed reads for several samples. Investigators are typically interested in discovering tags whose abundances significantly differ across samples.

One common characteristic of those studies is that the observed frequency of a feature depends on the sampling effort, or—for the analysis of text documents—the document length. This usually results in datasets characterized by 2 distinctive attributes. On the one hand, the datasets contain a high percentage of zero counts. A zero count can either indicate a missing trait in the population or be due to a limited sample. Furthermore, the datasets are highly variable, both with respect to the total number of counts per sample and to the total number of counts per feature. Thus, the observed distributions of counts are typically skewed and overdispersed, since a large number of features are recorded at low frequencies whereas a few features are recorded very frequently. The amount of overdispersion may also vary sample to sample [28, 59].

In order to take into account the skewness and heterogeneity of the data, some type of normalization and regularization is often necessary for conducting inference on the features' occurrence rates [4, 5, 51]. For example, [59] proposes a

Poisson log-linear model, where the Poisson intensities are robustly estimated after a normalization step which takes into account the total number of reads observed for each sample and each feature. More recently, [1] have proposed a Hierarchical Poisson modeling framework where the word occurrence rate is moderated by the document's length, and a low-dimensional representation of the data is achieved by means of a set of latent (mixture) components and sparsity inducing shrinkage priors.

In this paper, we propose a novel regularization approach for estimating occurrence rates in zero-inflated Poisson mixture models. Zero-inflated Poisson models have often been employed to fit count data characterized by overdispersion and a high number of zeros, both in the econometrics and the statistics literature [13, 14, 37]. Our proposal is characterized by priors that employ "soft" constraints on expected values of both samples' and features' scaling parameters, in order to normalize the information content of each sample. This is in contrast to typical approaches that use "hard" constraints as data-dependent plug-in estimates based on the observed counts, see for example [5], [12], [40], [42], and [59], which a priori condition the inference over unknown parameters on the observed counts. Such plug-in estimates can be regarded as maximum likelihood estimators in multiple-stage approaches and somewhat akin to Empirical Bayes methods, therefore relying on implicit assumptions of exchangeability of the observations, which may not be always justified in practice and can introduce bias in the estimation of posterior uncertainties [23, 41]. In addition, "soft" constraints based on a moderate amount of information have been shown to produce more flexible and less biased estimates [29]. Our approach is similar in spirit to the recent proposal by [48], who impose a stochastic constraint on the quantile functions of infinite mixtures, in order to avoid identifiability restrictions in Bayesian nonparametric approaches for quantile regression.

We further regularize the estimation problem by allowing priors that enable feature selection and discriminate samples into clusters. In the analysis of a text document, for example, our approach allows to identify a subset of discriminatory features (words) which are exclusive to particular topics, identified from clusters of documents, automatically balancing the influence of frequency and exclusivity of words across samples by virtue of our model-based inference [1, 50]. More specifically, to achieve feature selection we introduce latent discriminatory variables, similarly as in [56] and [46], who proposed the use of latent indicators for variable selection in the context of finite mixtures of Gaussian distributions. Samples are then clustered on the basis of the similarity of the resulting vectors of discriminatory features by means of an infinite Dirichlet process mixture (DPM). DPMs have been often employed in Bayesian modeling for clustering purposes, since they allow to estimate the number of clusters (ie, the number of mixing components) directly from the data [44]. Alternatively, one could consider a finite mixture model, and use reversible jump Markov chain Monte Carlo (RJMCMC) to determine the number of components, at the expense of increased computational cost [49, 56].

In recent Bayesian nonparametric literature, it has become common to employ an Indian Buffet Process (IBP) characterization to identify nonzero elements in a general matrix of n samples $\times p$ (latent) features [25]. For example, IBP Compound Dirichlet process models [58] and Beta-negative binomial processes [60] have been introduced for the analysis of count data in topic modeling. Differently than those, the constrained Poisson mixtures (CoPoM) approach we propose is aimed at normalizing the data through the use of mean constraints, and clustering the available samples based on the entire subset of selected observed features, rather than assigning single elements of the $n \times p$ matrix to either one of the mixture components (eg a topic) or none.

By means of a simulation study, we show how our approach improves on the accuracy of the clustering performance with respect to more standard approaches for the analysis of count data. We then present an application to a *bag-of-words* benchmark data set, where the features are represented by the frequencies of occurrence of each word.

The rest of the paper is organized as follows. In Section 2 we introduce the zero-inflated hierarchical mixture model and discuss the prior formulations. In Section 3 we briefly describe the MCMC algorithm and discuss the resulting posterior inference. In Section 4 we illustrate the performance of our method on simulated data and then present an application to document clustering, based on a *bag-of-words* benchmark data set. Section 5 concludes the paper with a discussion of the modeling choices, namely the use of a Poisson likelihood vs alternatives, and with future research directions.

2 | ZERO-INFLATED POISSON MIXTURE MODEL WITH FEATURE SELECTION

We consider a $n \times p$ matrix of counts, x_{ij}, $i = 1, \dots, n$, $j = 1, \dots, p$, observed on a set of p features and n samples. We assume that a large number of the counts is zero, either because the feature is truly missing in a subset of the population or due to limitations of the sampling effort. Thus, we start by considering a zero-inflated Poisson mixture model, that is, a mixture model where we constrain one of the kernels to be degenerate at zero,

$$x_{ij} \sim \pi\, \delta_0(x_{ij}) + (1 - \pi) \int \mathrm{Poi}(x_{ij}; \lambda)\, G(d\lambda), \tag{1}$$

where $\pi \in [0, 1]$, $\mathrm{Poi}(x; \lambda)$ denotes a Poisson distribution for the random variable x, with expectation λ, $\delta_c(\cdot)$ indicates a point mass distribution on $c \in \mathbb{R}$, and $G(\cdot)$ denotes a general mixing distribution, which we use to model the overdispersion of the data. Note that if $G(\cdot)$ is Gamma distributed, then Equation 1 defines a zero-inflated Negative Binomial [14]. Alternatively, we can write Equation 1 also by introducing latent indicator variables $r_{ij} \sim \text{Bern}(\pi)$, such that if $r_{ij} = 1$ then $x_{ij} = 0$, whereas if $r_{ij} = 0$ then $x_{ij} \sim \int \text{Poi}(x_{ij}; \lambda)\, G(d\lambda)$.

2.1 | Feature selection

We envision that only some of the features are relevant to discriminate the n samples into distinct clusters. In particular, we postulate the existence of a latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$, with $\gamma_j = 1$ if the counts associated to feature j are relevant to discriminate among the n samples, and $\gamma_j = 0$ otherwise. Let $p_\gamma = \sum_{j=1}^{p} \gamma_j$ denote the number of discriminatory features and, correspondingly, let $p - p_\gamma$ indicate the number of noninformative features. Our model formulation assumes that any count, say x_{ij}, that maps to a discriminatory feature is drawn from a zero-inflated Poisson distribution with intensity parameter λ_{ij}, while counts that map to nondiscriminatory features are drawn from a "null" model, which can be characterized as a zero-inflated Poisson distribution with intensity parameter λ_{ij}^0. In other words, our representation allows for features to be discriminatory, although their count may be zero in some samples. Conditioning on $r_{ij} = 0$, we can write our model as follows,

$$x_{ij} | r_{ij} = 0, \gamma_j, \lambda_{ij} \overset{ind}{\sim} \begin{cases} \text{Poi}(x_{ij}; \lambda_{ij}) & \textit{if}\ \gamma_j = 1 \\ \text{Poi}(x_{ij}; \lambda_{ij}^0) & \textit{if}\ \gamma_j = 0, \end{cases} \tag{2}$$

whereas if $r_{ij} = 1$ we assume $x_{ij} = 0$, irrespectively of γ_j, with $i = 1, \dots, n$ and $j = 1, \dots, p$. A common choice for the prior on the vector $\boldsymbol{\gamma}$ is to assume independent Bernoulli distributions on the individual components, with a common hyperparameter ω, that is, $\gamma_j | \omega \sim \text{Bern}(\omega)$, which is equivalent to a Binomial prior on the number of discriminatory features, that is, $p_\gamma | \omega \sim \text{Bin}(p, \omega)$. The hyperparameter ω can be elicited as the proportion of features expected a priori to be in the discriminatory set. This prior assumption can be further relaxed by formulating a $\text{Be}(a_\omega, b_\omega)$ hyperprior on ω, which leads to a beta-binomial prior for p_γ with expectation $p\, a_\omega / (a_\omega + b_\omega)$. A vague prior on ω can be obtained as in [10] by imposing the constraint $a_\omega + b_\omega = 2$, corresponding to a prior effective sample size of 2, with some desirable mean percentage of inclusion. Analogously, we assume $\pi \sim \text{Be}(a_\pi, b_\pi)$ as a prior on the zero-inflation weight.

In order to account for the variability observed across samples and across features, we further parameterize the intensity parameters of the Poisson distributions as the multiplicative effect of 3 random effects: (1) a scaling factor capturing how the sampling effort affects sample-specific occurrences across all features, denoted by s_i; (2) a scaling factor capturing feature-specific levels across all samples, denoted by g_j (eg a common usage preposition in the analysis of text documents); and (3) a term capturing the occurrence rate for each count, once all the previous global effects have been accounted for, denoted by d_{ij}. Specifically, we parameterize the Poisson intensity in Equation 2 as $\lambda_{ij} = s_i\, g_j\, d_{ij}$, if $\gamma_j = 1$, whereas we set $\lambda_{ij}^0 = s_i g_j d_0$ if $\gamma_j = 0$ in the "null" model. Here, d_0 is a parameter capturing homogenous background noise across samples and features. Since background noise should be the result of the variability captured by the scaling factors s_i and g_j, and also to ensure the identifiability of all model parameters, we assume $d_0 = 1$, so that the significance of each feature is completely revealed by the values of the random effects d_{ij} across the samples.

Multiplicative characterizations of the Poisson intensity parameter of model Equation 2 are typical both in the frequentist as well as in the Bayesian literature to account for latent heterogeneity and overdispersion in count data (eg, seethe latter for examples of this specification in spatial statistics)[1, 7, 14, 59]. To simplify both the prior specification and the computational algorithms, it is sometimes convenient to reparametrize model Equation 2 by using the logarithmic transformations $\tilde{\lambda}_{ij} = \log\{\lambda_{ij}\}$, and, consequently, $\tilde{\mathbf{s}}_i = \log\{s_i\}$, $\tilde{\mathbf{g}}_j = \log\{g_j\}$ and $\tilde{d}_{ij} = \log\{d_{ij}\}$, such that $\tilde{\lambda}_{ij} = \tilde{\mathbf{s}}_i + \tilde{\mathbf{g}}_j + \tilde{d}_{ij}$. Under this reparameterization, the base component is characterized by $\tilde{d}_0 = 0$. In Section 2.2, we describe a regularizing prior specification for the scaling factors that allows flexible modeling of count data, and also avoids some limiting assumptions commonly made when estimating those models.

2.2 | Estimation of scaling factors via mean constraints

In Poisson models with a multiplicative intensity λ_{ij} as defined in Section 2.1, the inferential interest is often limited to the estimation of the occurrence rates d_{ij}, whereas scaling factors are often normalized in order to regularize the inferential problem and ensure the identifiability of the relevant parameters d_{ij}. One typical choice is to estimate the scaling factors by means of plug-in estimators based on the observed counts. For example, in the context of the estimation of RNA-seq abundances, [40], [42], and [59] fix $\hat{s}_i = \sum_{j=1}^{p} x_{ij} / \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}$, so that $\sum_{i=1}^{n} s_i = 1$. Similarly, [5] propose $\hat{s}_i = m_i / \sum_{i=1}^{n} m_i$, where m_i is the median of the distribution of the ratios of the counts for observation i to their geometric mean. As a further example, [12] propose taking $\hat{s}_i = q_i / \sum_{i=1}^{n} q_i$, with q_i the 75th percentile of the counts for observation i. Many of the above examples further fix $\hat{g}_j = \sum_{i=1}^{n} x_{ij}$. While convenient, the use of plug-in estimates for estimating s_i and g_j has noticeable drawbacks, since a priori they condition the inference over unknown parameters on the observed data counts. Indeed, such plug-in estimates can be regarded as maximum likelihood estimators in multiple-stage approaches and somewhat akin to Empirical Bayes methods, therefore relying on implicit assumptions of exchangeability of the observations, which may not be always justified in practice and can introduce bias in the estimation of posterior uncertainties [23,41].

One of the contributions of this article is to provide an alternative normalization approach, through the use of priors for the vectors $\mathbf{s} = (s_1, \dots, s_n)$ and $\mathbf{g} = (g_1, \dots, g_p)$, that respectively capture the sampling and feature heterogeneity in the count data, and avoid fixing those values a priori. A simple choice would be to use conjugate priors, $s_i \sim \mathrm{Ga}(a_s, b_s)$ and $g_j \sim \mathrm{Ga}(a_g, b_g)$. However, the Poisson intensity in Equation 2 depends on the product $s_i\, g_j$, and further constraints are necessary to allow identifiability of the parameters and simultaneous inference on the occurrence rates d_{ij}. In the context of quantile regression, [48] impose a stochastic constraint on the quantile functions of infinite mixtures to ensure identifiability of the parameter estimates. Here, we consider priors on s_i and g_j that impose normalizing constraints on the expected values, but still provide a flexible estimate of the posterior densities.

