PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Cite this article: Franzolini B, Cremaschi A, van den Boom W, De Iorio M. 2023 Bayesian clustering of multiple zero-inflated outcomes. *Phil. Trans. R. Soc. A* **381**: 20220145. https://doi.org/10.1098/rsta.2022.0145

Received: 28 April 2022
Accepted: 15 September 2022

One contribution of 16 to a theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

Subject Areas:
statistics

Keywords:
conditional algorithm, excess-of-zeros data, enriched priors, hurdle model, finite mixtures, nested clustering

Author for correspondence:
Maria De Iorio
e-mail: m.deiorio@ucl.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6423945.

THE ROYAL SOCIETY PUBLISHING

Bayesian clustering of multiple zero-inflated outcomes

Beatrice Franzolini[1], Andrea Cremaschi[1], Willem van den Boom[2] and Maria De Iorio[1,2,3]

[1]Singapore Institute for Clinical Sciences (SICS), Agency for Science, Technology and Research (A∗STAR), Singapore, Republic of Singapore
[2]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Republic of Singapore
[3]Department of Statistical Science, University College London, London, UK

WvdB, 0000-0002-1777-3857; MDI, 0000-0003-3109-0478

Several applications involving counts present a large proportion of zeros (excess-of-zeros data). A popular model for such data is the hurdle model, which explicitly models the probability of a zero count, while assuming a sampling distribution on the positive integers. We consider data from multiple count processes. In this context, it is of interest to study the patterns of counts and cluster the subjects accordingly. We introduce a novel Bayesian approach to cluster multiple, possibly related, zero-inflated processes. We propose a joint model for zero-inflated counts, specifying a hurdle model for each process with a shifted Negative Binomial sampling distribution. Conditionally on the model parameters, the different processes are assumed independent, leading to a substantial reduction in the number of parameters as compared with traditional multivariate approaches. The subject-specific probabilities of zero-inflation and the parameters of the sampling distribution are flexibly modelled via an *enriched* finite mixture with random number of components. This induces a two-level clustering of the subjects based on the zero/non-zero patterns (outer clustering) and on the sampling distribution (inner clustering). Posterior inference is performed through tailored Markov chain Monte Carlo schemes. We demonstrate the proposed approach on an application involving the use of the messaging service WhatsApp.

This article is part of the theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

1. Introduction

Count data presenting excess of zeros are commonly encountered in applications. These can arise in several settings, such as healthcare, medicine or sociology. In this scenario, the observations carry structural information about the data-generating process, i.e. an *inflation* of zeros. The analysis of zero-inflated data requires the specification of models beyond standard count distributions, such as Poisson or Negative Binomial. Commonly used models are the zero-inflated [1], the hurdle [2] and the zero-altered [3] models. The first class assumes the existence of a probability mass at zero and a distribution over $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. This type of model explicitly differentiates between the zeros originating from a common underlying process, such as the utilization of a service, described by the sampling distribution on $\mathbb{N}_0$, and those arising from a structural phenomenon, such as the ineligibility to use the service, which are modelled by the point mass. Very popular zero-inflated models are the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) models, where the sampling distribution is chosen to be a Poisson and a negative binomial, respectively. These models allow for inflation in the number of zeros and departures from standard distributional assumptions on the moments of the sampling distribution. For instance, the ZIP model allows the mean and the variance of the distribution to be different from each other (as opposed to a standard Poisson distribution), while the ZINB additionally captures overdispersion in the data.

Hurdle models are a very popular choice of distributions for modelling zero-inflated counts. Differently from the zero-inflated ones, these models handle zeros and positive observations separately, assuming on the latter a sampling distribution with support on $\mathbb{N} = \mathbb{N}_0 \setminus \{0\}$. Thus the distribution of the count data is given by

$$\mathbb{P}(Y_i = y_i) = \begin{cases} (1 - p_i), & y_i = 0 \\ p_i g(y_i \mid \boldsymbol{\mu}_i), & y_i > 0 \end{cases} \tag{1.1}$$

where p_i and g now capture two distinct features of the data. Hurdle models present appealing features that can make them preferable to zero-inflated models. Firstly, hurdle distributions allow for both inflation and deflation of zero counts. Indeed, under a zero-inflated model, the probability of observing a zero is always greater than the corresponding probability under the sampling distribution, thus making it impossible to capture deflation in the number of zeros [4]. Secondly, and more importantly for our work, the probability of zero counts in hurdle models is independent of the parameters controlling the distribution of non-zero counts. This feature improves interpretability and facilitates parameter estimation. Note that the zero-altered model proposed by Heilbron [3] is a modified hurdle model in which the two parts are connected by specifying a direct link between the model parameters.

Univariate models for zero-inflated data can be extended to multivariate settings, where several variables presenting excess of zeros are recorded, e.g. in applications involving questionnaires or microbiome data analysis. In this context, a multivariate extension of the ZIP model has been proposed by Li *et al.* [5], through a finite mixture with ZIP marginals. In this construction, the number of parameters increases linearly as the number d of zero-inflated processes increases, as the total number of parameters is $3d + 2$. See also Liu *et al.* [6,7] and Tian *et al.* [8] for simplified versions of the previous construction involving a smaller number of parameters and better distributional properties.

In a Bayesian parametric setting, Fox [9] proposes the joint modelling of two related zero-inflated outcomes. Their strategy is based on the ZIP model, with the same Bernoulli component to capture the extra zeros for both processes. Correlation between subject-specific outcomes is accounted for through the specification of a joint random effect distribution for the parameters

governing the sampling distribution of the two processes. Alternatively, Lee *et al.* [10] model the binary variables indicating whether an observation is positive or not via a multivariate probit model [11,12]. In this approach, the vectors of latent continuous variables characterizing the multivariate probit are modelled jointly assuming a random unstructured correlation matrix describing their dependence.

In several applications, knowledge relative to the grouping of the subjects is also available, thus providing additional information that can be exploited in the model [13]. Moreover, the clustering structure can be estimated by assuming a prior distribution on the partition of the subjects, e.g. via the popular Dirichlet process [14] or a mixture with a random number of components as proposed by Hu *et al.* [15]. In the context of Bayesian semiparametric approaches, Shuler *et al.* [16] propose to model multivariate zero-inflated count data by linking different Dirichlet process mixtures of ZINB models through the use of the popular dependent Dirichlet process [17]. In particular, the probability of zeros and the sampling distribution are modelled via two distinct single-p DDP, where the location parameters of the mixture depend on a categorical covariate. The proposed approach yields flexible estimation of the partition of the subjects, although it does not allow for sharing of information *a priori* between the two components of the ZINB model, thus yielding two separate clustering structures. A different semiparametric approach is proposed by Arab *et al.* [18], which exploits the multivariate ZIP construction of Li *et al.* [5] to model bivariate count data, but the proportion of zeros and the intensity of the sampling distribution are modelled through the introduction of spline regression terms. The spline approach is flexible and computationally tractable when d is small. For larger dimensions, this model would induce a non-trivial computational burden.

The focus of this work is clustering of individuals based on multiple, possibly related, zero-inflated processes. To this end, we propose a Bayesian approach for joint modelling of zero-inflated count data, based on finite mixtures with a random number of components. In particular, we specify a hurdle model for each process with a shifted negative binomial sampling distribution on the positive integers. Let n denote the sample size and d is the number of processes under study. The subject-specific probabilities of zero-inflation p_{ij} for the ith individual and the jth process, $i = 1, \ldots, n$, $j = 1, \ldots, d$, and the parameter vector of the sampling distribution $\boldsymbol{\mu}_{ij}$ are flexibly modelled via an *enriched* mixture with a random number of components, borrowing ideas from the Bayesian non-parametric literature on the Dirichlet process. One of the main novelties of our work is to combine a recent representation of finite mixture models with a random number of components presented in Argiento & De Iorio [19] with a finite extension of the enriched non-parametric prior proposed by Wade *et al.* [20] to achieve a two-level clustering of the subjects, where at the *outer* level individuals are clustered based on the pattern of zero/non-zero observations, while within each outer cluster they are grouped at a finer level (which we refer to as *inner* level) according to the distribution of the non-zero counts. Figure 1 provides an illustration of the nested clustering structure.