For computational convenience, we consider the logarithmic transformation $\tilde{s}_i = \log\{s_i\}$ and $\tilde{g}_j = \log\{g_j\}$, so that $s_i g_j = \exp\{\tilde{s}_i + \tilde{g}_j\}$. We assume that the priors for $\tilde{s}_i$ and $\tilde{g}_j$ are mixture distributions,

$$\tilde{s}_i|\cdot \sim \sum_{m=1}^{M} \phi^s_m f^s_m(\tilde{s}_i|\cdot), \qquad \text{and} \qquad \tilde{g}_j|\cdot \sim \sum_{l=1}^{L} \phi^g_l f^g_l(\tilde{g}_j|\cdot), \tag{3}$$

with $0 \le \phi^s_m, \phi^g_l \le 1$, $\sum_{m=1}^{M} \phi^s_m = \sum_{l=1}^{L} \phi^g_l = 1$, and M, L are positive integers. The use of mixture distributions allows flexible estimation of the posterior density of $\tilde{s}_i$ and $\tilde{g}_j$. We further build each mixture so to satisfy the desired constraint, that is, $E[\tilde{\mathbf{s}}_i] = c_s$ and $E[\tilde{\mathbf{g}}_j] = c_g$, where c_s and c_g are some fixed values. [31] demonstrates that any distribution with a mean constraint can be generated by an infinite sum of 2-component mixture distributions, where the 2-component mixtures are constrained to the have the required expected value. Therefore, we assume that each $f^s_m(\cdot)$ and $f^g_l(\cdot)$ in (3) are themselves a 2-component Gaussian mixture, as

$$f^s_m(\tilde{s}_i|t_m, \eta_m) = t_m\, \mathrm{N}(\eta_m, \sigma^2_s) + (1 - t_m)\, \mathrm{N}\left(\frac{c_s - t_m\eta_m}{1 - t_m}, \sigma^2_s\right),$$
$$f^g_l(\tilde{g}_j|q_l, \mu_l) = q_l\, \mathrm{N}(\mu_l, \sigma^2_g) + (1 - q_l)\, \mathrm{N}\left(\frac{c_g - q_l\mu_l}{1 - q_l}, \sigma^2_g\right), \tag{4}$$

with $0 \le t_m, q_l \le 1$. It is immediate to check that the densities in (4) satisfy the desired constraint. Of course, one could consider other prior formulations satisfying that constraint. However, the proposed mixture-of-mixtures prior is attractive because it allows flexibility in the estimation of the unknown s_i and g_j's, by spanning a wide class of distributions, for example, skewed and multimodal densities.

Furthermore, if $M = L = \infty$, then (3) can be interpreted as Bayesian nonparametric infinite mixtures. In particular, DPM models have been extensively used in recent literature for flexible density estimation, both for continuous and discrete data (see, eg, [36,55,57]). The Dirichlet process assumes that the mixing distribution can be written as a discrete random measure, $F(\cdot) = \sum_{k=1}^{\infty} \phi_k\, \delta_{\theta^*_k}(\cdot)$, where the weights ϕ_k are defined by the [54] stick-breaking construction, ie, $\phi_1 = V_1$, $\phi_k = V_k \prod_{u=1}^{k-1}(1 - V_u)$, $V_k \sim \mathrm{Be}(1, \alpha)$, $k = 1, 2, \ldots$, and the atoms $\theta^*_k \sim F_0$, with F_0 a baseline parametric model describing the prior expectation of the Dirichlet process. In symbols, we write $F \sim \mathrm{DP}(\alpha, F_0)$. The concentration (or mass) parameter α provides a measure of the precision of the random measure around the baseline parametric model (see, for details on the Dirichlet process, [30, 44]). We note that theoretical results on large support and consistency of models based on discrete kernels have not been discussed in the literature. Indeed, given the discreteness of the support on the natural numbers, some technical issues make the derivation of such results more complex than for mixtures of continuous distributions.

We conclude this section by specifying the distributions of the hyper-parameters in (4). More specifically, we assume that the 2 mixtures are characterized by $\eta_m \sim \mathrm{N}(0, \tau_\eta)$, $t_m \sim \mathrm{Be}(a_t, b_t)$, and $\mu_l \sim \mathrm{N}(0, \tau_\mu)$, $q_l \sim \mathrm{Be}(a_q, b_q)$, whereas ϕ^s_m and ϕ^g_l are obtained according to the stick-breaking construction. Since the aim of this specification is simply to achieve an automatic normalization of the scaling factors, we further assume $\sigma^2_s = \sigma^2_g = 1$.

2.3 | Clustering selected features via DPMs

In the analysis of count data across multiple samples, one common objective is to characterize and cluster the observed samples into homogenous groups on the basis of the estimated features' occurrence rates. In this section we provide a method for clustering the n samples, based on the set of selected discriminatory features from Section 2.1. We use the superscript (γ) to index the set of discriminatory features, characterized by $\gamma_j = 1$ in (2). Similarly, (γ^c) indicates the set of non-discriminatory features, characterized by $\gamma_j = 0$ and the "null" Poisson distribution with intensity parameter $\lambda^0_{ij} = s_i\, g_j\, d_0$. Thus, each data sample is represented by the $1 \times p$ vector $\mathbf{x}_{i\cdot}$ of observations $\{x_{ij}\}$, with $\mathbf{x}^{(\gamma)}_{i\cdot}$ and $\mathbf{x}^{(\gamma^c)}_{i\cdot}$ indicating the subsets of features corresponding to $\gamma_j = 1$ and $\gamma_j = 0$, respectively. We assume that samples can be clustered based on the subset of selected features $\mathbf{x}^{(\gamma)}_{i\cdot}$, by means of a zero-inflated infinite mixture of Poisson distributed components. For that purpose, we introduce an auxiliary set of clustering allocation variables, $\mathbf{z} = \{z_1, \ldots, z_n\}$, defined so that $z_i = k$ if and only if the vector of observations $\mathbf{x}^{(\gamma)}_{i\cdot}$ belongs to cluster k, for some integer $k \ge 1$. Then, we can characterize the likelihood for the selected features as

$$x^{(\gamma)}_{ij}|z_i = k, r_{ij}, s_i, g_j \overset{ind}{\sim} r_{ij}\, \delta_0(x_{ij}) + (1 - r_{ij})\, \mathrm{Poi}(x_{ij}; \lambda^*_{ijk}) \tag{5}$$

where $\lambda^*_{ijk} = s_i\, g_j\, d^*_{kj}$, with $d^*_{kj} \sim \mathrm{Ga}(a, b)$ being a non-negative occurrence rate for feature j, common to all non-zero observations assigned to the kth cluster. We further assume that each $z_i \sim \sum_{k=1}^{\infty} \mathrm{w}_k\, \delta_{\{k\}}$, where the weights w_k are defined by the [54] stick-breaking construction, so that (5) effectively defines a zero-inflated conjugate DPM model. This modeling framework allows us to cluster the samples based on the vectors of selected features, by allowing $d_{ij} = d^*_{jk}$ when $r_{ij} = 0$, for some integer k, while at the same time the number of clusters is estimated as a by-product of the usual posterior inference. The dimension of the component-specific vector

Mixture model Likelihood:

$$\begin{aligned} x_{ij}|\gamma_j = 1, z_i, s_i, g_j, r_{ij} &\overset{ind}{\sim} r_{ij}\,\delta_0(x_{ij}) + (1-r_{ij})\,\mathrm{Poi}(x_{ij};\lambda^*_{ijz_i}) \quad \text{with } \lambda^*_{ijz_i} = s_i\, g_j\, d^*_{z_i j} \\ x_{ij}|\gamma_j = 0, s_i, g_j, r_{ij} &\overset{ind}{\sim} r_{ij}\,\delta_0(x_{ij}) + (1-r_{ij})\,\mathrm{Poi}(x_{ij};\lambda^0_{ij}) \quad \text{with } \lambda^0_{ij} = s_i\, g_j \end{aligned}$$

Zero-inflation latent indicator prior:

$$r_{ij} \sim \mathrm{Bern}(\pi), \qquad \pi \sim \mathrm{Be}(a_\pi, b_\pi)$$

Feature selection prior:

$$\gamma_j \sim \mathrm{Bern}(\omega), \qquad \omega \sim \mathrm{Be}(a_\omega, b_\omega)$$

Mixing distribution for selected discriminatory features:

$$\begin{aligned} z_i &\sim \sum_{k=1}^{\infty} \mathrm{w}_k\, \delta_{\{k\}} \\ \boldsymbol{d}_k^{*(\gamma)} &\sim \prod_{\{j:\gamma_j=1\}} \mathrm{Ga}(a,b) \\ \mathrm{w}_k &= V_k \prod_{u=1}^{k-1}(1-V_u),\ V_k \sim \mathrm{Be}(1,\alpha) \\ \alpha &\sim \mathrm{Ga}(a_\alpha, b_\alpha) \end{aligned}$$

Priors on subject- and feature-specific scaling factors:

$$\begin{aligned} \tilde{s}_i|\nu_i, \epsilon_i, \boldsymbol{t}, \boldsymbol{\eta} &\sim \sum_{m=1}^{M} \phi^s_m \left[t_m \mathrm{N}(\eta_m, \sigma_s^2) + (1-t_m)\mathrm{N}\left(\frac{c_s - t_m\eta_m}{1-t_m}, \sigma_s^2\right)\right] \\ \phi^s_m &= V_m \prod_{u=1}^{m-1}(1-V_u),\ V_m \sim \mathrm{Be}(a_m, b_m) \\ \eta_m &\sim \mathrm{N}(0, \tau_\eta) \\ t_m &\sim \mathrm{Be}(a_t, b_t) \\ \tilde{g}_j|\xi_j, \psi_j, \boldsymbol{q}, \boldsymbol{\mu} &\sim \sum_{l=1}^{L} \phi^g_l \left[q_l \mathrm{N}(\mu_l, \sigma_g^2) + (1-q_l)\mathrm{N}\left(\frac{c_g - q_l\mu_l}{1-q_l}, \sigma_g^2\right)\right] \\ \phi^g_l &= V_l \prod_{u=1}^{l-1}(1-V_u),\ V_l \sim \mathrm{Be}(a_l, b_l) \\ \mu_l &\sim \mathrm{N}(0, \tau_\mu) \\ q_l &\sim \mathrm{Be}(a_q, b_q) \end{aligned}$$

Fixed hyperparameters:

$$a, b, a_\alpha, b_\alpha, a_\omega, b_\omega, a_\pi, b_\pi, c_s, c_g, M, L, \sigma_s, \sigma_g, a_m, b_m, a_l, b_l, \tau_\eta, \tau_\mu, a_t, b_t, a_q, b_q$$

FIGURE 1 Hierarchical formulation of the proposed constrained Poisson mixture (CoPoM) model

$\mathbf{d}_k^{*(\gamma)} = \{d^*_{k1}, \dots, d^*_{kp_\gamma}\}$ depends on the outcome of the feature selection procedure. As a matter of fact, the joint prior probability of a given allocation of the n samples into K groups is given by

$$p(z_1, \dots, z_n) = \frac{\alpha^K\, \Gamma(\alpha) \prod_{k=1}^{K} \Gamma(n_k)}{\Gamma(\alpha + n)}, \tag{6}$$

that describes the Ewens distribution [6,18,21]. Here, α is the concentration parameter of the DPM model, which defines the expected number of clusters as $E(k) \approx \alpha \log(n)$ [6]. As $\alpha \to 0$, the number of clusters goes to 1, while for $\alpha \to \infty$ the number of clusters goes to n. We complete our prior specification by placing a $\mathrm{Ga}(a_\alpha, b_\alpha)$ hyperprior on α as in [20]. In the absence of prior information, we suggest choosing the values of a_α and b_α to obtain a fair degree of support for $\alpha \approx 1$.

The proposed model is summarized in Figure 1. Our zero-inflated Poisson mixture model provides a more flexible framework for density estimation of overdispersed count data with respect to widely used Negative Binomial models, which are indeed a special case of our framework.

3 | MODEL FITTING

We now briefly describe the MCMC algorithm for posterior inference. Our inferential strategy allows to simultaneously infer group structure in the samples while identifying the discriminatory features.

3.1 | MCMC algorithm

Our primary interest lies in the identification of the discriminatory features, via the vector $\boldsymbol{\gamma}$, and the estimation of the sample clustering allocations, via the vector $\mathbf{z}$. For this, we design a Markov chain Monte Carlo (MCMC) algorithm based on Metropolis search variable selection algorithms

[11, 24] and Gibbs sampling methods for DPM models [45]. We also sample the sample-specific and feature-specific scaling factors, **s** and **g**. We give full details of our MCMC algorithm in the Appendix A and report here a brief description of the most relevant updates.

Update of $\boldsymbol{\gamma}$: This is done via an *add-delete-swap* algorithm. In this approach, a new candidate vector, say $\boldsymbol{\gamma}^{new}$, is generated by randomly choosing between 2 types of moves. For the add/delete move, we select at random one of the elements in the current vector, say $\boldsymbol{\gamma}^{(old)}$, and change its value from 0 to 1, or vice versa. For the swap move, we select 2 elements in $\boldsymbol{\gamma}^{old}$ with different inclusion status and swap their values. Then, the Metropolis-Hastings ratio can be written as

$$\mathrm{m_{MH}} = \frac{p(\boldsymbol{\gamma}^{\mathrm{new}}|\mathbf{z}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X})}{p(\boldsymbol{\gamma}^{old}|\mathbf{z}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X})} \frac{J(\boldsymbol{\gamma}^{\mathrm{old}}|\boldsymbol{\gamma}^{\mathrm{new}})}{J(\boldsymbol{\gamma}^{\mathrm{new}}|\boldsymbol{\gamma}^{\mathrm{old}})},$$

where $J(\cdot|\cdot)$ indicates the proposal probability distribution for the selected move. The move is accepted with probability $\min(1, \mathrm{m_{MH}})$. We should notice that the feature selection and the cluster structure are determined simultaneously in the MCMC algorithm. Therefore, to improve mixing, it is necessary to allow the selection to stabilize for any visited cluster configuration. As suggested in [34], we repeat the above Metropolis step $E = 20$ times within each iteration. In the applications of this paper, no improvement in the MCMC performance was noticed beyond this value.