Enriched priors in Bayesian non-parametrics generalize concepts developed by Consonni & Veronese [21], who propose a general methodology for the construction of enriched conjugate families for the parametric natural exponential families. The idea underlying this approach is to decompose the joint prior distribution for a vector of parameters indexing a multivariate exponential family into tractable conditional distributions. In particular, distributions belonging to the multivariate natural exponential family satisfy the *conditional reducibility* property, which allows reparameterizing the distribution in terms of a parameter vector, whose components are variation and likelihood independent. Then, it is possible to construct an enriched standard conjugate family on the parameter vector, closed under i.i.d. sampling, which leads to the breaking down of the global inference procedure into several independent subcomponents. Such parameterization achieves greater flexibility in prior specification relative to the standard conjugate one, while still allowing for efficient computations (see, for example, [22]). An example of this class of parametric priors is the enriched Dirichlet distribution [23].

In a Bayesian non-parametric framework, Wade *et al.* [20] first propose an enrichment of the Dirichlet process [24] that is more flexible with respect to the precision parameter but still

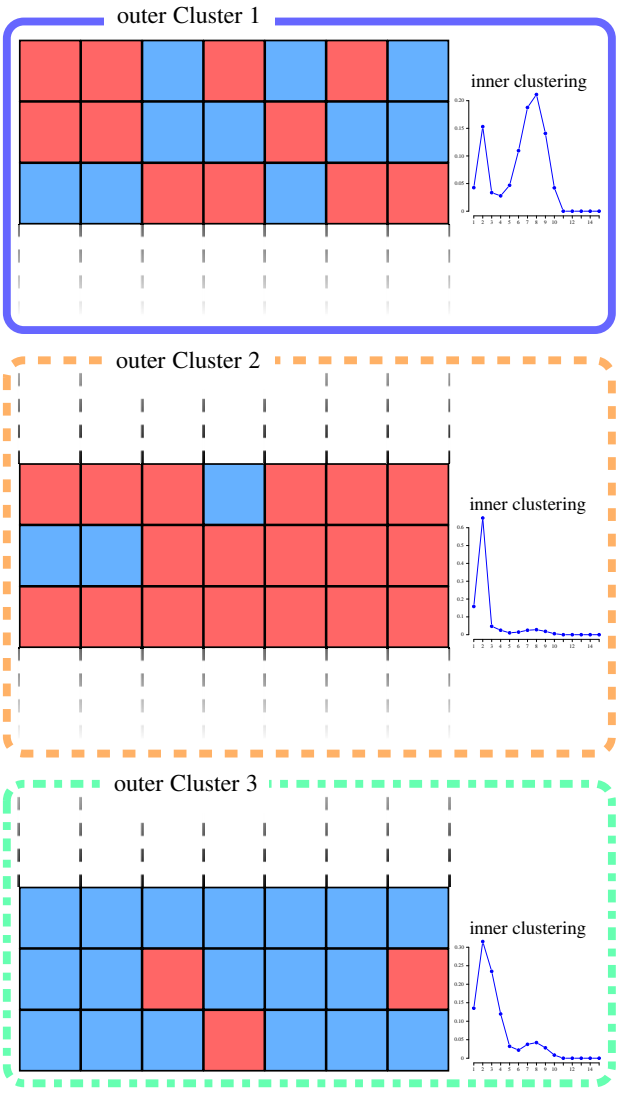

Figure 1. Example of two-level clustering induced by the enriched mixture with a random number of components. The observations are first clustered based on their zero/non-zero patterns. Within each outer cluster, subjects are grouped based on the sampling distribution of the non-zero observations. The inner clustering structure is here depicted via a multimodal discrete distribution, representing a finite mixture. (Online version in colour.)

conjugate, by defining a joint random probability measure on the measurable product space $(\mathcal{X}, \mathcal{Y})$ in terms of the marginal and conditional distributions, P_X and $P_{Y|X}$, and assigning independent Dirichlet process priors to each of these terms. The enriched Dirichlet process enables a nested clustering structure that is particularly appealing in our setting and allows for a finer control of the dependence structure between X and Y. This construction has been employed also in non-parametric regression problems to model the joint distribution of the response and the covariates [25,26], as well as in longitudinal data analysis [27] and causal inference [28]. Recently, Rigon *et al.* [29] proposed the enriched Pitman–Yor process, which leads to a more robust clustering estimation.

In this work, we consider the joint distribution of d zero-inflated process, where the d-dimensional vectors of probabilities $(p_{i1}, \ldots, p_{id})$ correspond to X, while the parameters of the

sampling distributions $\boldsymbol{\mu}_{ij}$ correspond to Y. The enrichment of the prior is achieved by modelling both P_X and $P_{Y|X}$ through a mixture with a random number of components (see, for instance, [30]). We exploit the recent construction by Argiento & De Iorio [19] based on normalized independent finite point processes (Norm-IFPP), which allows for a wider choice of prior distributions for the unnormalized weights of the mixture. Therefore, the proposed model offers more flexibility, while preserving computational tractability.

The motivating application for the proposed model is the analysis of multiple count data collected from a questionnaire on the frequency of use of the messaging service WhatsApp [31]. In particular, the questionnaire concerns the sharing of COVID-19-related information via WhatsApp messages, either directly or by forwarding, over the course of a week. For each subject, responses to the same seven questions are recorded over seven consecutive days, providing information on a subject's WhatsApp use (see the electronic supplementary material, Table S1). In this set-up, the multiple count processes correspond to the seven questions, all of which display an excess of zeros (see the electronic supplementary material, Figure S2).

The manuscript is organized as follows. Section 2 introduces a novel enriched prior process for multiple zero-inflated outcomes, while §3 describes the Markov chain Monte Carlo (MCMC) algorithm designed for posterior inference. We demonstrate the model on the WhatsApp application in §4. We conclude the paper in §5.

2. The model

(a) Likelihood

Let Y_{ij} be the count of subject $i = 1, \ldots, n$ for outcome $j = 1, \ldots, d$ and let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})$ be the d-dimensional vector of observations for subject i. To take into account the zero-inflated nature of the data, we assume for each outcome j a hurdle model. Each observed count Y_{ij} is equal to zero with probability $1 - p_{ij}$, while with probability p_{ij} it is distributed according to a probability mass function (pmf) $g(\cdot \mid \boldsymbol{\mu}_{ij})$ with support on $\mathbb{N}$. Assuming conditional independence among responses, the likelihood for a subject is given by

$$\mathbb{P}(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{p}_i, \boldsymbol{\mu}_i) = \prod_{j=1}^{d} f(y_{ij} \mid p_{ij}, \boldsymbol{\mu}_{ij}) \quad f(y \mid p, \boldsymbol{\mu}) = \begin{cases} 1 - p, & y = 0 \\ p\, g(y \mid \boldsymbol{\mu}), & y > 0 \end{cases} \tag{2.1}$$

with $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{id}) \in (0,1)^d$, $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}, \ldots, \boldsymbol{\mu}_{id})$, $i = 1, \ldots, n$. In what follows, we set g to be a shifted negative binomial distribution with parameters $\boldsymbol{\mu}_{ij} = (r_{ij}, \theta_{ij})$ and pmf:

$$g(y \mid r_{ij}, \theta_{ij}) = \frac{(y + r_{ij} - 2)!}{(r_{ij} - 1)!(y - 1)!} \theta_{ij}^{y-1} (1 - \theta_{ij})^{r_{ij}}, \quad y \in \mathbb{N}, \tag{2.2}$$

where $r_{ij} \in \mathbb{N}$ and $\theta_{ij} \in (0,1)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Different parametric choices for g are possible (e.g. a shifted Poisson), or even non-parametric alternatives could be employed. Note that the conditional independence assumption among the multiple processes leads to a significant reduction in the number of parameters as compared with multivariate zero-inflated models.

(b) Enriched finite mixture model

In this work, we propose an *enriched* extension of the Norm-IFPP of Argiento & De Iorio [19] and specify a joint prior for $(\boldsymbol{p}_i, \boldsymbol{\mu}_i)$ as conditionally dependent processes. This allows us to account for interindividual heterogeneity, overdispersion and outliers and induces data-driven nested clustering of the observations. Each subject is first assigned to an *outer* cluster, and then clustered again at an *inner* level, providing increased interpretability. Differently from previous work on Bayesian non-parametric enriched processes, we opt for a finite mixture with a random number of components, where the weights are obtained through the normalization of a finite point process. Finite mixture models with a random number of components have received increasing attention

Algorithm 1. Conditional algorithm.