Update of z: Since we have assumed a conjugate baseline parametric distribution, $\prod_{\{j:\gamma_j=1\}} \mathrm{Ga}(a, b)$, we can integrate analytically over the cluster-specific parameters $\mathbf{d}_k^{*(\gamma)}$ and directly sample the cluster assignment indicators of the selected features, **z**, according to Algorithm 3 of [45]. More specifically, the Gibbs sampler iteratively samples the full conditionals,

$$\begin{aligned} &p(z_i|\mathbf{z}_{-i}, \boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X}) \\ &\quad = \begin{cases} \frac{n_{k,-i}}{n-1+\alpha} f(\mathbf{x}_{i\cdot}|z_i = k, \mathbf{z}_{-i}, \boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X}_{-i\cdot}) \\ \qquad\qquad \text{for } z_i = k, k = 1, \dots, K_{-i}, \\ \frac{\alpha}{n-1+\alpha} f(\mathbf{x}_{i\cdot}|\boldsymbol{\gamma}, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot}) \quad \text{for } z_i = K_{-i} + 1, \end{cases} \end{aligned} \tag{7}$$

where $\mathbf{z}_{-i}$ denotes all the elements in **z** excluding the *i*th one, $n_{k,-i}$ is the size of cluster k in $\mathbf{z}_{-i}$, and K_{-i} is the number of unique values in $\mathbf{z}_{-i}$. Note that $f(\mathbf{x}_{i\cdot}|z_i = k, \mathbf{z}_{-i}, \boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X}_{-i\cdot})$ is the integrated likelihood, with updated $\mathbf{d}_k^{*(\gamma)}$ based on its prior and all observations except the *i*th one. See details in the Appendix A.

Update of s and g: The prior distribution for the scaling factors **s** and **g** is a DPM on the log-transformed values $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{g}}$. Since the mixture of log-normal distribution that we have assumed as baseline measure in the Dirichlet process is not conjugate to the Poisson likelihood, we consider a finite truncation of the Dirichlet process, so that M and L in Equation 3 are large but finite [33]. Such a choice allows to simplify computations considerably, and still achieves flexible estimate of posterior densities, such as those commonly obtained in a Bayesian nonparametric framework. We employ Metropolis-Hastings steps for all $\tilde{\mathbf{s}}_i$'s and $\tilde{\mathbf{g}}_j$'s.

Update of R: The full conditional for the zero-inflation latent indicators takes into account that we only need to update those r_{ij}'s that correspond to zero counts. For positive counts, necessarily $r_{ij} = 0$. We use a Metropolis-Hasting within Gibbs sampling step for updating r_{ij} and π, after sampling d_{kj}^* for those features corresponding to $\gamma_j = 0$. Details are given in the Appendix A.

3.2 | Posterior inference

We obtain posterior inference on the parameters by postprocessing of the MCMC samples after burn-in. We start by obtaining a probabilistic assessment of the cluster allocations by analyzing the MCMC samples of **z**. One way to summarize the posterior distribution of **z** is via the *maximum-a-posteriori* (MAP) estimate that can be calculated as

$$\begin{aligned} \hat{\mathbf{z}}_{\mathrm{MAP}} &= \underset{1 \le b \le B}{\operatorname{argmax}}\, p(\mathbf{z}^{(b)}|\boldsymbol{\gamma}^{(b)}, \tilde{\mathbf{s}}^{(b)}, \tilde{\mathbf{g}}^{(b)}, \mathbf{R}^{(b)}, \mathbf{X}) \\ &= \underset{1 \le b \le B}{\operatorname{argmax}}\, p(\mathbf{z}) \prod_{i=1}^{n} f \\ &\quad \times (\mathbf{x}_{i\cdot}|z_i^{(b)}, \mathbf{z}_{-i}^{(b)}, \boldsymbol{\gamma}^{(b)}, \tilde{\mathbf{s}}^{(b)}, \tilde{\mathbf{g}}^{(b)}, \mathbf{R}^{(b)}, \mathbf{X}_{-i\cdot}), \end{aligned}$$

with $b = 1, \dots, B$ indicating the MCMC iterations, after burn-in, and where the marginal posterior probability that sample i is allocated to cluster k can be calculated through Equation 7. An alternative estimate relies on the computation of a matrix of posterior pairwise probabilities of co-clustering, that is, the probabilities that observation i and observation i' are assigned to the same cluster, $p_{ii'} = p(z_i = z_{i'}|\boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X})$, as suggested by [19], among others. These probabilities can be estimated by computing empirical frequencies of co-clustering based on the MCMC samples, resulting in an $n \times n$ symmetric pairwise probability matrix (PPM). Then, a point estimate for the cluster memberships, $\hat{\mathbf{z}}_{\mathrm{PPM}}$, is obtained by minimizing the sum of squared deviations of its association matrix from the PPM, that is,

$$\hat{\mathbf{z}}_{\mathrm{PPM}} = \underset{\mathbf{z}}{\operatorname{argmin}} \sum_{i<i'} \left[I(z_i = z_{i'}) - p_{ii'}\right]^2.$$

As for feature selection, the MAP estimate of $\boldsymbol{\gamma}$ can be obtained by enumerating all visited MCMC samples $\boldsymbol{\gamma}^{(b)}$ and then considering the set of features that maximizes the posterior density. Alternatively, we can estimate the marginal posterior probability of inclusion (PPI) of single features as the proportion of MCMC iterations, after burn-in, in which the corresponding γ_j were equal to 1, that is $\mathrm{PPI}(j) = \sum_{u=1}^{B} (\gamma_j^{(b)}|\mathbf{z}^{(b)}, \tilde{\mathbf{s}}^{(b)}, \tilde{\mathbf{g}}_j^{(b)}, \mathbf{r}_{\cdot j}^{(b)}, \mathbf{x}_{\cdot j})/B$. A point estimate of $\boldsymbol{\gamma}$ is then obtained by identifying those PPI values that exceed a given threshold. The optimal threshold is typically chosen based on a decision theoretic criterion, for example, to maximize power under a constraint on the number of false positives [27, 43].

4 | APPLICATIONS

We first explore performances on simulated data and then show results on a *bag-of-words* benchmark data set. We also demonstrate the superiority of our CoPoM model over other widely adopted methods for the analysis of overdispersed count data.

4.1 | Simulation study

Data were generated with $n = 30$ samples and $K = 3$ clusters. We then simulated observations from a zero-inflated mixture of Poisson distributions, assuming 3 mixture components and $p_\gamma = 50$ discriminating features,

$$\begin{aligned} x_{ij} \sim\ & 0.25\,\delta_0(x_{ij}) + 0.75 \Big\{ I(1 \le i \le 6)\,\mathrm{Poi}(s_i g_j d^*_{1j}) \\ & + I(7 \le i \le 21)\,\mathrm{Poi}(s_i g_j d^*_{2j}) \\ & + I(22 \le i \le 30)\,\mathrm{Poi}(s_i g_j d^*_{3j}) \Big\}, \end{aligned}$$

where the first 6 observations were drawn from the first distribution, the next 15 from the second and the last 9 from the third distribution. We added 950 noisy features, which we generated from a zero-inflated Poisson model, $x_{ij} \sim 0.25\,\delta_0(x_{ij}) + 0.75\,\mathrm{Poi}(s_i g_j)$. We simulated the s_i's as independent and identically distributed $s_i \sim \mathrm{U}(0.5, 1.5)$ and the g_j's as $g_j \sim \mathrm{Exp}(1/3)$. For the mixture components, we simulated $\tilde{d}^*_{kj}, k = 1, 2, 3$ from a standard Normal distribution $\mathrm{N}(0, \sigma^2)$ with $\sigma^2 = 1$.

As for hyperparameter settings, we used the following default settings. We set the hyperparameters that control the base distribution of the mixture DP prior to $a = b = 1$, which leads to a Gamma distribution with mean and variance equal to 1, and $a_\alpha = b_\alpha = 1$, setting the expectation and variance of the concentration parameter α to 1. As for the Beta prior on the feature selection parameter ω, we set $a_\omega = 0.2, b_\omega = 1.8$, resulting in the proportion of features expected a-priori to discriminate the different groups to be $a_\omega/(a_\omega + b_\omega) = 10\%$. For the priors on $\tilde{\mathbf{s}}_i$'s and $\tilde{\mathbf{g}}_j$'s, we use the following default settings: $M = n/2 = 15, L = p/2 = 500$, $c_s = c_g = 0$, $\sigma_s = \sigma_g = 1$, $\tau_\eta = \tau_\mu = 1$, $a_m = b_m = 1$, $a_l = b_l = 1$, $a_t = b_t = 1$, and $a_q = b_q = 1$. Results we report below were obtained by running one MCMC chain with 10 000 iteration, discarding the first 1000 as burn-in. We started the chain from a model with 2 randomly chosen γ_j's set to 1 and with each observation assigned to a different cluster.

We first describe posterior inference on the relevant parameters as a result of our normalization and regularization approach. Figure 2 shows the heatmap of the pairwise posterior probabilities, $p(z_i = z_{i'} | \boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X})$, of allocating observations i and i' to the same cluster, after burn-in. It is evident from the map that the inspection of the highest posterior allocation probabilities allows to reconstruct the true allocation structure quite well.

As for the feature selection, Figure 3 shows the marginal PPI of each feature $p(\gamma_j = 1 | \mathbf{z}, \tilde{\mathbf{s}}, g_j, \mathbf{r}_{\cdot j}, \mathbf{x}_{\cdot j})$, after burn-in. The red dots indicate the truly discriminatory features. A threshold of 0.5 on the marginal probabilities results in a median model that includes 18 features, all of which are included in the set of discriminatory features used in the data generation process. Figure 4 shows the estimated marginal posterior distribution of the parameter d_{ij}, for 2 of the discriminating features. These were obtained by post-MCMC compositional sampling of the parameters as in [28]. Notice that the distributions are skewed, with high probability mass on low counts and small mass on extreme values, confirming that our modeling approach is able to capture the overdispersion characterizing the simulated data.

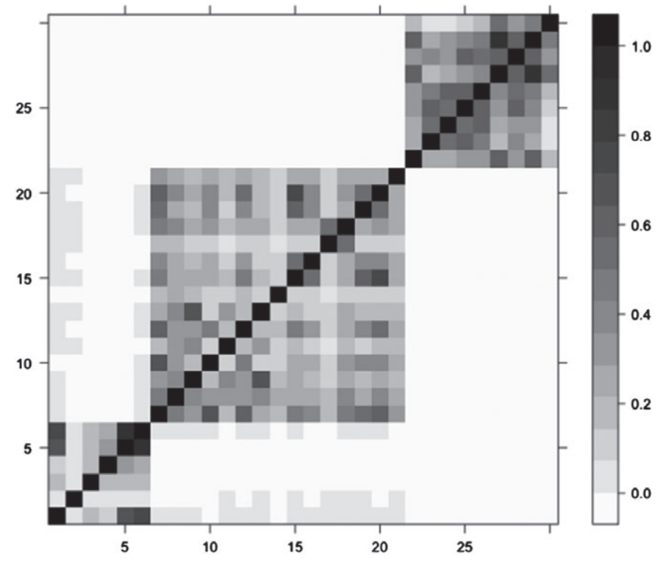

FIGURE 2 Simulation study: Heatmap of the pairwise posterior probabilities $p(z_i = z_{i'} | \cdot)$

In order to further quantify the accuracy of our algorithm, and to compare its performances with other methods available in the literature, we looked at results under different scenarios in terms of sample size, n, number of noisy features, $p{-}p_\gamma$, and cluster dispersion, σ. In particular, we considered $\sigma^2 = 1, 0.5$, $n = 21, 99$ (equally distributed among the $K = 3$ mixture components) and $p = 200, 1000$ (with $p_\gamma = 50$ discriminating features). For each of these $2 \times 2 \times 2 = 8$ scenarios, we simulated 50 data sets and ran our MCMC algorithm with the default settings describe above and ran our MCMC algorithm with the default settings described above.

For the analysis of the clustering results, we quantified performance via the (ARI; [32]), a variant of the Rand index [47], on the basis of the vector of point estimates $\hat{\mathbf{z}}_{\mathrm{PPM}} = (\hat{z}^{\mathrm{PPM}}_1, \ldots, \hat{z}^{\mathrm{PPM}}_n)$. Let $A = \sum_{i>i'} I(z_i = z_{i'}) I(\hat{z}^{\mathrm{PPM}}_i = \hat{z}^{\mathrm{PPM}}_{i'})$ be the number of pairs of observations that belong to the same group in both the true clustering and the estimated one; $B = \sum_{i>i'} I(z_i = z_{i'}) I(\hat{z}^{\mathrm{PPM}}_i \neq \hat{z}^{\mathrm{PPM}}_{i'})$ be the number of pairs which belong to the same group in the true clustering but different groups in the estimated one; $C = \sum_{i>i'} I(z_i \neq z_{i'}) I(\hat{z}^{\mathrm{PPM}}_i = \hat{z}^{\mathrm{PPM}}_{i'})$ be the number of pairs in different groups in the true partition but assigned to the same group in the estimated one; $D = \sum_{i>i'} I(z_i \neq z_{i'}) I(\hat{z}^{\mathrm{PPM}}_i \neq \hat{z}^{\mathrm{PPM}}_{i'})$, the number of pairs

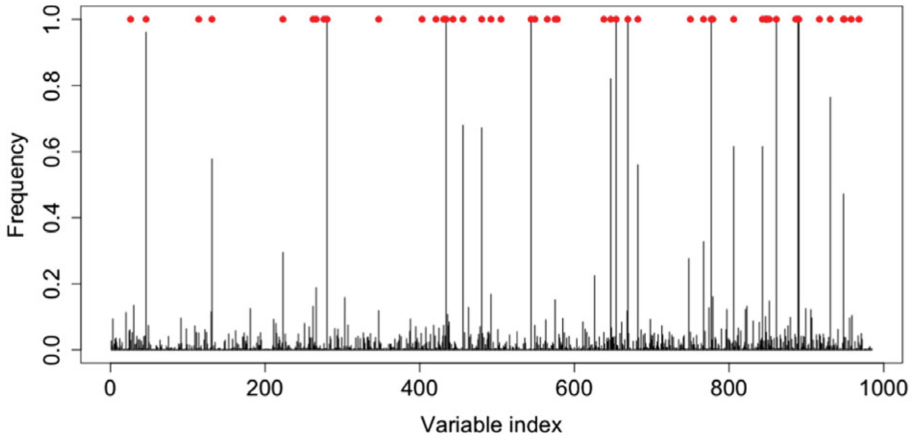

FIGURE 3 Simulation study: Marginal posterior probabilities of inclusion $p(\gamma_j = 1|\cdot)$, with the red dots indicating the truly discriminatory features