Input: $(y_{ij})_{ij}$ and parameter initialization

Output: posterior distribution of cluster allocation and other parameters

for i in 1:n **do**
Sample c_i and z_i from

$$\mathbb{P}[c_i = m, z_i = s \mid \text{rest}] \propto \Gamma_m \, \Delta_{ms} \prod_{j=1}^{d} f\left(y_{ij} \mid \boldsymbol{p}^\star_m, \boldsymbol{r}^\star_{ms}, \boldsymbol{\theta}^\star_{ms}\right)$$

end for
Compute K, the number of allocated components at the outer level
Relabel the outer level clusters so that the first K components of the mixture are allocated
Sample the latent variable $\tilde{u}$ from Gamma $\left(n, \sum_{m=1}^{M} \Gamma_m\right)$
Set $M = K + x$, where

$$x \sim q_x \qquad q_x \propto \frac{(x+K)!}{x!} \psi_{\text{out}}(\tilde{u})^x q_M(K+x) \qquad \text{for } x = 0, 1, \ldots$$

Sample the unnormalized weights of the outer measure from

$$\mathbb{P}[\Gamma_m \in d\omega \mid \text{rest}] \propto \omega^{n_m} \, \mathrm{e}^{-\omega\tilde{u}} h_{\text{out}}(\omega) \, \mathrm{d}\omega \quad \text{for } m = 1, \ldots, M$$

where n_m is the cardinality of outer level cluster m and $n_m = 0$ for $m > K$
for m in 1:K **do**
Sample $\boldsymbol{p}^\star_m$ from the full conditional.
Compute the number K_m of allocated components at the inner level
Relabel the inner level clusters so that the first K_m components are allocated
Sample the latent variable u_m from Gamma $\left(n_m, \sum_{s=1}^{S_m} \Delta_{ms}\right)$
Set $S_m = K_m + x$ where

$$x \sim q_x \; q_x \propto \frac{(x+K_m)!}{x!} \psi_{\text{in}}(u_m)^x q_S(K_m + x) \quad \text{for } x = 0, 1, \ldots$$

Sample the unnormalized weights of the mth inner mixture from

$$\mathbb{P}[\Delta_{ms} \in dq \mid \text{rest}] \propto q^{n_{ms}} \, \mathrm{e}^{-\omega u_m} h_{\text{in}}(q) \, \mathrm{d}q \quad \text{for } s = 1, \ldots, S_m$$

where n_{ms} is the cardinality of inner level cluster s and $n_{ms} = 0$ for $s > K_m$
for s in 1:K_m **do**
Sample $(\boldsymbol{r}^\star_{ms}, \boldsymbol{\theta}^\star_{ms})$ from the full conditional
end for
for s in $(K_m + 1)$:S_m **do**
Sample $\boldsymbol{r}^\star_{ms}$ from the prior
Sample $\theta^\star_{ms}$ from the prior
end for
end for
for m in $(K + 1)$:M **do**
Sample $\boldsymbol{p}^\star_m$ and S_m from the prior
for s in 1:S_m **do**
Sample Δ_{ms} from the prior
Sample $\boldsymbol{r}^\star_{ms}$ from the prior
Sample $\theta^\star_{ms}$ from the prior
end for
end for

in the last years (see, for example, [30,32]). The representation of Argiento & De Iorio [19] allows for the specification of a wide range of distributions for the weights and simultaneously leads to effective and widely applicable MCMC schemes on which algorithms 1 and 2 are based. More specifically, they show that a finite mixture model is equivalent to a realization of a stochastic process with random dimension and infinite-dimensional support, leading to flexible distributions for the weights of the mixture given by the normalization of a finite point process. We thus employ this approach as it allows for efficient computations via a conditional algorithm, as compared with labour-intensive reversible jump algorithms common in mixture models. An alternative efficient conditional sampler for mixtures with a random number of components is the recently proposed telescopic sampler [33].

In the proposed framework, the observations are assumed to be sampled from a mixture with an inner and an outer component. As kernel of the mixture, we assume the hurdle model in (2.1), which distinguishes between the probabilities of being non-zero $\boldsymbol{p}_i$ and the parameters of the

Algorithm 2. Marginal algorithm.

Input: $(y_{ij})_{ij}$ and parameter initialization
Output: posterior distribution of cluster allocation and other posterior summaries

for i in 1:n **do**
Sample c_i

$$\mathbb{P}[c_i = m \mid \boldsymbol{c}^{-(i)}, \boldsymbol{z}^{-(i)}, \bar{U}, U_1, \ldots, U_K]$$

$$\propto \begin{cases} \left(n_m^{-(i)} + \gamma_M\right) \prod_{j=1}^{d} \dfrac{\mathcal{M}_{\text{Bern}}\left(y^*_{j\mathcal{C}_m^{+(i)}}\right)}{\mathcal{M}_{\text{Bern}}\left(y^*_{j\mathcal{C}_m^{-(i)}}\right)} \left(\dfrac{n_{ms}^{-(i)} + \gamma_S}{L_m^{-(i)}} \prod_{j=1}^{d} \dfrac{\mathcal{M}_{\text{NB}}\left(y^*_{j\mathcal{C}_{ms}^{+(i)}}\right)}{\mathcal{M}_{\text{NB}}\left(y^*_{j\mathcal{C}_{ms}^{-(i)}}\right)} \right. \\ \qquad \left. + \dfrac{L_m^{-(i)} - n_m^{-(i)} - \gamma_S}{L_m^{-(i)}} \prod_{j=1}^{d} \mathcal{M}_{\text{NB}}\left(y_{ij}\right) \right) & \text{if } m = m_{\text{old}} \\ \dfrac{\Lambda_M + (K^{-(i)} + 1)(\bar{u} + 1)^{\gamma_M}}{\Lambda_M + K^{-(i)}(\bar{u} + 1)^{\gamma_M}} \dfrac{\Lambda_M \gamma_M}{(\bar{u} + 1)^{\gamma_M}} \prod_{j=1}^{d} \mathcal{M}_{\text{Bern}}\left(y_{ij}\right) \mathcal{M}_{\text{NB}}\left(y_{ij}\right) & \text{otherwise} \end{cases}$$

where $n_m^{-(i)}$ and $n_{ms}^{-(i)}$ are the cardinalities of outer and inner clusters after removing the ith observation, $\mathcal{C}_m^{-(i)} = \mathcal{C}_m \setminus \{i\}$ and $\mathcal{C}_m^{+(i)} = \mathcal{C}_m \cup \{i\}$, and similarly for $\mathcal{C}_{ms}^{+(i)}$, $\mathcal{C}_{ms}^{-(i)}$, $K^{(-i)}$ and $K_m^{-(i)}$. Here the subscript 'old' denotes an existing (occupied) cluster and

$$L_m^{-(i)} = \frac{\Lambda_S + \left(K_m^{-(i)} + 1\right)(u_m + 1)^{\gamma_S}}{\Lambda_S + K_m^{-(i)}(u_m + 1)^{\gamma_S}} \frac{\Lambda_S \gamma_S}{(u_m + 1)^{\gamma_S}} + n_m^{-(i)} + \gamma_S$$

Sample z_i

$$\mathbb{P}[z_i = s \mid \boldsymbol{c}, \boldsymbol{z}^{-(i)}, U_m]$$

$$\propto \begin{cases} (n_{ms}^{-(i)} + \gamma_S) \prod_{j=1}^{d} \dfrac{\mathcal{M}_{\text{NB}}\left(y^*_{j\mathcal{C}_{ms}^{+(i)}}\right)}{\mathcal{M}_{\text{NB}}\left(y^*_{j\mathcal{C}_{ms}^{-(i)}}\right)} & \text{if } s = s_{\text{old}} \\ \dfrac{\Lambda_S + \left(K_m^{-(i)} + 1\right)(u_m + 1)^{\gamma_S}}{\Lambda_S + K_m^{-(i)}(u_m + 1)^{\gamma_S}} \dfrac{\Lambda_S \gamma_S}{(u_m + 1)^{\gamma_S}} \prod_{j=1}^{d} \mathcal{M}_{\text{NB}}\left(y_{ij}\right) & \text{otherwise} \end{cases}$$

Note that when a subject i is assigned to a new outer cluster, then the full conditional distribution of z_i is degenerate and a new auxiliary variable U_m has to be sampled before moving to the next subject $i + 1$.
end for
Sample the latent variables $\bar{U}$ and $U_1, \ldots, U_K$ from their full conditional:

$$\mathbb{P}\left[\bar{U} = \bar{u} \mid \text{rest}\right] \propto \left(\frac{\Lambda_M}{(\bar{u} + 1)^{\gamma_M}} + K\right) \exp\left\{\frac{\Lambda_M}{(\bar{u} + 1)^{\gamma_M}}\right\} \frac{\bar{u}^{n-1}}{(\bar{u} + 1)^{n + K\gamma_M}}, \quad \bar{u} > 0$$

$$\mathbb{P}\left[U_m = u_m \mid \text{rest}\right] \propto \left(\frac{\Lambda_S}{(u_m + 1)^{\gamma_S}} + K_m\right) \exp\left\{\frac{\Lambda_s}{(u_m + 1)^{\gamma_S}}\right\} \frac{(u_m)^{n_m - 1}}{(u_m + 1)^{n_m + K_m \gamma_S}}, \quad u_m > 0$$

sampling distribution $(\boldsymbol{r}_i, \boldsymbol{\theta}_i)$. The components of the outer mixture are determined by different probabilities of non-zero outcomes, denoted with $\boldsymbol{p}^{\star}_{\boldsymbol{m}} = (p^{\star}_{m1}, \ldots, p^{\star}_{md})$, for $m = 1, \ldots, M$, with M the number of outer mixture components. The components of the inner mixtures are characterized by distinct parameters of the sampling distribution, denoted with $\boldsymbol{r}^{\star}_{\boldsymbol{ms}} = (r^{\star}_{ms1}, \ldots, r^{\star}_{msd})$ and $\boldsymbol{\theta}^{\star}_{ms} = (\theta^{\star}_{ms1}, \ldots, \theta^{\star}_{msd})$, for $s = 1, \ldots, S_m$ and $m = 1, \ldots, M$, where S_m is the number of mixture

components within the mth outer mixture component. Letting $\boldsymbol{\psi}^{\star}_{msj} = (p^{\star}_{mj}, r^{\star}_{msj}, \theta^{\star}_{msj})$ and $\boldsymbol{\psi}^{\star}_{ms} = (\boldsymbol{\psi}^{\star}_{ms1}, \ldots, \boldsymbol{\psi}^{\star}_{msd})$, the mixture model is as follows:

$$\begin{aligned}
\boldsymbol{Y}_i \mid \{\boldsymbol{\psi}^{\star}_{ms}\}, \boldsymbol{w}, \{\boldsymbol{q}_m\} &\overset{\text{iid}}{\sim} \underbrace{\sum_{m=1}^{M} w_m}_{\text{outer level}} \quad \underbrace{\sum_{s=1}^{S_m} q_{ms}}_{\text{inner level}} \prod_{j=1}^{d} f(y_{ij} \mid \boldsymbol{\psi}^{\star}_{msj}) \\
\boldsymbol{q}_m = (q_{m1}, \ldots, q_{mS_m}) \mid S_m &\sim \text{Dirichlet}_{S_m}(\gamma_S, \ldots, \gamma_S) \\
\boldsymbol{w} = (w_1, \ldots, w_M) \mid M &\sim \text{Dirichlet}_M(\gamma_M, \ldots, \gamma_M) \\
\boldsymbol{p}^{\star}_m &\overset{\text{iid}}{\sim} \prod_{j=1}^{d} \text{Beta}(\alpha, \beta) \\
\boldsymbol{r}^{\star}_{ms} &\overset{\text{iid}}{\sim} \prod_{j=1}^{d} \text{Geometric}(\zeta) \\
\boldsymbol{\theta}^{\star}_{ms} &\overset{\text{iid}}{\sim} \prod_{j=1}^{d} \text{Beta}(\eta, \lambda) \\
S_1, \ldots, S_M \mid M &\overset{\text{iid}}{\sim} \text{Poi}_0(\Lambda_S) \\
M &\sim \text{Poi}_0(\Lambda_M)
\end{aligned} \tag{2.3}$$

where the kernel $f(y_{ij} \mid \boldsymbol{\psi}^{\star}_{msj})$ is defined via conditionally independent hurdle models in (2.1)–(2.2). Here $\text{Dirichlet}_M(\gamma_M, \ldots, \gamma_M)$ denotes the symmetric Dirichlet distribution defined on the $(M-1)$-dimensional simplex with mean $1/M$, which is the distribution of the normalized mixture weights. $\text{Beta}(\alpha, \beta)$ indicates the Beta distribution with mean $\alpha/(\alpha+\beta)$ and variance $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$, $\text{Geometric}(\zeta)$ the Geometric distribution with mean $1/\zeta$, and $\text{Poi}_0(\Lambda)$ the shifted Poisson distribution, such that if $X \sim \text{Poi}_0(\Lambda)$ then $X-1$ has a Poisson distribution with mean Λ. Moreover, M and S_m, for $m = 1, \ldots, M$, indicate the random number of components at the outer and inner level of the enriched Norm-IFPP, respectively.

The outer mixture is a mixture of multivariate Bernoulli distributions, and coincides with the widely used latent class model [34]. Moreover, being conditionally independent of the actual values of the non-zero observations, it offers further computation advantages as shown in §3.

Model (2.3) induces a partition of the subject indices $\{1, \ldots, n\}$ at an outer and an inner level. Let c_i and z_i, for $i = 1, \ldots, n$, denote the allocation variables which indicate to which component of the mixture each subject is assigned to at the outer and inner level, respectively. When two subjects, i and l, are assigned to the same component of the outer level mixture, then the probabilities of observing a zero for the two subjects are the same, $\boldsymbol{p}_i = \boldsymbol{p}_l$, and the two subjects are assigned to the same cluster, i.e. $c_i = c_l$. Moreover, if the two subjects are also assigned to the same component of the inner level mixture, we have $z_i = z_l$ and $\boldsymbol{\mu}_i = \boldsymbol{\mu}_l$ (with obviously $c_i = c_l$). However, the vectors of parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_l$ characterizing the sampling distribution might be different even when $c_i = c_l$ and, consequently, the two subjects might be assigned to different clusters at the inner level. This is reflected in the components of the vectors of parameters $(\boldsymbol{p}_i, \boldsymbol{\mu}_i)$ and $(\boldsymbol{p}_l, \boldsymbol{\mu}_l)$, which might share only the component corresponding to the probability of zero outcomes or both components.

Using allocation variables, the conditional dependence structure between outer and inner levels is the following. Let

$$\widetilde{Y}_{ij} = \begin{cases} 1 & \text{if } Y_{ij} > 0 \\ 0 & \text{if } Y_{ij} = 0 \end{cases}, \tag{2.4}$$

$\widetilde{\boldsymbol{Y}}_i = (\widetilde{Y}_{i1}, \ldots, \widetilde{Y}_{id})$, $\mathcal{C}_m = \{i : c_i = m\}$ and $\mathcal{C}_{ms} = \{i : c_i = m, z_i = s\}$.