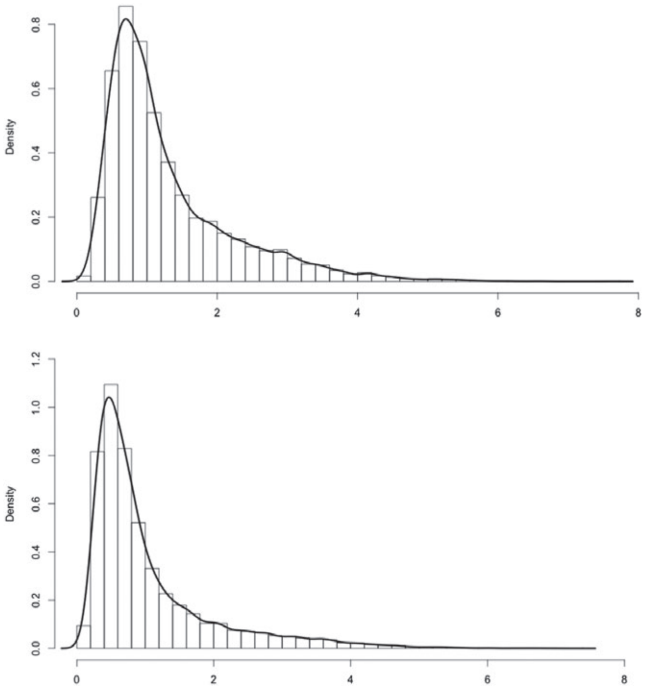

FIGURE 4 Simulation study: Estimated marginal posterior distribution of $E[d_{ij}]$ for 2 of the discriminatory features, obtained by post-MCMC compositional sampling. The distributions are skewed, with high probability mass on low counts and small mass on extreme values, capturing the overdispersion of the simulated data

assigned to different groups in both the truth and the estimate. Then the Rand index (RI) is defined as

$$RI = \frac{A+D}{A+B+C+D},$$

and the ARI as

$$ARI = \frac{\binom{n}{2}(A+D) - [(A+B)(A+C)+(C+D)(B+D)]}{\binom{n}{2}^2 - [(A+B)(A+C)+(C+D)(B+D)]}.$$

The RI yields values between 0 and 1, while the ARI can yield negative values [53]. The larger the index, the more accurate the clustering result.

For comparison, we selected 2 commonly employed estimation methods for count data, based on Poisson mixtures and implemented in the freely available R packages edgeR [52] and PoiClaClu [59]. Both edgeR and PoiClaClu incorporate plug-in estimates of scaling factors for normalization purposes: edgeR considers a negative binomial likelihood, whereas PoiClaClu aims at clustering count data by using a regularized Poisson log-linear model. Unlike our modeling approach, which allows to directly estimate the cluster assignments, through the latent $\hat{\mathbf{z}}$, those methods do not provide individual allocation estimates, but rather yield a dissimilarity matrix that can be transformed into a tree via hierarchical clustering. In order to make the comparison with our CoPoM model feasible, we considered those estimates that achieved the maximum ARI values. Table 1 reports results on clustering performances of all methods in terms of ARI values, averaged over the 50 replicates, under the different simulated scenarios. For our method, we report the results obtained by using the PPM estimates for cluster allocation. The MAP estimates performed similarly (not shown). In all replicates and scenarios considered, we generated data from a zero-inflated Poisson model, fixing $\pi = 0.25$. The percentage of observed zeros was around 44%. Results show that CoPoM consistently outperforms competing methods in terms of clustering accuracy. This is to be expected, since the competing methods do not incorporate the variable selection, and one might expect the inclusion of noisy features to mask the recovery of the true clustering structure. Table 2 shows the performances of our algorithm when data were generated from a Poisson mixture with 3 components (ie, $\pi = 0$). Also in this setting, our zero-inflated Poisson model performs favorably or similarly with respect to the other methods in all cases. The widely used edgeR method shows the worst performances, especially for weakening signal strength, that is, decreasing σ or increasing p, whereas our CoPoM method and PoiClaClu show the best performances.

Since our CoPoM model also allows the selection of a subset of discriminatory features, we quantified its performances in terms of averaged false positive rate (FPR) and true positive

TABLE 1 Simulation study with data from a zero-inflated Poisson mixture: Adjusted Rand index (ARI) values, averaged over 50 replicates, achieved by the R packages edgeR, PoiClaClu and by our CoPoM method under different simulated scenarios. Standard deviations are indicated in parentheses

Scenario			Competing methods		
n	*p*	σ^2	edgeR	PoiClaClu	CoPoM (PPM)
99	200	1.00	0.412 (0.129)	0.550 (0.162)	0.894 (0.129)
99	200	0.50	0.257 (0.100)	0.227 (0.095)	0.895 (0.116)
99	1000	1.00	0.307 (0.125)	0.179 (0.107)	0.925 (0.122)
99	1000	0.50	0.127 (0.103)	0.043 (0.029)	0.907 (0.106)
21	200	1.00	0.492 (0.194)	0.516 (0.210)	0.788 (0.120)
21	200	0.50	0.287 (0.155)	0.280 (0.169)	0.744 (0.128)
21	1000	1.00	0.240 (0.131)	0.134 (0.107)	0.716 (0.081)
21	1000	0.50	0.135 (0.095)	0.101 (0.068)	0.670 (0.130)

TABLE 2 Simulation study with data from a Poisson mixture: Adjusted Rand index (ARI) values, averaged over 50 replicates, achieved by the R packages edgeR, PoiClaClu and by our CoPoM method under different simulated scenarios. Standard deviations are indicated in parentheses

Scenario			Competing methods		
n	*p*	σ^2	edgeR	PoiClaClu	CoPoM (PPM)
99	200	1.00	0.590 (0.133)	1.000 (0.000)	1.000 (0.000)
99	200	0.50	0.437 (0.106)	0.998 (0.007)	1.000 (0.000)
99	1000	1.00	0.394 (0.111)	0.999 (0.006)	1.000 (0.000)
99	1000	0.50	0.232 (0.082)	0.947 (0.067)	0.991 (0.061)
21	200	1.00	0.701 (0.167)	1.000 (0.000)	1.000 (0.000)
21	200	0.50	0.556 (0.138)	0.983 (0.059)	0.997 (0.021)
21	1000	1.00	0.477 (0.143)	0.987 (0.058)	1.000 (0.000)
21	1000	0.50	0.312 (0.150)	0.863 (0.190)	1.000 (0.000)

TABLE 3 Simulation study: Average false-positive rates (FPRs), true positive rates (TPRs), and areas under ROC curves (AUCs), achieved by the CoPoM model under different simulated scenarios, for different values of the threshold, τ, on the PPIs

			CoPoM (PPI)								
Scenario			$\tau = 0.2$		$\tau = 0.4$		$\tau = 0.6$		$\tau = 0.8$		
n	*p*	σ^2	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	AUC
99	200	1.00	0.075	0.888	0.019	0.827	0.005	0.790	0.001	0.743	0.964
99	200	0.50	0.057	0.803	0.014	0.724	0.004	0.667	0.001	0.603	0.936
99	1000	1.00	0.008	0.803	0.002	0.760	0.001	0.724	0.000	0.667	0.961
99	1000	0.50	0.006	0.677	0.001	0.625	0.000	0.580	0.000	0.514	0.929
21	200	1.00	0.354	0.872	0.110	0.700	0.033	0.576	0.008	0.474	0.881
21	200	0.50	0.269	0.706	0.073	0.505	0.020	0.389	0.005	0.305	0.806
21	1000	1.00	0.054	0.579	0.014	0.476	0.004	0.414	0.001	0.350	0.855
21	1000	0.50	0.044	0.431	0.011	0.327	0.003	0.266	0.001	0.216	0.793

rate (TPR) achieved under the different simulated scenarios, for different values of the threshold on the PPIs. Results are shown in Table 3, and the corresponding receiver operating characteristic (ROC) curves are shown in Figure 5. Each subfigure, corresponding to different values of n and p, shows that the estimate becomes more accurate with increasing separation between the clusters, captured by the between-cluster variability parameter σ.

Unlike the edgeR and PoiClaClu methods, which use plug-in estimates, our model formulation is characterized by priors on the samples' and features' scaling parameters that impose soft constraints on the expected values. To appreciate the effect of such prior formulation, we looked at the frequentist coverages of our estimates and compared those to an implementation of our model that employs simple conjugate Gamma priors on the s_i's and g_j's parameters. For example, for the scenario with $n = 21, p = 200, \sigma^2 = 1$, the average frequentist coverages of the 95% credible intervals obtained with our model with constrained priors, calculated over 100 simulated datasets, were .91, .92, .92, for the s_i's, g_j's and $s_i g_j$'s parameters, respectively, while those obtained with unconstrained Gamma priors were 0, .001, .8, for the s_i's, g_j's and $s_i g_j$'s parameters, respectively.

We conclude this section by conducting a sensitivity analysis on the prior specification of our method and reporting results for different hyperparameter values of the priors for

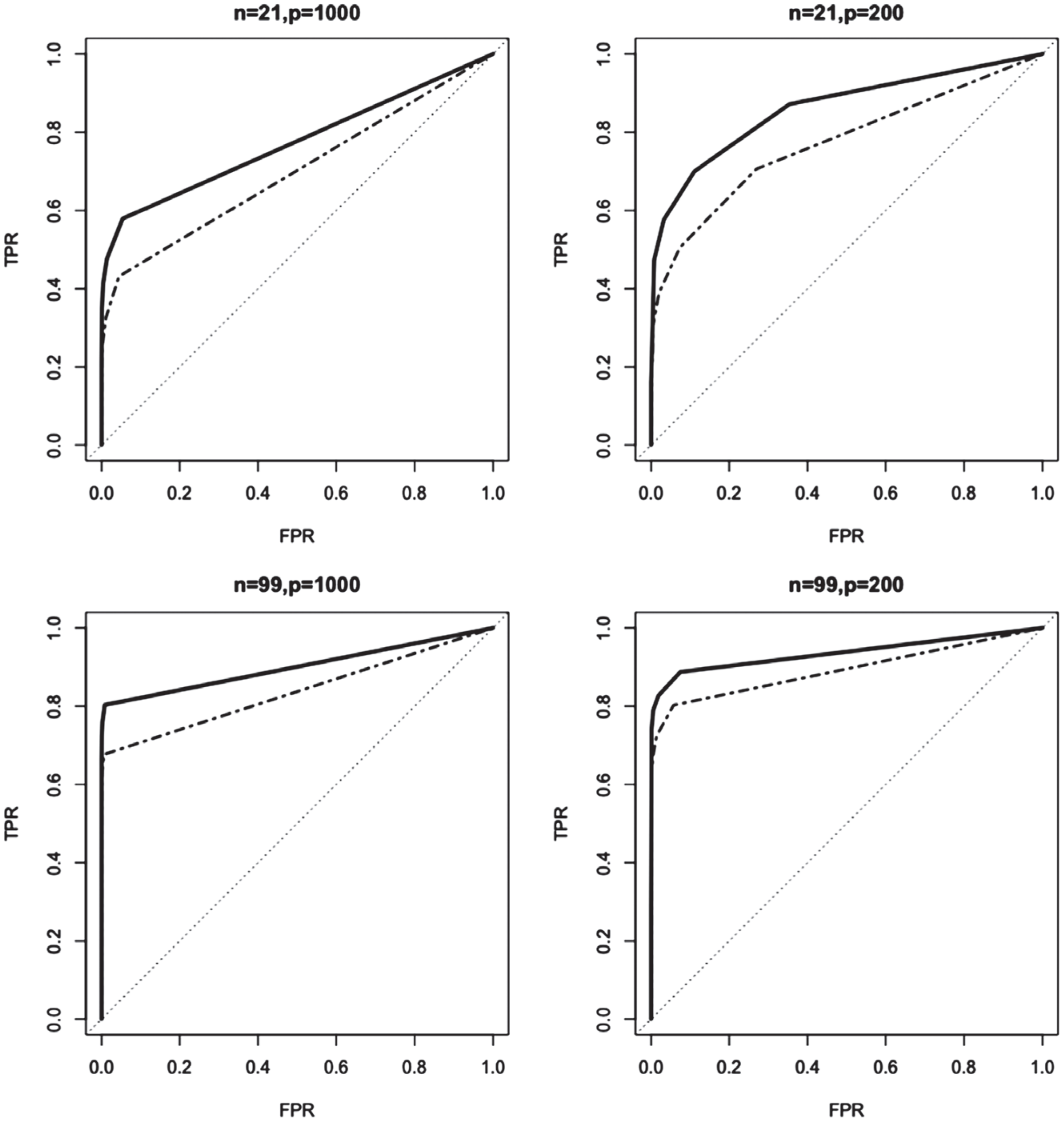

FIGURE 5 Simulation study: Receiver operating characteristic (ROC) curves, for different values of the threshold on the PPIs, obtained by CoPoM under different simulated scenarios. The bold curve corresponds to $\sigma^2 = 1$, whereas the dashed curve corresponds to $\sigma^2 = 0.5$

ω and α. In particular, we considered 4 scenarios: $(a_\omega = 0.02, b_\omega = 1.98)$, $(a_\omega = 1, b_\omega = 1)$, $(a_\alpha = 1, b_\alpha = 10)$, and $(a_\alpha = 1, b_\alpha = 0.1)$, while setting all the other hyperparameter to default values. Table 4 shows that clustering and feature selection performance, in terms of the ARI and the AUC, are fairly robust to the choice of the hyperparameters. When no prior information is available, we suggest to use the default choice $a_\omega = 2p_{\text{prior}}/p$, $b_\omega = 2 - a_\omega$, and $a_\alpha = b_\alpha = 1$.