Outer mixture:

$$\begin{aligned}
\widetilde{\boldsymbol{Y}}_i \mid \boldsymbol{p}_i &\sim \prod_{j=1}^{d} p_{ij}^{\widetilde{y}_{ij}} (1-p_{ij})^{\widetilde{y}_{ij}}, \quad \widetilde{y}_{ij} \in \{0,1\} \\
\boldsymbol{p}_i &= \boldsymbol{p}^{\star}_{c_i} \\
\boldsymbol{p}^{\star}_1, \ldots, \boldsymbol{p}^{\star}_M \mid M &\overset{\text{iid}}{\sim} \prod_{j=1}^{d} \text{Beta}(\alpha, \beta) \\
\Pr(c_i = m) &\propto \Gamma_m, \quad m = 1, \ldots, M \\
\Gamma_1, \ldots, \Gamma_M &\overset{\text{iid}}{\sim} \text{Gamma}(\gamma_M, 1) \\
M &\sim \text{Poi}_0(\Lambda_M)
\end{aligned} \tag{2.5}$$

Inner mixture:

$$\begin{aligned}
\boldsymbol{Y}_i \mid M, c_i = m, \boldsymbol{p}^{\star}_m, \boldsymbol{r}_{mi}, \boldsymbol{\theta}_{mi} &\sim \prod_{j=1}^{d} f(y_{ij} \mid p^{\star}_{mj}, r_{mij}, \theta_{mij}) \\
(\boldsymbol{r}_{mi}, \boldsymbol{\theta}_{mi}) &= (\boldsymbol{r}^{\star}_{mz_i}, \boldsymbol{\theta}^{\star}_{mz_i}) \\
\boldsymbol{r}^{\star}_{m1}, \ldots, \boldsymbol{r}^{\star}_{mS_m} \mid S_m &\overset{\text{iid}}{\sim} \prod_{j=1}^{d} \text{Geometric}(\zeta) \\
\boldsymbol{\theta}^{\star}_{m1}, \ldots, \boldsymbol{\theta}^{\star}_{mS_m} \mid S_m &\overset{\text{iid}}{\sim} \prod_{j=1}^{d} \text{Beta}(\eta, \lambda) \\
\Pr(z_i = s \mid c_i = m) &\propto \Delta_{ms}, \quad i \in \mathcal{C}_m, \quad s = 1, \ldots, S_m \\
\Delta_{m1}, \ldots, \Delta_{mS_m} &\overset{\text{iid}}{\sim} \text{Gamma}(\gamma_S, 1) \\
S_1, \ldots, S_M \mid M &\overset{\text{iid}}{\sim} \text{Poi}_0(\Lambda_S),
\end{aligned} \tag{2.6}$$

where, as before, we denote with $\boldsymbol{p_m}^{\star}$, $\boldsymbol{r_{ms}}^{\star}$ and $\boldsymbol{\theta_{ms}}^{\star}$ the component-specific parameters, which are assumed *a priori* independent and Gamma(α, β) is the Gamma distribution with mean α/β and variance α/β^2. The choice of Gamma distribution for the unnormalized weight of the mixture leads to the standard Dirichlet distribution for the normalized weights. In this setting, the computations are greatly simplified by the introduction of a latent variable, conditionally on which the unnormalized weights are independent. See Argiento & De Iorio [19] for details. Note that the inner mixture is here defined conditionally on the probabilities $p_{m,j}$ of being zero and not on $\widetilde{\boldsymbol{Y}}_i$. Thus, while conditioning on $p_{m,j}$, $\boldsymbol{Y}_i$ is still allowed to present zero entries. Finally, we highlight that representations (2.3) and (2.5)–(2.6) are equivalent.

3. Inference

Posterior inference can be performed through both a conditional and a marginal algorithm, derived by extending the algorithms by Argiento & De Iorio [19] to the enriched set-up. The conditional algorithm is described in algorithm 1, while in algorithm 2 we present the marginal one.

The conditional algorithm is very flexible and allows for different prior distributions on the weights of the two mixtures as well as on M and S_m (see [19] for details). In algorithm 2, we use the notation q_M and q_S to denote the prior on M and S_m, respectively, and we set them both

equal to a shifted Poisson for the application in §4. Furthermore, h_{out} and h_{in} denote the prior distribution on the unnormalized weights (in our case Gamma distributions) of the outer and inner mixture, respectively, $\psi_{\text{out}}(u)$ and $\psi_{\text{in}}(u)$ denote the corresponding Laplace transforms of h_{out} and h_{in} (in our case $\psi_{\text{out}}(u) = (u+1)^{-\gamma_M}$ and $\psi_{\text{in}}(u) = (u+1)^{-\gamma_S}$).

To implement the marginal algorithm, we need to derive the marginal likelihood of the data, conditionally on cluster membership. The likelihood in equation (2.3) can be written as

$$\prod_{i=1}^{n}\prod_{j=1}^{d}\left\{(1-p_{ij})^{1-\widetilde{y}_{ij}}\, p_{ij}^{\widetilde{y}_{ij}}\left\{\frac{(y_{ij}+r_{ij}-2)!}{(r_{ij}-1)!(y_{ij}-1)!}\theta_{ij}^{y_{ij}-1}(1-\theta_{ij})^{r_{ij}}\right\}^{\widetilde{y}_{ij}}\right\}. \tag{3.1}$$

Recall that c_i and z_i denote the labels of the clusters to which the ith subject belongs to in the outer and the inner clustering, respectively. The marginal likelihood of the data conditionally on the cluster allocation is obtained marginalizing with respect to the prior distributions defined in (2.5) and (2.6). For a vector of counts y, we obtain:

$$\mathcal{M}(y \mid \boldsymbol{c}, \boldsymbol{z}) = \prod_{j=1}^{d}\left\{\prod_{m=1}^{K}\left\{\mathcal{M}_{\text{Bern}}(y^*_{j\mathcal{C}_m})\prod_{s=1}^{K_m}\mathcal{M}_{\text{NB}}(y^*_{j\mathcal{C}_{ms}})\right\}\right\}$$

$$\mathcal{M}_{\text{Bern}}(y) = \frac{B(\alpha+n^1, \beta+n^0)}{B(\alpha,\beta)}$$

and

$$\mathcal{M}_{\text{NB}}(y) = \sum_{r=1}^{+\infty}\left\{\frac{B\left(\eta+\sum_i(y_i-1)\widetilde{y}_i, \lambda + r\sum_i \widetilde{y}_i\right)}{B(\eta,\lambda)}\prod_i\left(\frac{(y_i+r-2)!}{(r-1)!(y_i-1)!}\right)^{\widetilde{y}_i}(1-\zeta)^{r-1}\zeta\right\}$$

where $\mathcal{C}_m = \{i : c_i = m\}$, $\mathcal{C}_{ms} = \{i : c_i = m, z_i = s\}$, $y^*_{j\mathcal{C}_m}$ is the vector of observations y_{ij} such that $c_i = m$, for $j = 1, \ldots, d$. Similarly, $y^*_{j\mathcal{C}_{ms}}$ is the vector of observations y_{ij} such that $c_i = m$ and $z_i = s$. Moreover, $B(\cdot,\cdot)$ denotes the Beta function, $n^1 = \sum_i \widetilde{y}_i$, $n^0 = \sum_i(1-\widetilde{y}_i)$, $\widetilde{y}_i$ is defined as in equation (2.4) and the last two summations run over the elements of the vector $\widetilde{y}$. Here K and K_m are the numbers of clusters at the outer and inner level, respectively. Note that by cluster we mean an occupied component (i.e. a mixture component to which at least one observation has been assigned), with $K \leq M$ and $K_m \leq S_m, m = 1, \ldots, M$.

When implementing the marginal algorithm, after updating the latent variables $\bar{U}$ and U_m, we could add an extra step involving a shuffle of the nested partition structure as suggested by Wade *et al.* [25] to improve mixing. More details and an empirical comparison of the two algorithms are provided in Section S3 of the electronic supplementary material.

4. Application to WhatsApp use during COVID-19

(a) Data description and preprocessing

We apply our model to a dataset on WhatsApp use during COVID-19 [31]. The data consist of a questionnaire filled out by participants living in India. Each subject answers the same $d = 7$ questions for $T = 7$ consecutive days on the number of ($j = 1$) COVID-19 messages forwarded, ($j = 2$) WhatsApp groups to which COVID-19 messages were forwarded, ($j = 3$) people to whom COVID-19 messages were forwarded, ($j = 4$) unique forwarded messages received in personal chats, ($j = 5$) people from whom forwarded messages were received, ($j = 6$) personal chats that discussed COVID-19, ($j = 7$) WhatsApp groups that mentioned COVID-19. Table S1 in the electronic supplementary material provides the list of the questions, as well as a brief description. In what follows, the first replicate ($t = 1$) corresponds to Sunday for all subjects, $t = 2$ to Monday, up to $T = 7$ corresponding to Saturday. The questionnaire responses were collected in June and July 2021, during India's infection wave of the Delta variant of the SARS-CoV-2 virus that causes coronavirus disease 2019 (COVID-19).

From the initial 1156 respondents, we remove two subjects for which no answers are available, resulting in a final sample size of $n = 1154$. Moreover, 19% of the observations are missing. We also

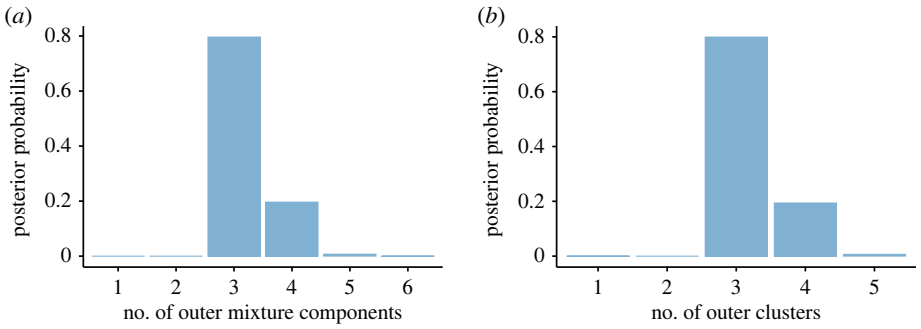

Figure 2. Posterior distribution of the number of outer mixture components M (*a*) and clusters K, i.e. number of occupied components to which at least one observation is assigned (*b*). (Online version in colour.)

treat counts higher than 400, which are very rare (seven observations out of 56 546), as missing data as they are very far from the range of the majority of the data. We handle missing data using a two-step procedure. Firstly, whenever possible, we recover missing zeros using deterministic imputation based on respondents' answers to other sections of the questionnaire. For instance, if the answer to the question 'Did you send any message of this kind today?' is 'No' and there is a missing value for the question 'How many?', we can reasonably assume that the answer to the latter question is zero. In this way, we can recover 0.5% of the missing observations. Secondly, the remaining missing values are imputed using random forest imputation (as implemented in the `R` package `mice` [35]). In Section S2 of the electronic supplementary material, we provide more details on the data imputation technique and we present an empirical study to quantify the impact of data imputation on the results presented in the next section. Figure S2 of the electronic supplementary material displays the data after imputation.

To account for the fact that T repeated observations are available for each subject and process, we need to slightly modify model (2.3). We do so by assuming that the different time points are independent of each other, so that repeated observations can be straightforwardly included in the proposed model. Let Y_{ijt} denote the count for the ith subject and the jth process at time t, $i = 1, \ldots, n$, $j = 1, \ldots, d$ and $t = 1, \ldots, T$. We assume that Y_{ijt} are conditionally independent, given the parameters of the model. Thus, the likelihood contribution of each subject i is given by $\prod_{t=1}^{T} \prod_{j=1}^{d} f(y_{ijt} \mid \boldsymbol{\psi}^{\star}_{msj})$. It must be highlighted that we are clustering individuals based on the pattern of all their observations, at each time point t and for each process j.

Finally we note that, thanks to the probabilistic structure of the hurdle model for zero-inflated data, $\boldsymbol{p}_i$ and the sampling distribution $g(\cdot \mid \boldsymbol{\mu}_i)$ reflect two distinct features of the respondents' behaviour: $\boldsymbol{p}_i$ represents the probability of engaging in some COVID-19 related WhatsApp activity, while $g(\cdot \mid \boldsymbol{\mu}_i)$ captures the behaviour of those subjects who have actually engaged in the activity.

(b) Results

Posterior inference is performed through the conditional algorithm described in algorithm 1. We run the algorithm for 15 000 MCMC iterations, discarding the first 5000 as burn-in.

Figure 2 shows that, at the outer level, the posterior distributions of the number of both components and clusters present a mode at the value three.

As point estimate of the cluster allocation, we report the configuration that minimizes the posterior expectation of Binder's loss function [36] under equal misclassification costs, which is a common choice in the applied Bayesian non-parametrics literature [37]. Briefly, this expectation of the loss measures the difference for all possible pairs of subjects between the posterior probability of co-clustering and the estimated cluster allocation. We refer to the resulting cluster allocation as the Binder estimate.

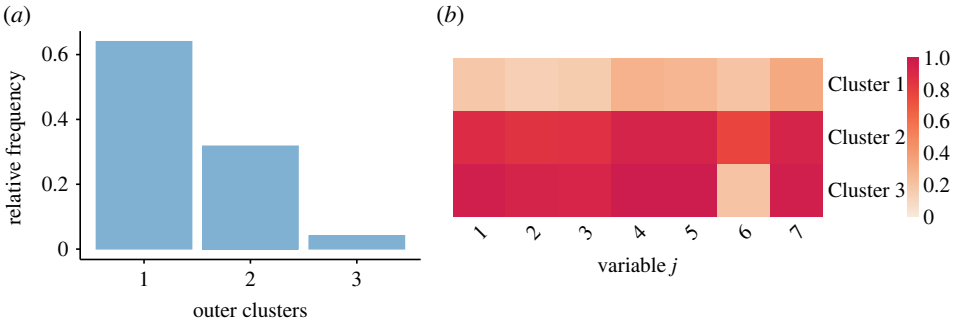

Figure 3. Relative frequency of the outer clusters (*a*) and the posterior means of the cluster-specific probabilities of a non-zero count $p^{\star}_{mj}$ (*b*) corresponding to the posterior estimate of the clustering allocation obtained by minimizing Binder's loss function. (*a*) Outer clusters relative frequencies and (*b*) Outer level Bernoulli parameters $p^{\star}_{mj}$. (Online version in colour.)

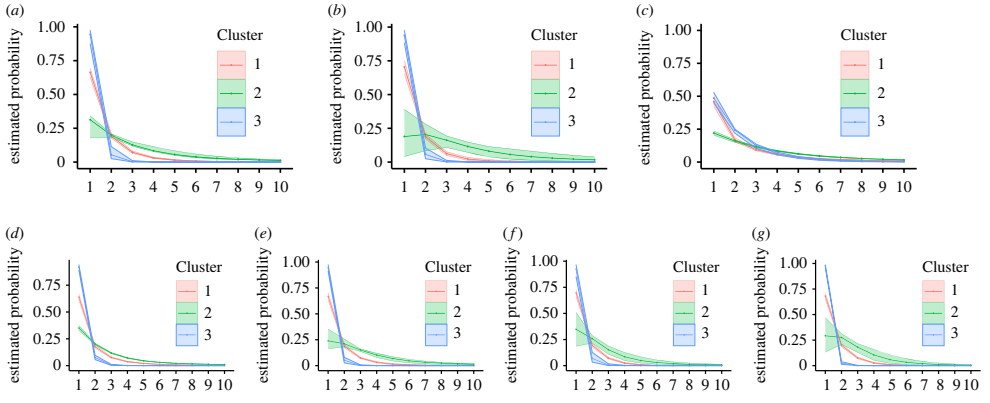

Figure 4. Estimated pmfs for the seven questions within each outer cluster (conditionally on the counts being positive) corresponding to the posterior estimate of the clustering allocation obtained by minimizing Binder's loss function. Shaded areas represent the 95% credible intervals. (*a*) Question 1, (*b*) Question 2, (*c*) Question 3, (*d*) Question 4, (*e*) Question 5, (*f*) Question 6 and (*g*) Question 7. (Online version in colour.)

The Binder estimate of the outer clustering contains three clusters, whose characteristics are summarized in figures 3 and 4. The largest cluster corresponds to WhatsApp users who on most days report a zero count for all $d = 7$ questions. The individuals in the other two clusters use WhatsApp more frequently when it comes to forwarding COVID-19 messages ($j = 1, 2$), receiving forwarded messages ($j = 3, 4, 5$) and having COVID-19 mentioned in their WhatsApp groups ($j = 7$). The main feature distinguishing Cluster 2 from Cluster 3 in terms of probabilities $\boldsymbol{p}_i$ of non-zero counts is that on most days Cluster 2, unlike Cluster 3, discusses COVID-19 also in personal chats (question $j = 6$).