TABLE 4 Simulation study: Sensitivity analysis on the values of the hyperparameters of the priors on α and π

Scenario				Performance	
a_ω	b_ω	a_α	b_α	ARI	AUC
0.02	1.98	1.00	1.00	0.498	0.879
0.20	1.80	1.00	0.10	0.384	0.864
0.20	1.80	1.00	1.00	0.562	0.846
0.20	1.80	1.00	10.0	0.370	0.833
1.00	1.00	1.00	1.00	0.452	0.869

4.2 | Application to the analysis of *bag-of-words* data

Bag-of-words data sets report the frequencies of occurrence of each word in a text document. Clustering of documents on the basis of a subset of relevant words is one of the most common tasks when analyzing *bag-of-words* data. Here we illustrate the performance of the CoPoM approach using the widely

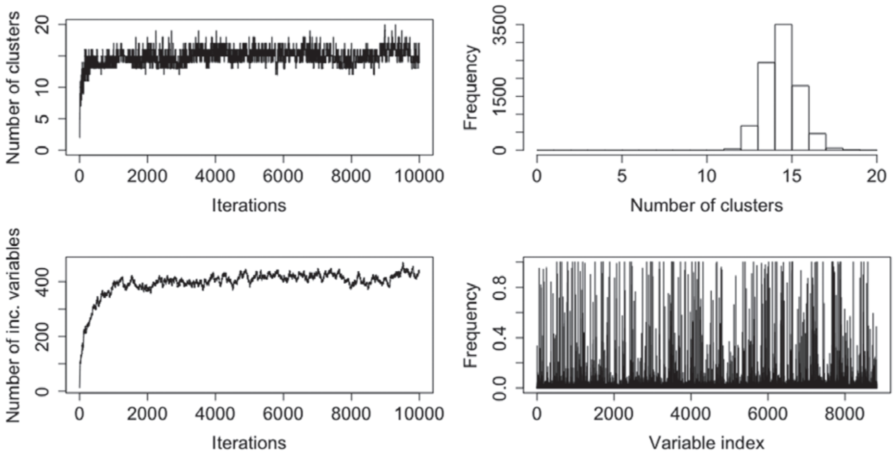

FIGURE 6 Brown Corpus data set: (A) trace plot of the number of clusters K; (B) histogram of the number of clusters K; (C) trace plot of the number of selected variables p_γ; (D) marginal posterior probabilities of inclusion $p(\gamma_j = 1|\cdot)$, with the red dots indicating the truly discriminatory features

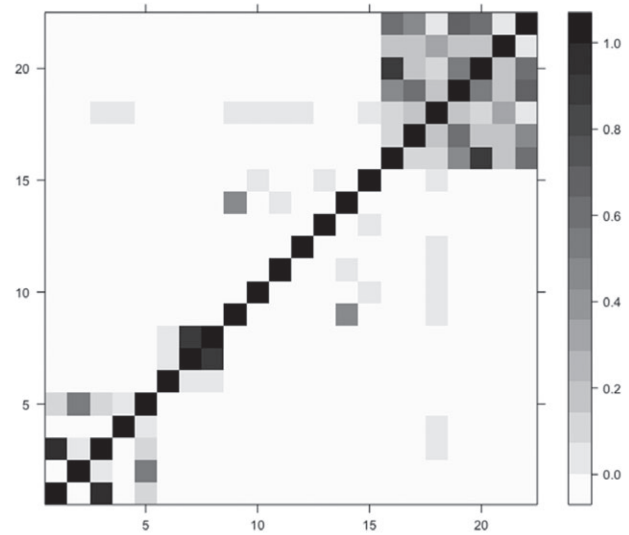

FIGURE 7 Brown Corpus data set: Heatmap of the pairwise posterior probabilities $p(z_i = z_{i'}|\cdot)$

employed *Brown Corpus* [35]. In linguistics, a corpus defines a large and structured collection of texts, which are often used for conducting statistical analyses and tests of hypotheses about a linguistic variety or particular characteristics of a language. The Brown Corpus consists of 500 samples, distributed across 15 genres. Each sample begins at a random sentence-boundary in the article and continues up to the first sentence boundary after 2000 words. The total vocabulary is about 50 000 words and half of them occur equal or less than once in the corpus. Thus, the data are typically characterized by an excessive number of zeros and overdispersion.

For the application of this paper, we selected a subset of the Brown Corpus, composed of 5 sports reportage, 3 society reportage, 7 spot news, and 7 editorials, whose length range from 2200 to 2374 words, for a total of 22 texts and 8826 features. The data are quite sparse, as about 90% of the counts are zero.

We report results obtained by running the MCMC algorithm described in Section 3.1 with the same default settings used in the simulations. Figures 6A and B shows the posterior inference on the number of clusters, more specifically the MCMC trace plot across iterations (A) and the resulting estimate of the posterior distribution of K, after burn-in (B). Figures 6C and D reports the results of the feature selection. More specifically, Figure 6C reports the trace plot for the total number of included features, whereas Figure 6D shows the estimated marginal PPIs of each single feature, $p(\gamma_j = 1|\mathbf{z}, \tilde{\mathbf{s}}, g_j, \mathbf{r}_{\cdot j}, \mathbf{x}_{\cdot j})$, after burn-in. A threshold of 0.5 on the marginal probabilities results in a median model that includes 203 features (2.3% of the total). Finally, Figure 7 shows the heatmap of the pairwise posterior probabilities of co-clustering, $p(z_i = z_{i'}|\boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X})$. Table 5 summarizes the posterior sample allocations $\hat{\mathbf{z}}$, based on the PPM estimate, and the corresponding ARI value. Our results suggest that the proposed CoPoM model is able to roughly recover distinctive features of the 4 broad categories.

For comparison, we already pointed out that the `R` packages edgeR and PoiClaClu do not provide a single point estimate, $\hat{\mathbf{z}}$, but yield a dissimilarity matrix that can be used as input in a hierarchical clustering algorithm. Within our approach, the squared Euclidean distance between each pair of observations, based on the selected subset of discriminatory features, can be defined as

$$d(\mathbf{x}_{i\cdot}^{(\gamma)}, \mathbf{x}_{i'\cdot}^{(\gamma)}) = \sqrt{\sum_{\{j:\gamma_j=1\}} \left(d^{*}_{z_i j} - d^{*}_{z'_i j}\right)^2},$$

TABLE 5 Brown Corpus data set: Point estimate $\hat{\mathbf{z}}_{\text{PPM}}$ of cluster membership obtained from the pairwise probability matrix of co-clustering. See Section 3.2 for details

	Sports					Society			Spot news							Editoral							ARI
$\hat{\mathbf{z}}_{\text{PPM}}$	1	1	1	1	1	2	2	2	3	3	4	4	5	3	5	6	6	6	6	6	6	6	0.788

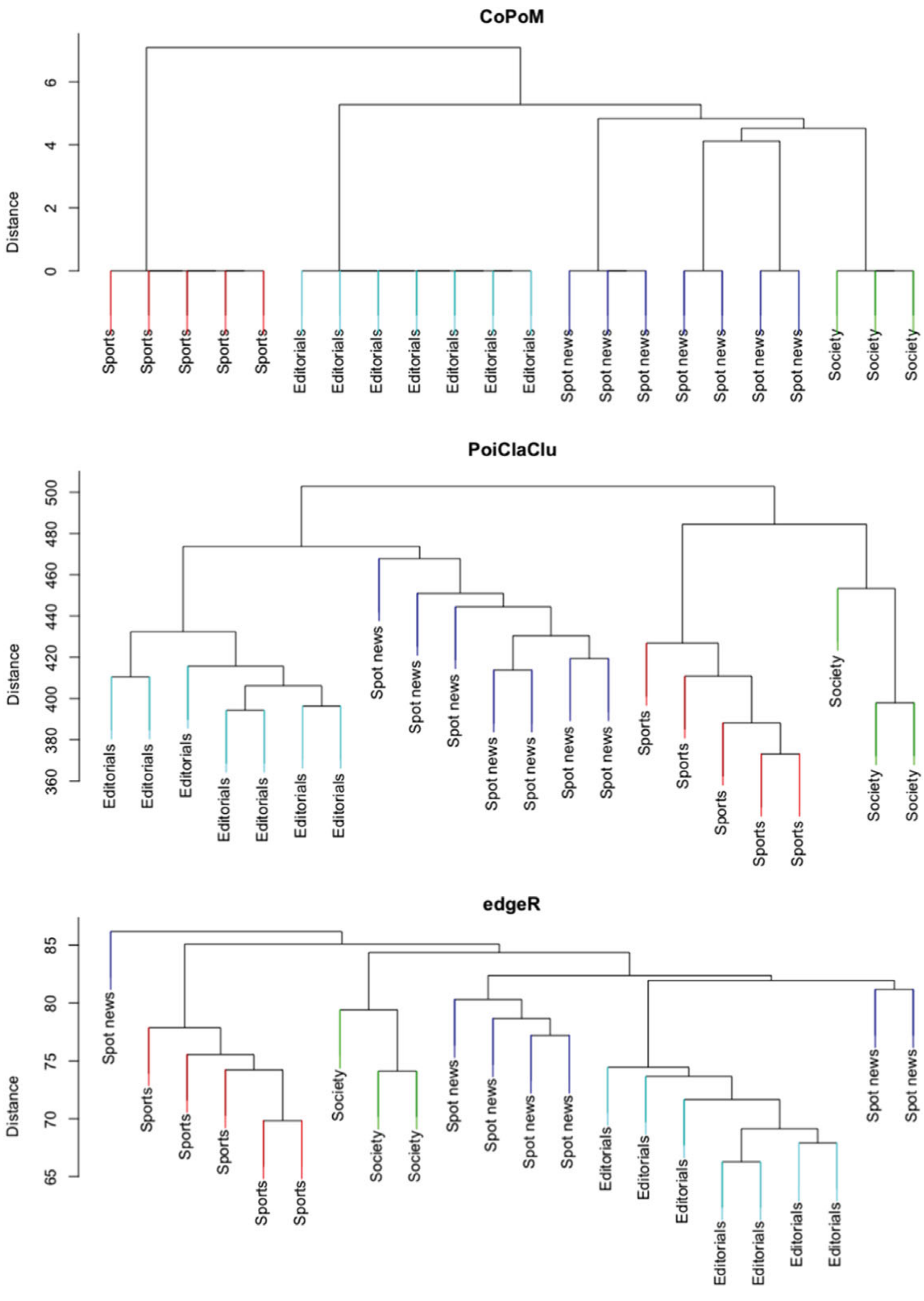

FIGURE 8 Bag-of-words data: The hierarchical clustering dendrograms using the Euclidean distance measure and the complete agglomeration method achieved by the CoPoM model, PoiClaClu and edgeR, respectively

and an estimate of d^*_{kj} can be computed as

$$\hat{d}^*_{kj} = \frac{\sum_{\{i:\hat{z}_i=k,r_{ij}=0\}} x_{ij}}{\hat{g}_j \sum_{\{i:\hat{z}_i=k,r_{ij}=0\}} \hat{s}_i},$$

with $\hat{\mathbf{z}}_{\mathrm{PPM}}$ and $\hat{\mathbf{s}}$ and $\hat{\mathbf{g}}$ the MAP estimates of the parameters. Note that the distance between each pair of observations within the same cluster is zero. Figure 8 shows the dendrogram of the hierarchical cluster analysis for the CoPoM model and the 2 R packages edgeR and PoiClaClu [59], employing the Euclidean distance and the complete agglomeration method. Our CoPoM model and Witten's methods appear to be the most efficient to separate the 4 genres.

5 | DISCUSSION

In this paper, we have introduced a zero-inflated CoPoM model for the analysis of count data, where multiple features are observed as counts over a number of samples. We assume that the data are characterized by an excessive number of zeros and by overdispersion, that is a large number of features do not get frequently observed, whereas a few features are characterized by large occurrence rates in 1 or more samples. We assume that the amount of overdispersion may also vary from sample to sample. Our CoPoM approach proposes to regularize the estimation of the feature occurrence rates by means of a model-based normalization approach, which employs prior distributions with mean constraints on some of the model parameters. It also incorporates a feature selection mechanism, for the simultaneous identification of a parsimonious set of discriminatory features and group structures in the samples.

When applied to simulated data, the CoPoM model has shown improved accuracy of the clustering performance with respect to more standard approaches. We have also presented an application to a *bag-of-words* benchmark data set, where the data matrix is represented by the frequencies of word occurrences in multiple documents. We have shown the good performance of our Poisson mixture model with mean constraints in terms of feature selection and the clustering of the samples into larger topical groups.

Our modeling framework enables discrimination of samples based on a subset of selected features. Alternative feature selection approaches are often used to describe overlapping clustering of (latent) features across samples, [9,38,39], with resulting post hoc interpretation of the subset of nonexclusive pairings. Furthermore, several models have been proposed recently for the analysis of overdispersed count data as an alternative to Poisson mixtures. For example, in the Bayesian nonparametric literature, rounded Gaussian kernel models have been introduced as a more flexible and robust choice for analyzing both underdispersed and overdispersed count data [15, 16]. To date rounded Gaussian kernel methods have been developed mostly for density estimation. In our article, instead, we are concerned with the estimation and comparison of features' occurrence rates across multiple samples, and feature selection, which require some degree of regularization due to the nature of the data generating mechanism. Future work will aim at exploring how our inferential aims can be comprised into the domain of those more recent approaches.

Although here we have focused on the analysis of text documents, our methodology is quite general. In particular, the class of prior distributions with mean constraints that we propose can be successfully employed on other types of high-dimensional count data sets, such as those encountered in ecology, genomics and spatial statistics, where the normalization of Poisson intensities is commonly employed to account for the overdispersion and heterogeneity observed across samples and across features. Future work will explore how the approaches presented here can be extended to the analysis of multivariate vectors of dependent count data, observed spatially or longitudinally in time.

REFERENCES

1. E. M. Airoldi and J. M. Bischof, *Improving and Evaluating Topic Models and Other Models of Text*, J. Am. Stat. Assoc. **111** (2016), 1381–1403.
2. E. M. Airoldi et al., *Who wrote Ronald Reagan's radio addresses?*, Bayesian Anal. **1** (2006), 289–319.
3. E. M. Airoldi et al., *Reconceptualizing the classification of PNAS articles*, Proc. Natl Acad. Sci. **107** (2010), 899–904.
4. E. M. Airoldi et al. (eds.), *Handbook of mixed membership models and their applications*, CRC Press, 2014.
5. S. Anders and W. Huber, *Differential expression analysis for sequence count data*, Genome Biol. **11** (2010), no. 10, R106.
6. C. E. Antoniak, *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*, Ann. Stat. **2** (1974), no. 6, 1152–1174.
7. S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modeling and analysis for spatial data,* CRC Press, 2014.
8. S. Blackshaw et al., *Genomic analysis of mouse retinal development*, PLoS Biol. **2** (2004), e247.
9. T. Broderick, J. Pitman, and M. I. Jordan, *Feature allocations, probability functions, and paintboxes*, Bayesian Anal. **12** (2013), no. 4, 801–836.
10. P. Brown, M. Vannucci, and T. Fearn, *Bayesian wavelength selection in multicomponent analysis*, J. Chemom. **12** (1998), no. 3, 173–182.
11. P. J. Brown, M. Vannucci, and T. Fearn, *Multivariate Bayesian variable selection and prediction*, J. R. Stat. Soc. B Stat. Methodol. **60** (1998), no. 3, 627–641.
12. J. H. Bullard et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments*, BMC Bioinform. **11** (2010), no. 1, 94.
13. A. C. Cameron and P. K. Trivedi, *Microeconometrics: methods and applications,* Cambridge University Press, 2005.
14. *Regression analysis of count data*, vol. 53 Cambridge University Press, 2013.
15. A. Canale and D. B. Dunson, *Bayesian kernel mixtures for counts*, J. Am. Stat. Assoc. **106** (2011), no. 496, 1528–1539.
16. A. Canale and I. Prünster, *Robustifying Bayesian nonparametric mixtures for count data*, Biometrics **73** (2017), no. 1, 174–184.
17. J. Chen and H. Li, *Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis*, Ann. Appl. Stat. **7** (2013).
18. H. Crane and etaltext, *The ubiquitous ewens sampling formula*, Stat. Sci. **31** (2016), no. 1, 1–19.
19. D. B. Dahl, *Model-based clustering for expression data via a Dirichlet process mixture model*, In Bayesian Inference for Gene Expression and Proteomics (Kim-Anh Do, Peter Mueller, and Marina Vannucci, eds.), Cambridge University Press.
20. M. D. Escobar and M. West, *Bayesian density estimation and inference using mixtures*, J. Am. Stat. Assoc. **90** (1995), no. 430, 577–588.
21. W. J. Ewens, *The sampling theory of selectively neutral alleles*, Theor. Popul. Biol. **3** (1972), no. 1, 87–112.
22. J. M. Gee et al., *Effects of organic enrichment on meiofaunal abundance and community structure in sublittoral soft sediments*, J. Exp. Mar. Biol. Ecol. **91** (1985), 247–262.
23. A. Gelman, *Objections to Bayesian statistics*, Bayesian Anal. **3** (2008), no. 3, 445–449.
24. E. I. George and R. E. McCulloch, *Approaches for Bayesian variable selection*, Stat. Sin. **7** (1997), no. 2, 339–373.
25. T. L. Griffiths and Z. Ghahramani, *The Indian Buffet process: an introduction and review*, J. Mach. Learn. Res. **12** (2011), 1185–1224.
26. T. L. Griffiths and M. Steyvers, *Finding scientific topics*, Proc. Natl. Acad. Sci. **101** (2004), no. 1, 5228–5235.
27. M. Guindani, P. Müller, and S. Zhang, *A Bayesian discovery procedure*, J. R. Stat. Soc. B Stat. Methodol. **71** (2009), no. 5, 905–925.
28. M. Guindani et al., *A Bayesian semiparametric approach for the differential analysis of sequence counts data*, J. R. Stat. Soc. C Appl. Stat.) **63** (2014), no. 3, 385–404.

29. P. Gustafson, *On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with discussion)*, Stat. Sci. **20** (2005), 111–140.
30. N. L. Hjort et al., *Bayesian nonparametrics*, vol. 28 Cambridge University Press, 2010.
31. P. Hoff, *Nonparametric estimation of convex models via mixture*, Ann. Stat. **31** (2003), 174–200.
32. L. Hubert and P. Arabie, *Comparing partitions*, J. Classif. **2** (1985), no. 1, 193–218.
33. H. Ishwaran and L. F. James, *Gibbs sampling methods for stick-breaking priors*, J. Am. Stat. Assoc. **96** (2001), no. 453, 161–173.
34. S. Kim, M. G. Tadesse, and M. Vannucci, *Variable selection in clustering via Dirichlet process mixture models*, Biometrika **93** (2006), no. 4, 877–893.
35. H. Kucera and N. Francis, *Computational analysis of present-day American English,* Brown University Press, 1967.
36. M. Kyung, J. Gill, and G. Casella, *Sampling schemes for generalized linear Dirichlet process random effects models*, Stat. Methods Appl. **20** (2011), no. 3, 259–290.
37. D. Lambert, *Zero-inflated Poisson regression, with an application to defects in manufacturing*, Technometrics **34** (1992), 1–14.
38. J. Lee et al., *A Bayesian feature allocation model for tumor heterogeneity*, Ann. Appl. Stat **06** (2015), no. 2, 621–639.
39. J. Lee et al., *Bayesian inference for intratumour heterogeneity in mutations and copy number variation*, J. R. Stat. Soc. C Appl. Stat. **65** (2016), no. 4, 547–563.
40. J. C. Marioni et al., *RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays*, Gen. Res. **18** (2008), 1509–1517.
41. C. N. Morris, *Parametric empirical Bayes inference: theory and applications*, J. Am. Stat. Assoc. **78** (1983), no. 381, 47–55.
42. A. Mortazavi et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq*, Nat. Methods **5** (2008), no. 7, 621–628.
43. P. Müller et al., *Optimal sample size for multiple testing: the case of gene expression microarrays*, J. Am. Stat. Assoc. **99** (2004), no. 468, 990–1001.
44. P. Müller et al., *Bayesian nonparametric data analysis,* Springer, 2015.
45. R. M. Neal, *Markov chain sampling methods for Dirichlet process mixture models*, J. Comput. Graph. Stat. **9** (2000), no. 2, 249–265.
46. A. E. Raftery and N. Dean, *Variable selection for model-based clustering*, J. Am. Stat. Assoc. **101** (2006), no. 473, 168–178.
47. W. M. Rand, *Objective criteria for the evaluation of clustering methods*, J. Am. Stat. Assoc. **66** (1971), no. 336, 846–850.
48. B. J. Reich, H. D. Bondell, and H. J. Wang, *Flexible Bayesian quantile regression for independent and clustered data*, Biostatistics **11** (2010), no. 2, 337–352.
49. S. Richardson and P. J. Green, *On Bayesian analysis of mixtures with an unknown number of components (with discussion)*, J. R. Stat. Soc. B Stat. Methodol.) **59** (1997), no. 4, 731–792.
50. M. E. Roberts, B. M. Stewart, and E. M. Airoldi, *A model of text for experimentation in the social sciences*, J. Am. Stat. Assoc. **111** (2016), 988–1003.
51. M. D. Robinson and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-Seq data*, Genome Biol. **11** (2010), no. 3, R25.
52. M. D. Robinson, D. J. McCarthy, and G. K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*, Bioinformatics **26** (2010), no. 1, 139–140.
53. J. M. Santos and M. Embrechts, *On the use of the adjusted Rand index as a metric for evaluating supervised classification*, 2009. Artificial Neural Networks–ICANN 2009, Springer.
54. J. Sethuraman, *A constructive definition of Dirichlet priors*, Stat. Sin. **4** (1994), 639–650.
55. M. A. Taddy and A. Kottas, *Mixture modeling for Marked Poisson processes*, Bayesian Anal. **7** (2012), no. 2, 335–362.
56. M. G. Tadesse, N. Sha, and M. Vannucci, *Bayesian variable selection in clustering high-dimensional data*, J. Am. Stat. Assoc. **100** (2005), no. 470, 602–617.
57. L. Trippa and G. Parmigiani, *False discovery rates in somatic mutation studies of cancer*, Ann. Appl. Stat. (2011), 1360–1378.
58. S. Williamson et al., *The IBP compound Dirichlet process and its application to focused topic modeling*, 2010. Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel.
59. D. M. Witten, *Classification and clustering of sequencing data using a Poisson model*, Ann. Appl. Stat. **5** (2011), 2493–2518.
60. M. Zhou, *Beta-negative binomial process and exchangeable random partitions for mixed-membership modeling*, Neural Information Processing Systems (NIPS2014), Montreal, Canada, 2014.

How to cite this article: Li Q, Guindani M, Reich B, Bondell H, Vannucci M. A Bayesian Mixture Model for Clustering and Selection of Feature Occurrence Rates under Mean Constraints. *Stat Anal Data Min: The ASA Data Sci Journal.* 2017;10:393–409. https://doi.org/10.1002/sam.11350.

APPENDIX A: DETAILS OF THE MCMC ALGORITHM

We start by writing the marginal likelihood for each sample i, $i = 1, \dots, n$, after integrating out the parameters $\mathbf{d}^*_{kj}$.

$$
\begin{aligned}
& f(\mathbf{x}_{i\cdot}|\boldsymbol{\gamma}, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot}) \\
&\quad = \int f(\mathbf{x}_{i\cdot}|z_i = k, \boldsymbol{\gamma}, \mathbf{d}^*_{k\cdot}, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot}) p(\mathbf{d}^*_{k\cdot})\, \mathrm{d}\mathbf{d}^*_{k\cdot} \\
&\quad = \int \prod_{\{j:\gamma_j=1, r_{ij}=0\}} \mathrm{Poi}(x_{ij}; s_i g_j d^*_{kj}) \prod_{\{j:\gamma_j=0, r_{ij}=0\}} \mathrm{Poi}(x_{ij}; s_i g_j) \\
&\qquad \times \prod_{\{j:\gamma_j=1, r_{ij}=0\}} \mathrm{Ga}(d^*_{kj}; a, b)\, \mathrm{d}d^*_{kj} \\
&\quad = \prod_{\{j:r_{ij}=0\}} \frac{(s_i g_j)^{x_{ij}}}{x_{ij}!} \exp\left\{ -s_i \sum_{\{j:\gamma_j=0, r_{ij}=0\}} g_j \right\} \\
&\qquad \times \prod_{\{j:\gamma_j=1, r_{ij}=0\}} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + x_{ij})}{(b + s_i g_j)^{a + x_{ij}}}.
\end{aligned}
$$

where $s_i = \exp\{\tilde{\mathbf{s}}_i\}$, $g_j = \exp\{\tilde{\mathbf{g}}_j\}$, and $d^*_{kj} = \exp\{\tilde{d}^*_{kj}\}$

In terms of each feature $j, j = 1, \dots, p$, we write

$$
f(\mathbf{x}_{\cdot j}|\mathbf{z}, \gamma_j = 0, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}_j, \mathbf{r}_{\cdot j}) = \prod_{\{i:r_{ij}=0\}} \frac{(s_i g_j)^{x_{ij}}}{x_{ij}!} \exp\left\{ -g_j \sum_{\{i:r_{ij}=0\}} s_i \right\}
$$

and

$$
\begin{aligned}
f(\mathbf{x}_{\cdot j}|\mathbf{z}, \gamma_j = 1, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}_j, \mathbf{r}_{\cdot j}) &= \prod_{\{i:r_{ij}=0\}} \frac{(s_i g_j)^{x_{ij}}}{x_{ij}!} \left(\frac{b^a}{\Gamma(a)} \right)^K \\
&\quad \times \prod_{k=1}^{K} \frac{\Gamma(a + \sum_{\{i:z_i=k, r_{ij}=0\}} x_{ij})}{(b + g_j \sum_{\{i:z_i=k, r_{ij}=0\}} s_i)^{a + \sum_{\{i:z_i=k, r_{ij}=0\}} x_{ij}}}.
\end{aligned}
$$

At each MCMC iteration, we perform the following steps:

Update of the set of discriminatory features $\boldsymbol{\gamma}$: We randomly perform an *add-delete-swap* step. We repeat this step 20 times to ensure validation of the feature selection for each given cluster assignment. The general Hasting ratio can be written as

$$\mathrm{r} = \frac{p(\boldsymbol{\gamma}^* | \mathbf{z}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}_j, \mathbf{R}, \omega, \mathbf{X})}{p(\boldsymbol{\gamma}^{(b-1)} | \mathbf{z}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}_j, \mathbf{R}, \omega, \mathbf{X})} \frac{J(\boldsymbol{\gamma}^{(b-1)} | \boldsymbol{\gamma}^*)}{J(\boldsymbol{\gamma}^* | \boldsymbol{\gamma}^{(b-1)})}$$

$$= \frac{f(\mathbf{X} | \mathbf{z}, \boldsymbol{\gamma}^*, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}_j, \mathbf{R})}{f(\mathbf{X} | \mathbf{z}, \boldsymbol{\gamma}^{(b-1)}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}_j, \mathbf{R})} \frac{p(\boldsymbol{\gamma}^* | \omega)}{p(\boldsymbol{\gamma}^{(b-1)} | \omega)} \frac{J(\boldsymbol{\gamma}^{(b-1)} | \boldsymbol{\gamma}^*)}{J(\boldsymbol{\gamma}^* | \boldsymbol{\gamma}^{(b-1)})}.$$

For both add/delete and swap steps, the proposal density ratio equals to 1.