Figures 5 and 6 display the main characteristics of the inner clusters. We are interested in the posterior distribution of the number of the inner clusters per outer cluster, as well as the inner clustering within each outer cluster. To this end, we run the MCMC algorithm fixing the outer cluster allocation to its Binder estimate, thus obtaining the conditional posterior distribution of the inner clustering. The results reveal substantial variability in the distribution of non-zero counts within outer Clusters 1 and 2 (see figure 5*c*). The majority of counts in outer Cluster 1 are zero, leaving little variation in the counts for the inner clustering. As most individuals present zero counts (for most processes) at an inner cluster level, it becomes difficult to detect specific patterns as it is also evident from the fact that many co-clustering probabilities are in the range 0.3–0.6

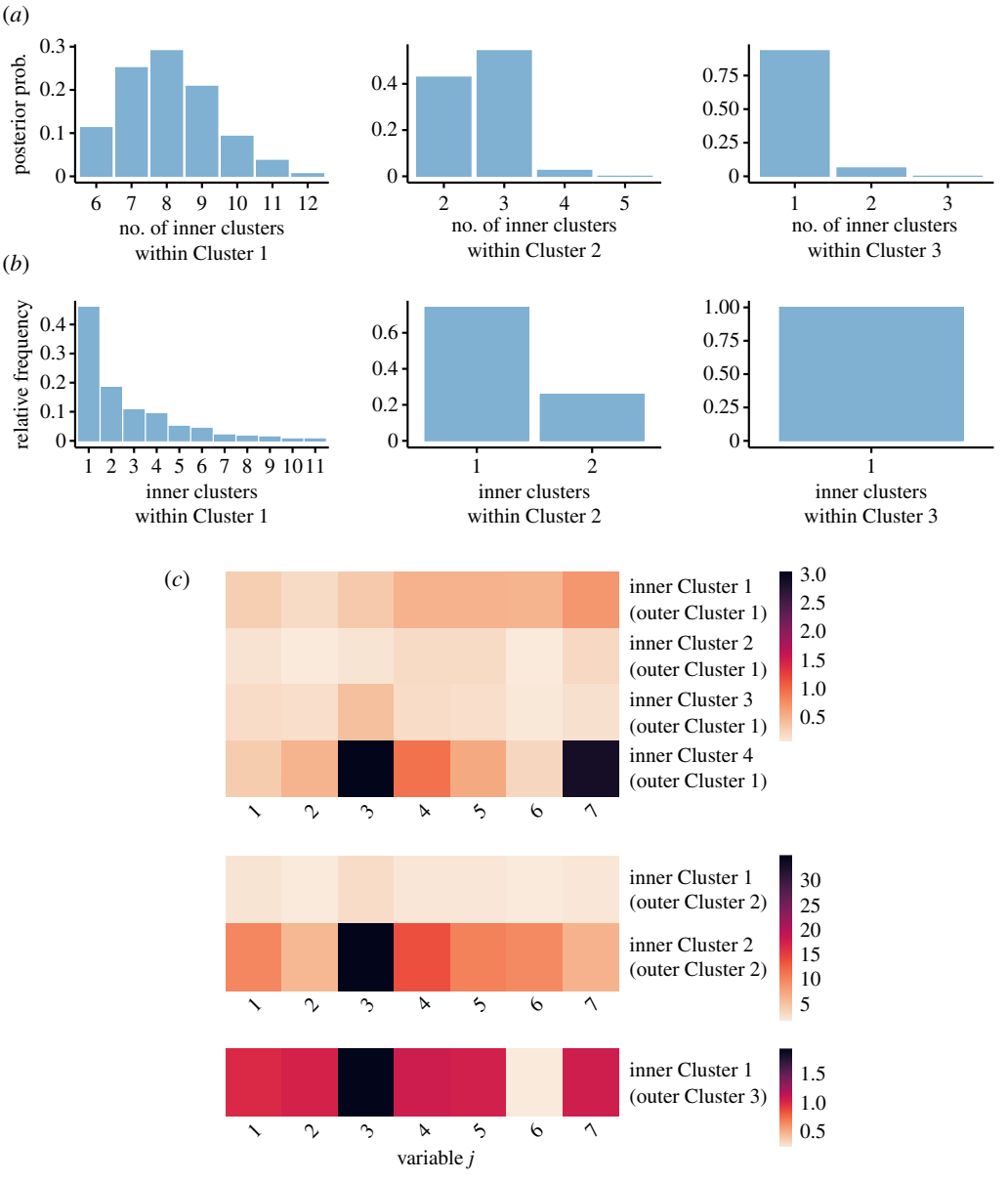

Figure 5. Posterior distribution of the number of inner clusters per outer cluster (*a*), relative frequency of the inner clusters corresponding to the Binder estimate of the inner cluster allocation (*b*), cluster-specific empirical means of the counts (*c*). For outer Cluster 1, the latter is only shown for the four largest inner clusters for visualization purposes. Results are obtained conditionally on the Binder estimate of the outer clustering. (Online version in colour.)

(see figure 6). Notably, around a quarter of the individuals in outer Cluster 2, as captured by its inner Cluster 2, forward COVID-19 messages to many more people (question $j = 3$) than subjects in inner Cluster 1 of outer Cluster 2. Figure 4 also supports the fact that outer Cluster 2 engages with WhatsApp in a much more persistent manner than the other outer clusters. These results highlight that a sizeable minority of WhatsApp users has a relatively large propensity to spread COVID-19 messages during a critical phase of the pandemic. This is in line with a similar survey in Singapore [38] and findings on 'superspreaders' on other social media.

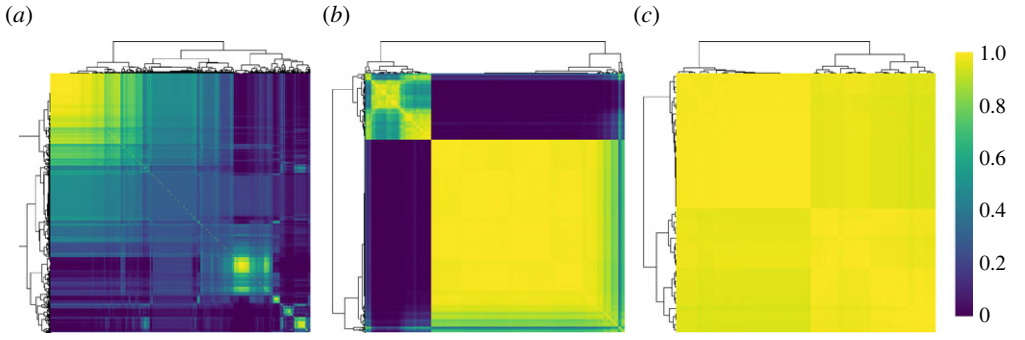

Figure 6. Heatmaps of the posterior co-clustering probabilities for the inner clusters per outer cluster. Results are obtained conditionally on the Binder estimate of the outer cluster allocation. Observations are reordered based on the co-clustering probability profiles, through hierarchical clustering. (*a*) Outer Cluster 1, (*b*) outer Cluster 2 and (*c*) outer Cluster 3. (Online version in colour.)

5. Conclusion

In this work, we propose a Bayesian model for multiple zero-inflated count data, building on the well-established hurdle model and exploiting the flexibility of finite mixture models with a random number of components. The main contribution of this work is the construction of an *enriched* finite mixture with a random number of components, which allows for two-level (nested) clustering of the subjects based on their pattern of counts across different processes. This structure enhances interpretability of the results and has the potential to better capture important features of the data. We design a conditional and a marginal MCMC sampling scheme to perform posterior inference. The proposed methodology has wide applicability, since excess-of-zeros count data arise in many fields. Our motivating application involves answers to a questionnaire on the use of WhatsApp in India during the COVID-19 pandemic. Our analysis identifies a two-level clustering of the subjects: the outer cluster allocation reflects daily probabilities of engaging in different WhatsApp activities, while the inner level informs on the number of messages conditionally on the fact that the subject is indeed receiving/sending messages on WhatsApp. Any two subjects are clustered together if they show a similar pattern across the multiple responses. We find three different well-distinguished respondent behaviours corresponding to the three outer clusters: (i) subjects with low probability of daily utilization; (ii) subjects with high probability of sending/receiving all types of messages and (iii) subjects with high probability for all considered messages except for non-forwarded messages in personal chats. Interestingly, the inner level clustering and the outer cluster-specific estimates of the sampling distribution g highlight similarities between the outer Clusters 1 and 3, where subjects tend to send/receive fewer messages compared with outer Cluster 2. Moreover, we are able to identify those subjects with a high propensity to spread COVID-19 messages during the critical phase of the pandemic and for these subjects we do not find notable differences in terms of types of messages sent or received. Our results are in line with existing literature on the topic. Future work involves the development of more complex clustering hierarchies and techniques able to identify processes that most inform the clustering structure.