More specifically, for the add-detele step, we randomly sample a feature $j, j = 1, \ldots, p$ and propose to change the value of γ_j with probability $\min(1, \mathrm{m}_{\mathrm{MH}})$, where

$$\mathrm{m}_{\mathrm{MH}}^{\mathrm{add}} = \frac{\left(\frac{b^a}{\Gamma(a)}\right)^K \prod_{k=1}^K \frac{\Gamma(a + \sum_{i:z_i=k, r_{ij}=0} x_{ij})}{(b + g_j \sum_{i:z_i=k, r_{ij}=0} s_i)^{a + \sum_{i:z_i=k, r_{ij}=0} x_{ij}}}}{\exp\left(-g_j \sum_{\{i:r_{ij}=0\}} s_i\right)} \frac{\omega}{1-\omega},$$

$$\mathrm{m}_{\mathrm{MH}}^{\mathrm{del}} = \frac{\exp\left(-g_j \sum_{\{i:r_{ij}=0\}} s_i\right)}{\left(\frac{b^a}{\Gamma(a)}\right)^K \prod_{k=1}^K \frac{\Gamma(a + \sum_{i:z_i=k, r_{ij}=0} x_{ij})}{(b + g_j \sum_{i:z_i=k, r_{ij}=0} s_i)^{a + \sum_{i:z_i=k, r_{ij}=0} x_{ij}}}} \frac{1-\omega}{\omega}.$$

For the swap step, we randomly swap 2 features if applicable, that is, change the value of a feature γ_{j_1} from 1 to 0, and, correspondingly, of another currently nondiscriminatory feature j_2, that is, change the value of γ_{j_2} from 0 to 1, with probability $\min(1, \mathrm{r}_{\mathrm{MH}}^{\mathrm{swp}})$, where

$$\mathrm{m}_{\mathrm{MH}}^{\mathrm{swp}} = \frac{\exp\left(-g_{j_1} \sum_{\{i:r_{ij}=0\}} s_i\right)}{\prod_{k=1}^K \frac{\Gamma(a + \sum_{\{i:z_i=k, r_{ij}=0\}} x_{ij_1})}{(b + g_{j_1} \sum_{\{i:z_i=k, r_{ij}=0\}} s_i)^{a + \sum_{\{i:z_i=k, r_{ij}=0\}} x_{ij_1}}}}$$

$$\times \frac{\prod_{k=1}^K \frac{\Gamma(a + \sum_{\{i:z_i=k, r_{ij}=0\}} x_{ij_2})}{(b + g_{j_2} \sum_{\{i:z_i=k, r_{ij}=0\}} s_i)^{a + \sum_{\{i:z_i=k, r_{ij}=0\}} x_{ij_2}}}}{\exp\left(-g_{j_2} \sum_{\{i:r_{ij}=0\}} s_i\right)}.$$

Finally, we update the hyperparameter ω for inclusion probability,

$$\omega | \boldsymbol{\gamma}, a_\omega, b_\omega \sim \mathrm{Be}(a_\omega + p_\gamma, b_\omega + n - p_\gamma).$$

Update of the cluster allocation for the selected features z: At each iteration, we perform a Gibbs sampling to update z_i sequentially from observation 1 to n according to Algorithm 3 in [45]. In order to do this, we first need to integrate out the parameter $\mathbf{d}_{k\cdot}^*$ based on the its prior and all observations except itself, that is, all of $\mathbf{x}_{i'\cdot}$ for which $i' \neq i$.

$$f(\mathbf{x}_{i\cdot} | z_i = k, \mathbf{z}_{-i}, \boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X}_{-i\cdot})$$

$$= \int f(\mathbf{x}_{i\cdot} | z_i = k, \boldsymbol{\gamma}, \mathbf{d}_{k\cdot}^*, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot})$$

$$\times p(\mathbf{d}_{k\cdot}^* | \mathbf{z}_{-i}, \boldsymbol{\gamma}, \tilde{\mathbf{s}}_{-i}, \tilde{\mathbf{g}}, \mathbf{R}_{-i\cdot}, \mathbf{X}_{-i\cdot}) \, \mathrm{d}\mathbf{d}_{k\cdot}^*$$

$$= \int \prod_{\{j:\gamma_j=1, r_{ij}=0\}} \mathrm{Poi}(x_{ij}; s_i g_j d_{kj}^*) \prod_{\{j:\gamma_j=0, r_{ij}=0\}} \mathrm{Poi}(x_{ij}; s_i g_j)$$

$$\prod_{\{j:\gamma_j=1, r_{ij}=0\}} \mathrm{Ga}\left(d_{kj}^*; a + \sum_{\{i':z_{i'}=k, r_{i'j}=0, i'\neq i\}}\right.$$

$$\left. \times x_{i'j}, b + g_j \sum_{\{i':z_{i'}=k, r_{i'j}=0, i'\neq i\}} s_i'\right) \mathrm{d}d_{kj}^*$$

$$= \prod_{\{j:r_{ij}=0\}} \frac{(s_i g_j)^{x_{ij}}}{x_{ij}!} \exp\left\{-s_i \sum_{\{j:\gamma_j=0, r_{ij}=0\}} g_j\right\}$$

$$\prod_{\{j:\gamma_j=1, r_{ij}=0\}} \frac{\left(b + g_j \sum_{\{i':z_{i'}=k, r_{i'j}=0, i'\neq i\}} s_i'\right)^{a + \sum_{\{i':z_{i'}=k, r_{i'j}=0, i'\neq i\}} x_{i'j}}}{\Gamma(a + \sum_{\{i':z_{i'}=k, r_{i'j}=0, i'\neq i\}} x_{i'j})}$$

$$\prod_{\{j:\gamma_j=1, r_{ij}=0\}} \frac{\Gamma\left(a + \sum_{\{i':z_{i'}=k, r_{i'j}=0\}} x_{i'j}\right)}{\left(b + g_j \sum_{\{i':z_{i'}=k, r_{i'j}=0\}} s_i'\right)^{a + \sum_{\{i':z_{i'}=k, r_{i'j}=0\}} x_{i'j}}}.$$

Then, we propose to either:

- form a new cluster (change the value of z_i to $K_{-i} + 1$) with probability $\mathrm{p}_{K_{-i}+1}$, where

$$\mathrm{p}_{K_{-i}+1} = \frac{\frac{\alpha}{n-1+\alpha} f(\mathbf{x}_{i\cdot} | \boldsymbol{\gamma}, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot})}{\frac{\alpha}{n-1+\alpha} f(\mathbf{x}_{i\cdot} | \boldsymbol{\gamma}, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot}) + \sum_{k=1}^K \frac{n_{k,-i}}{n-1+\alpha} \times f(\mathbf{x}_{i\cdot} | z_i = k, \mathbf{z}_{-i}, \boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X}_{-i\cdot})}.$$

- allocate the observation to any of the existing clusters, say cluster $k \in \{1, \ldots, K_{-i} + 1\}$, with probability p_k, where

$$\mathrm{p}_k = \frac{\frac{n_{k,-i}}{n-1+\alpha} f(\mathbf{x}_{i\cdot} | z_i = k, \mathbf{z}_{-i}, \boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X}_{-i\cdot})}{\frac{\alpha}{n-1+\alpha} f(\mathbf{x}_{i\cdot} | \boldsymbol{\gamma}, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot}) + \sum_{k=1}^K \frac{n_{k,-i}}{n-1+\alpha} \times f(\mathbf{x}_{i\cdot} | z_i = k, \mathbf{z}_{-i}, \boldsymbol{\gamma}, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}, \mathbf{R}, \mathbf{X}_{-i\cdot})}.$$

Finally, we update the concentration parameter α, by following the algorithm in [20], that is we generate an auxiliary variable $\eta_\alpha | \alpha \sim \mathrm{Be}(\alpha + 1, n)$ and then we sample α from a mxiture of 2 gamma densities,

$$\alpha | \eta_\alpha, \mathbf{z} \sim \pi_\eta \mathrm{Ga}(a_\alpha + K, b - \log(\eta_\alpha))$$
$$+ (1 - \pi_\eta) \mathrm{Ga}(a_\alpha + K - 1, b_\alpha - \log(\eta_\alpha)),$$

with the weights π_η defined by $\pi_\eta / (1 - \pi_\eta) = (a_\alpha + K - 1) / (n(b_\alpha - \log(\eta_\alpha)))$.

Update of the scaling factors: We can rewrite the prior distribution in Equations 3 to 4 by introducing latent auxiliary variables, that specify how the $\tilde{s}_i$ and $\tilde{g}_j$ are assigned to any of the inner and outer mixture components. More specifically, we can introduce a $n \times 1$ vector of assignment indicators, $\boldsymbol{\nu}$, with $\nu_i = m$ indicating that $\tilde{s}_i$ is a sample from $f_m^s(\tilde{s}_i|t_m, \eta_m)$. The weights ϕ_m^s determine the probability of each value $\nu_i = m$, with $m = 1, 2, \ldots$. Correspondingly, we can consider a $n \times 1$ vector $\boldsymbol{\epsilon}$ of binary elements ϵ_i, where $\epsilon_i = 1$ indicates that, given $\nu_i = m$, $\tilde{s}_i$ is drawn from a $\mathrm{N}(\eta_m, \sigma_s^2)$ with probability t_m, and $\epsilon_i = 0$ indicates that $\tilde{s}_i = 0$ is drawn from the right component of $f_m^s(\tilde{s}_i|t_m, \eta_m)$, that is, $\mathrm{N}\left(\frac{c_s - t_m\eta_m}{1-t_m}, \sigma_s^2\right)$, with probability $1 - t_m$. Similarly, we can introduce a $p \times 1$ vector $\boldsymbol{\xi}$, with $\xi_j = l$ indicating that $\tilde{g}_j$ is sampled from $f_l^g(\tilde{g}_j|q_l, \mu_l)$, $l = 1, 2, \ldots$, and $\phi_l^g = p(\boldsymbol{\xi} = l)$. Correspondingly, given the assignments obtained in the vector $\boldsymbol{\xi}$, we can define a $p \times 1$ vector $\boldsymbol{\psi}$ of binary elements ψ_j, where $\psi_j = 1$ indicates that, given $\xi_j = l$, then $\tilde{g}_j$ is drawn from $\mathrm{N}(\mu_l, \sigma_g^2)$, whereas $\psi_j = 0$ indicates that $\tilde{g}_j$ is from $\mathrm{N}\left(\frac{c_g - q_l\mu_l}{1-q_l}, \sigma_g^2\right)$. Thus, the prior model Equations 3 and 4 can be rewritten as

$$\tilde{s}_i|\nu_i, \epsilon_i, \mathbf{t}, \boldsymbol{\eta} \sim \mathrm{N}\left(\epsilon_i\eta_{\nu_i} + (1-\epsilon_i)\frac{c_s - t_{\nu_i}\eta_{\nu_i}}{1 - t_{\nu_i}}, \sigma_s^2\right) \tag{8}$$

and

$$\tilde{g}_j|\xi_j, \psi_j, \mathbf{q}, \boldsymbol{\mu} \sim \mathrm{N}\left(\psi_j\mu_{\xi_j} + (1-\psi_j)\frac{c_g - q_{\xi_j}\mu_{\xi_j}}{1 - q_{\xi_j}}, \sigma_g^2\right), \tag{9}$$

where $\mathbf{t}$, $\boldsymbol{\eta}$, $\mathbf{q}$, and $\boldsymbol{\mu}$ denote the vectors of t_m, η_m, q_l, and μ_l, respectively. Therefore, the update of the sample- and feature-specific scaling factors s_i and g_j can proceed as follows, after logarithmic transformation:

a) Update of the $\tilde{\mathbf{s}}_i$'s: We perform Metropolis sampling to update $\tilde{\mathbf{s}}_i$, where $\tilde{\mathbf{s}}_i = \log\{s_i\}$, sequentially from observation 1 to n. We propose a new $\tilde{s}_i^*$ from $\mathrm{N}(\tilde{s}_i^{(b-1)}, \tau_s^2)$ and accept it with probability $\min(1, \mathrm{m_{MH}})$, where

$$\begin{aligned}
\mathrm{m_{MH}} &= \frac{p(\tilde{\mathbf{s}}_i^*|\boldsymbol{\gamma}, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot}, \nu_i, \epsilon_i, \mathbf{t}, \boldsymbol{\eta}, \mathbf{x}_{i\cdot})}{p(\tilde{\mathbf{s}}_i^{(b-1)}|\boldsymbol{\gamma}, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot}, \nu_i, \epsilon_i, \mathbf{t}, \boldsymbol{\eta}, \mathbf{x}_{i\cdot})} \frac{J(\tilde{s}_i^{(b-1)}|\tilde{s}_i^*)}{J(\tilde{s}_i^*|\tilde{s}_i^{(b-1)})} \\
&= \frac{f(\mathbf{x}_{i\cdot}|\boldsymbol{\gamma}, \tilde{\mathbf{s}}_i^*, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot})}{f(\mathbf{x}_{i\cdot}|\boldsymbol{\gamma}, \tilde{\mathbf{s}}_i^{(b-1)}, \tilde{\mathbf{g}}, \mathbf{r}_{i\cdot})} \frac{p(\tilde{s}_i^*|\nu_i, \epsilon_i, \mathbf{t}, \boldsymbol{\eta})}{p(\tilde{s}_i^{(b-1)}|\nu_i, \epsilon_i, \mathbf{t}, \boldsymbol{\eta})} \\
&= \frac{s_i^{*\sum_{\{j: r_{ij}=0\}} x_{ij}} \exp\left(-s_i^* \sum_{\{j:\gamma_j=0, r_{ij}=0\}} g_j\right)}{s_i^{(b-1)\sum_{\{j: r_{ij}=0\}} x_{ij}} \exp\left(-s_i^{(b-1)} \sum_{\{j:\gamma_j=0, r_{ij}=0\}} g_j\right)} \\
&\quad \times \frac{\prod_{j:\gamma_j=1, r_{ij}=0}(b + s_i^* g_j)^{-a-x_{ij}}}{\prod_{j:\gamma_j=1, r_{ij}=0}(b + s_i^{(b-1)} g_j)^{-a-x_{ij}}} \\
&\quad \times \frac{\mathrm{N}\left(\tilde{\mathbf{s}}_i^*; \epsilon_i\eta_{\nu_i} + (1-\epsilon_i)\frac{c_s - t_{\nu_i}\eta_{\nu_i}}{1-t_{\nu_i}}, \sigma_s^2\right)}{\mathrm{N}\left(\tilde{\mathbf{s}}_i^{(b-1)}; \epsilon_i\eta_{\nu_i} + (1-\epsilon_i)\frac{c_s - t_{\nu_i}\eta_{\nu_i}}{1-t_{\nu_i}}, \sigma_s^2\right)}.
\end{aligned}$$

Since $\boldsymbol{\nu}$, $\boldsymbol{\epsilon}$, $\mathbf{t}$, and $\boldsymbol{\eta}$ have conjugate full conditionals, we use Gibbs sampling to update them one after another

- Gibbs sampling for updating $\nu_i, i = 1, \ldots, n$:

$$\pi(\nu_i = l|\boldsymbol{\nu}_{-i}, \epsilon_i, \mathbf{t}, \boldsymbol{\eta}, \tilde{\mathbf{s}}_i) \propto \phi_m^s \mathrm{N} \times \left(\tilde{\mathbf{s}}_i; \epsilon_i\eta_m + (1-\epsilon_i)\frac{c_s - t_m\eta_m}{1 - t_m}, \sigma_s^2\right).$$

- Gibbs sampling for updating $\epsilon_i, i = 1, \ldots, n$:

$$\pi(\epsilon_i|\nu_i = m, \boldsymbol{\epsilon}_{-i}, \mathbf{t}, \boldsymbol{\eta}, \tilde{\mathbf{s}}_i) \propto \begin{cases} (1-t_m)\mathrm{N}\left(\tilde{\mathbf{s}}_i; \frac{c_s - t_m\eta_m}{1-t_m}, \sigma_s^2\right) & \textit{if } \epsilon_i = 0 \\ t_m\mathrm{N}\left(\tilde{\mathbf{s}}_i; \eta_m, \sigma_s^2\right) & \textit{if } \epsilon_i = 1 \end{cases}.$$

- Gibbs sampling for updating $t_m, m = 1, \ldots, M$:

$$t_m|\boldsymbol{\nu}, \boldsymbol{\epsilon} \sim \mathrm{Be}\left(a_t + \sum_{i=1}^n I(\nu_i = m)I(\epsilon_i = 1), b_t \times + \sum_{i=1}^n I(\nu_i = m)I(\epsilon_i = 0)\right).$$

- Gibbs sampling for updating $\eta_m, m = 1, \ldots, M$:

$$\eta_m|\boldsymbol{\nu}, \boldsymbol{\epsilon}, \mathbf{t}, \tilde{\mathbf{s}} \sim \mathrm{N}\left(\frac{c_m/\sigma_s^2}{e_m/\sigma_s^2 + 1/\tau_\eta^2}, \frac{1}{e_m/\sigma_s^2 + 1/\tau_\eta^2}\right),$$

  where $c_m = \sum_{\{i:\nu_i=m, \epsilon_i=1\}} \tilde{\mathbf{s}}_i - \frac{t_m}{1-t_m}\sum_{\{i:\nu_i=m, \epsilon_i=0\}}\left(\tilde{\mathbf{s}}_i - \frac{c_s}{1-t_m}\right)$ and $e_m = \sum_{i=1}^n I(\nu_i = m)I(\epsilon_i = 1) + \sum_{\{i:\nu_i=m, \epsilon_i=0\}}\left(\frac{t_m}{1-t_m}\right)^2$.
- Gibbs Sampling for updating $\phi_m^s, m = 1, \ldots, M$ by stick-breaking process [33]:

$$\begin{aligned}
\phi_1^s &= v_1, \\
\phi_2^s &= (1 - v_1)v_2, \\
&\vdots \\
\phi_M^s &= (1 - v_1)\cdots(1 - v_{M-1})v_M,
\end{aligned}$$

where $v_m|\boldsymbol{\nu} \sim \mathrm{Be}\left(a_m + \sum_{i=1}^n I(\nu_i = m), b_m + \sum_{i=1}^n I(\nu_i > m)\right)$.

b) Update of the g_j's: We perform Metropolis sampling to update $\tilde{\mathbf{g}}_j$, where $\tilde{\mathbf{g}}_j = \log\{g_j\}$, sequentially from feature 1 to p. We propose a new $\tilde{g}_j^*$ from $\mathrm{N}(\tilde{g}_j^{(b-1)}, \tau_g^2)$ and accept it with probability $\min(1, \mathrm{m_{MH}})$, where

$$\begin{aligned}
\mathrm{m_{MH}} &= \frac{p(\tilde{\mathbf{g}}_j^*|\mathbf{z}, \gamma_j, \tilde{\mathbf{s}}, \mathbf{r}_{\cdot j}, \xi_j, \psi_j, \mathbf{q}, \boldsymbol{\mu}, \mathbf{x}_{\cdot j})}{p(\tilde{\mathbf{g}}_j^{(b-1)}|\mathbf{z}, \gamma_j, \tilde{\mathbf{s}}, \mathbf{r}_{\cdot j}, \xi_j, \psi_j, \mathbf{q}, \boldsymbol{\mu}, \mathbf{x}_{\cdot j})} \frac{J(\tilde{g}_j^{(b-1)}|\tilde{\mathbf{g}}_j^*)}{J(\tilde{g}_j^*|\tilde{g}_j^{(b-1)})} \\
&= \frac{f(\mathbf{x}_{\cdot j}|\mathbf{z}, \gamma_j, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}_j^*, \mathbf{r}_{\cdot j})}{f(\mathbf{x}_{\cdot j}|\mathbf{z}, \gamma_j, \tilde{\mathbf{s}}, \tilde{\mathbf{g}}_j^{(b-1)}, \mathbf{r}_{\cdot j})} \frac{p(\tilde{g}_j^*|\xi_j, \psi_j, \mathbf{q}, \boldsymbol{\mu})}{p(\tilde{g}_j^{(b-1)}|\xi_j, \psi_j, \mathbf{q}, \boldsymbol{\mu})}
\end{aligned}$$

$$= \begin{cases} \dfrac{g_j^{*\sum_{\{i:r_{ij}=0\}} x_{ij}}}{g_j^{(b-1)\sum_{\{i:r_{ij}=0\}} x_{ij}}} \dfrac{\exp\left\{-g_j^* \sum_{\{i:r_{ij}=0\}} s_i\right\}}{\exp\left\{-g_j^{(b-1)} \sum_{\{i:r_{ij}=0\}} s_i\right\}} & \\ \qquad \times \dfrac{\mathrm{N}\left(\tilde{\mathbf{g}}_j^*;\psi_j\mu_{\xi_j}+(1-\psi_j)\frac{c_g-q_{\xi_j}\mu_{\xi_j}}{1-q_{\xi_j}},\sigma_g^2\right)}{\mathrm{N}\left(\tilde{\mathbf{g}}_j^{(b-1)};\psi_j\mu_{\xi_j}+(1-\psi_j)\frac{c_g-q_{\xi_j}\mu_{\xi_j}}{1-q_{\xi_j}},\sigma_g^2\right)} & \mathit{if}\ \gamma_j = 0 \\ \dfrac{g_j^{*\sum_{\{i:r_{ij}=0\}} x_{ij}} \prod_{k=1}^K (b+g_j^* \sum_{\{i:z_i=k,r_{ij}=0\}} s_i)^{-a-\sum_{\{i:z_i=k,r_{ij}=0\}} x_{ij}}}{g_j^{(b-1)\sum_{\{i:r_{ij}=0\}} x_{ij}} \prod_{k=1}^K (b+g_j^{(b-1)} \sum_{\{i:z_i=k,r_{ij}=0\}} s_i)^{-a-\sum_{\{i:z_i=k,r_{ij}=0\}} x_{ij}}} & \\ \qquad \times \dfrac{\mathrm{N}\left(\tilde{\mathbf{g}}_j^*;\psi_j\mu_{\xi_j}+(1-\psi_j)\frac{c_g-q_{\xi_j}\mu_{\xi_j}}{1-q_{\xi_j}},\sigma_g^2\right)}{\mathrm{N}\left(\tilde{\mathbf{g}}_j^{(b-1)};\psi_j\mu_{\xi_j}+(1-\psi_j)\frac{c_g-q_{\xi_j}\mu_{\xi_j}}{1-q_{\xi_j}},\sigma_g^2\right)} & \mathit{if}\ \gamma_j = 1 \end{cases}.$$

Since $\boldsymbol{\xi}$, $\boldsymbol{\psi}$, $\mathbf{q}$, and $\boldsymbol{\mu}$ have conjugate full conditionals, we use Gibbs sampling to update them one after another

- Gibbs sampling for updating $\xi_j, j = 1, \ldots, p$:

$$p(\xi_j = l|\boldsymbol{\xi}_{-j}, \psi_j, \mathbf{q}, \boldsymbol{\mu}, \tilde{\mathbf{g}}_j)$$
$$\propto \phi_l^g \mathrm{N}\left(\tilde{\mathbf{g}}_j; \psi_j\mu_l + (1-\psi_j)\frac{c_g - q_l\mu_l}{1-q_l}, \sigma_g^2\right).$$

- Gibbs sampling for updating $\psi_j, j = 1, \ldots, p$:

$$\pi(\psi_j|\xi_j = l, \boldsymbol{\psi}_{-j}, \mathbf{q}, \boldsymbol{\mu}, \tilde{\mathbf{g}}_j)$$
$$\propto \begin{cases} (1-q_l)\mathrm{N}\left(\tilde{\mathbf{g}}_j; \frac{c_g-q_l\mu_l}{1-q_l}, \sigma_g^2\right) & \mathit{if}\ \psi_j = 0 \\ q_l\mathrm{N}\left(\tilde{\mathbf{g}}_j; \mu_l, \sigma_g^2\right) & \mathit{if}\ \psi_j = 1 \end{cases}.$$

- Gibbs sampling for updating $q_l, l = 1, \ldots, L$:

$$q_l|\boldsymbol{\xi}, \boldsymbol{\psi} \sim \mathrm{Be}\left(a_q + \sum_{j=1}^p I(\xi_j = l)I(\psi_j = 1), b_q + \sum_{j=1}^p I(\xi_j = l)I(\psi_j = 0)\right).$$

- Gibbs sampling for updating $\mu_l, l = 1, \ldots, L$:

$$\mu_l|\boldsymbol{\xi}, \boldsymbol{\psi}, \mathbf{q}, \tilde{\mathbf{g}} \sim \mathrm{N}\left(\frac{c_l/\sigma_g^2}{e_l/\sigma_g^2 + 1/\tau_\mu^2}, \frac{1}{e_l/\sigma_g^2 + 1/\tau_\mu^2}\right),$$

where $c_l = \sum_{\{j:\xi_j=l,\psi_j=1\}} \tilde{\mathbf{g}}_j - \frac{q_l}{1-q_l}\sum_{\{j:\xi_j=l,\psi_j=0\}}\left(\tilde{\mathbf{g}}_j - \frac{c_g}{1-q_l}\right)$ and $e_l = \sum_{j=1}^p I(\xi_j = l)I(\psi_j = 1) + \sum_{\{j:\xi_j=l,\psi_j=0\}}\left(\frac{q_l}{1-q_l}\right)^2$.

- Gibbs Sampling for updating $\phi_l^g, l = 1, \ldots, L$ by stick-breaking process [33]:

$$\begin{aligned} \phi_1^g &= v_1, \\ \phi_2^g &= (1-v_1)v_2, \\ &\vdots \\ \phi_L^g &= (1-v_1)\cdots(1-v_{L-1})v_L, \end{aligned}$$

where $v_l|\boldsymbol{\xi} \sim \mathrm{Be}\left(1 + \sum_{j=1}^p I(\xi_j = l), 1 + \sum_{j=1}^p I(\xi_j > l)\right)$.

Update of zero-inflation latent indicator r_{ij}: The full conditionals of the r_{ij}'s can be obtained after considering that we need to consider only those cases for which $x_{ij} = 0$, and

$$\begin{aligned} &p(r_{ij}|x_{ij} = 0, z_i = k, \gamma_j, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}_j, d_{kj}^*, \pi) \\ &\quad \propto f(x_{ij} = 0|z_i = k, \gamma_j, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}_j, d_{kj}^*, r_{ij})p(r_{ij}|\pi) \\ &\quad = (\exp\left\{-s_ig_jd_{kj}^{*\,\gamma_j}\right\})^{1-r_{ij}}\pi^{r_{ij}}(1-\pi)^{1-r_{ij}} \\ &\quad = \pi^{r_{ij}}\left[(1-\pi)\exp\left\{-s_ig_jd_{kj}^{*\,\gamma_j}\right\}\right]^{1-r_{ij}}. \end{aligned}$$

In order to sample from the above full conditional, we proceed with a Metropolis-Hasting within Gibbs approach, where within each iteration we first propose a value d_{kj}^* for which $\gamma_j = 0$ by sampling

$$d_{kj}^*|\cdot \sim \mathrm{Ga}\left(a + \sum_{\{i:z_i=k,r_{ij}=0\}} x_{ij}, b + g_j \sum_{\{i:z_i=k,r_{ij}=0\}} s_i\right),$$

and then propose a value for r_{ij} by sampling

$$p(r_{ij}|x_{ij} = 0, z_i = k, \gamma_j, \tilde{\mathbf{s}}_i, \tilde{\mathbf{g}}_j, d_{kj}^*, \pi)$$
$$= \begin{cases} \dfrac{\pi^{r_{ij}}\left[(1-\pi)e^{-s_ig_j}\right]^{1-r_{ij}}}{\pi+(1-\pi)\exp\left\{-s_ig_j\right\}} & \mathit{if}\ \gamma_j = 0 \\ \dfrac{\pi^{r_{ij}}\left[(1-\pi)\exp\left\{-s_ig_jd_{kj}^*\right\}\right]^{1-r_{ij}}}{\pi+(1-\pi)\exp\left\{-s_ig_jd_{kj}^*\right\}} & \mathit{if}\ \gamma_j = 1. \end{cases}$$

After the Metropolis-Hasting step, we use a Gibbs sampling step to update π, as

$$\pi|\mathbf{R}, a_\pi, b_\pi \sim \mathrm{Be} \times \left(a_\pi + \sum_{i=1}^n\sum_{j=1}^p r_{ij}, b_\pi + \sum_{i=1}^n\sum_{j=1}^p I(x_{ij} = 0) - \sum_{i=1}^n\sum_{j=1}^p r_{ij}\right).$$