Data accessibility. Due to ethical and regulatory constraints on sharing human subject data, the dataset is not available.

Authors' contributions. B.F.: methodology; A.C.: methodology; W.v.d.B.: data curation, writing—original draft; M.D.I.: conceptualization, writing—review and editing. All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was partially supported by the NUS Centre for Trusted Internet and Community (grant no. CTIC-RP-20-09).

Acknowledgements. We thank Dr Jean Liu and the Synergy Lab at Yale-NUS College for providing the data.

References

1. Lambert D. 1992 Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14. (doi:10.2307/1269547)
2. Mullahy J. 1986 Specification and testing of some modified count data models. *J. Econometr.* **33**, 341–365. (doi:10.1016/0304-4076(86)90002-3)
3. Heilbron DC. 1994 Zero-altered and other regression models for count data with added zeros. *Biom. J.* **36**, 531–547. (doi:10.1002/bimj.4710360505)
4. Min Y, Agresti A. 2005 Random effect models for repeated measures of zero-inflated count data. *Stat. Model.* **5**, 1–19. (doi:10.1191/1471082X05st084oa)
5. Li CS, Lu JC, Park J, Kim K, Brinkley PA, Peterson JP. 1999 Multivariate zero-inflated Poisson models and their applications. *Technometrics* **41**, 29–38. (doi:10.1080/00401706.1999.10485593)
6. Liu Y, Tian GL. 2015 Type I multivariate zero-inflated Poisson distribution with applications. *Comput. Stat. Data Anal.* **83**, 200–222. (doi:10.1016/j.csda.2014.10.010)
7. Liu Y, Tian GL, Tang ML, Yuen KC. 2019 A new multivariate zero-adjusted Poisson model with applications to biomedicine. *Biom. J.* **61**, 1340–1370. (doi:10.1002/bimj.201700144)
8. Tian GL, Liu Y, Tang ML, Jiang X. 2018 Type I multivariate zero-truncated/adjusted Poisson distributions with applications. *J. Comput. Appl. Math.* **344**, 132–153. (doi:10.1016/j.cam.2018.05.014)
9. Fox JP. 2013 Multivariate zero-inflated modeling with latent predictors: modeling feedback behavior. *Comput. Stat. Data Anal.* **68**, 361–374. (doi:10.1016/j.csda.2013.07.003)
10. Lee KH, Coull BA, Moscicki AB, Paster BJ, Starr JR. 2020 Bayesian variable selection for multivariate zero-inflated models: application to microbiome count data. *Biostatistics* **21**, 499–517. (doi:10.1093/biostatistics/kxy067)
11. Chib S, Greenberg E. 1998 Analysis of multivariate probit models. *Biometrika* **85**, 347–361. (doi:10.1093/biomet/85.2.347)
12. García-Zattera MJ, Jara A, Lesaffre E, Declerck D. 2007 Conditional independence of multivariate binary data with an application in caries research. *Comput. Stat. Data Anal.* **51**, 3223–3234. (doi:10.1016/j.csda.2006.11.021)
13. Choo-Wosoba H, Gaskins J, Levy S, Datta S. 2018 A Bayesian approach for analyzing zero-inflated clustered count data with dispersion. *Stat. Med.* **37**, 801–812. (doi:10.1002/sim.7541)
14. Li Q, Guindani M, Reich BJ, Bondell HD, Vannucci M. 2017 A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Stat. Anal. Data Min. ASA Data Sci. J.* **10**, 393–409. (doi:10.1002/sam.11350)
15. Hu G, Yang HC, Xue Y, Dey DK. 2022 Zero-inflated Poisson model with clustered regression coefficients: application to heterogeneity learning of field goal attempts of professional basketball players. *Can. J. Stat.* (doi:10.1002/cjs.11684)
16. Shuler K, Verbanic S, Chen IA, Lee J. 2021 A bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study. *J. R. Stat. Soc. C Appl. Stat.* **70**, 961–979. (doi:10.1111/rssc.12493)
17. MacEachern SN. 1999 Dependent nonparametric processes. In *ASA Proc. of the Section on Bayesian Statistical Science, Baltimore, MD, August 8–12*, vol. 1, pp. 50–55.
18. Arab A, Holan SH, Wikle CK, Wildhaber ML. 2012 Semiparametric bivariate zero-inflated Poisson models with application to studies of abundance for multiple species. *Environmetrics* **23**, 183–196. (doi:10.1002/env.1142)
19. Argiento R, De Iorio M. 2022 Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *Ann. Stat.* **50,** 2641–2663. (doi:10.1214/22-AOS2201)
20. Wade S, Mongelluzzo S, Petrone S. 2011 An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Anal.* **6**, 359–385. (doi:10.1214/ba/1339616468)
21. Consonni G, Veronese P. 2001 Conditionally reducible natural exponential families and enriched conjugate priors. *Scand. J. Stat.* **28**, 377–406. (doi:10.1111/1467-9469.00243)

22. Consonni G, Veronese P, Gutiérrez-Pena E. 2004 Reference priors for exponential families with simple quadratic variance function. *J. Multivariate Anal.* **88**, 335–364. (doi:10.1016/S0047-259X(03)00095-2)
23. Connor RJ, Mosimann JE. 1969 Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Stat. Assoc.* **64**, 194–206. (doi:10.1080/01621459.1969.10500963)
24. Ferguson TS. 1973 A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230. (doi:10.1214/aos/1176342360)
25. Wade S, Dunson DB, Petrone S, Trippa L. 2014 Improving prediction from Dirichlet process mixtures via enrichment. *J. Mach. Learn. Res.* **15**, 1041–1071.
26. Gadd C, Wade S, Boukouvalas A. 2020 Enriched mixtures of generalised Gaussian process experts. In *Proc. of the Twenty Third Int. Conf. on Artificial Intelligence and Statistics* (eds S Chiappa, R Calandra), vol. 108 of *Proceedings of Machine Learning Research, 26–28 August*, pp. 3144–3154. PMLR Online.
27. Zeldow B, Flory J, Stephens-Shields A, Raebel M, Roy JA. 2021 Functional clustering methods for longitudinal data with application to electronic health records. *Stat. Methods Med. Res.* **30**, 655–670. (doi:10.1177/0962280220965630)
28. Roy J, Lum KJ, Zeldow B, Dworkin JD, Re III VL, Daniels MJ. 2018 Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics* **74**, 1193–1202. (doi:10.1111/biom.12875)
29. Rigon T, Scarpa B, Petrone S. 2022 Enriched Pitman-Yor processes. (http://arxiv.org/abs/2003.12200v2)
30. Miller JW, Harrison MT. 2018 Mixture models with a prior on the number of components. *J. Am. Stat. Assoc.* **113**, 340–356. (doi:10.1080/01621459.2016.1255636)
31. ClinicalTrialsgov. 2021 WhatsApp in India during the COVID-19 pandemic. Identifier NCT04918849. U.S. National Library of Medicine. Available from https://clinicaltrials.gov/ct2/show/NCT04918849.
32. Malsiner-Walli G, Frühwirth-Schnatter S, Grün B. 2016 Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26**, 303–324. (doi:10.1007/s11222-014-9500-2)
33. Frühwirth-Schnatter S, Malsiner-Walli G, Grün B. 2021 Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Anal.* **16**, 1279–1307. (doi:10.1214/21-BA1294)
34. Lazarsfeld PF, Henry NW. 1968 *Latent structure analysis*. New York, NY: Houghton Mifflin.
35. van Buuren S, Groothuis-Oudshoorn K. 2011 mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67. (doi:10.18637/jss.v045.i03)
36. Binder DA. 1978 Bayesian cluster analysis. *Biometrika* **65**, 31–38. (doi:10.1093/biomet/65.1.31)
37. Lau JW, Green PJ. 2007 Bayesian model-based clustering procedures. *J. Comput. Graph. Stat.* **16**, 526–558. (doi:10.1198/106186007X238855)
38. Tan EY, Wee RR, Saw YE, Heng KJ, Chin JW, Tong EM, Liu JC. 2021 Tracking private WhatsApp discourse about COVID-19 in Singapore: longitudinal infodemiology study. *J. Med. Internet Res.* **23**, e34218. (doi:10.2196/34218)