Electronic Journal of Statistics
Vol. 10 (2016) 3287–3309
ISSN: 1935-7524
DOI: 10.1214/16-EJS1177

A Bayesian nonparametric model for white blood cells in patients with lower urinary tract symptoms

William Barcella

Department of Statistical Science
University College London
e-mail: william.barcella.13@ucl.ac.uk

Maria De Iorio

Department of Statistical Science
University College London
e-mail: m.deiorio@ucl.ac.uk

Gianluca Baio

Department of Statistical Science
University College London
e-mail: g.baio@ucl.ac.uk

and

James Malone-Lee

Division of Medicine
University College London
e-mail: james.malone-lee@ucl.ac.uk

Abstract: Lower Urinary Tract Symptoms (LUTS) affect a significant proportion of the population and often lead to a reduced quality of life. LUTS overlap across a wide variety of diseases, which makes the diagnostic process extremely complicated. In this work we focus on the relation between LUTS and Urinary Tract Infection (UTI). The latter is detected through the number of White Blood Cells (WBC) in a sample of urine: WBC$\geq$ 1 indicates UTI and high levels may indicate complications. The objective of this work is to provide the clinicians with a tool for supporting the diagnostic process, deepening the available knowledge about LUTS and UTI. We analyze data recording both LUTS profile and WBC count for each patient. We propose to model the WBC using a random partition model in which we specify a prior distribution over the partition of the patients which includes the clustering information contained in the LUTS profile. Then, within each cluster, the WBC counts are assumed to be generated by a zero-inflated Poisson distribution. The results of the predictive distribution allows to identify the symptoms configuration most associated with the presence of UTI as well as with severe infections.

Keywords and phrases: Bayesian nonparametric, zero-inflated Poisson distribution, Dirichlet process mixture model, random partition model, clustering with covariates.

Received December 2015.

1. Introduction

Lower Urinary Tract Symptoms (LUTS) define a group of symptoms that comprises urgency, pain, stress incontinence and voiding problems. They particularly affect elderly population with 40% of the men and 28% of the women with age between 70 and 79 years (Irwin et al. (2006)) suffering from them. This group of symptoms is related to a number of diseases (from neurological pathologies to anxiety and stress) which are not directly identified by disjoint groups of LUTS, making the diagnostic process complicated. Often LUTS indicate the presence of Urinary Tract Infection (UTI), a condition that may lead to chronic problems when not readily diagnosed and consequently require time consuming and expensive treatments.

Given the difficulty in interpreting LUTS, specific exams are commonly employed in order to assess the presence of the infection. The published data show that the best biological indicator of UTI available is pyuria (≥ 1 White Blood Cell count (WBC) μl^{-1}) detected by microscopy of a fresh unspun, unstained specimen of urine (Khasriya et al. (2010); Kupelian et al. (2013)). In the presence of symptoms, any pyuria (≥ 1 WBC μl^{-1}) correlates with other independent inflammatory and microbiological markers distinguishing patients from controls (Khasriya et al. (2010); Kupelian et al. (2013); Gill et al. (2015)). This procedure allows counting the WBC, but on the other side it can only be performed in specific laboratories, requiring time to return the results as well as representing a consistent cost for the health system. Therefore, it is common practice to use dipsticks for examining urine samples, which can reveal the presence of White Blood Cells (WBC) which in turn indicate UTI. Dipsticks can be used by non-specialized clinicians and deliver a result in few instants. However, Khasriya et al. (2010) investigated the diagnostic power of dipstick urinalysis and identified deficiencies. Thus, an infection can be present much earlier than being diagnosed using a dipstick increasing significantly the risk of chronicity.

For all these reasons, it is valuable to study the relation between LUTS and UTI from a statistical point of view, in order to provide tools for assisting the clinicians during the diagnostic process. This is the broad objective of this work.

The starting point of our analysis is a dataset containing information about patients affected by LUTS for which the counts of the WBC from the microanalysis have been recorded together with the symptoms profiles. The latter are vectors of binary indicators which indicate the presence of the symptoms. The WBC counts in the dataset are zero more than 50% of the time, *i.e.* more than half of the patients do not show microscopic evidence of UTI. We thus propose an approach to model the relation between the WBC counts (response) and the LUTS profiles (covariates), which extends nonparametrically the well known class of the zero-inflated distributions (Neelon, O'Malley and Normand (2010)). This class of distributions has been extensively employed in a number of applications: it involves the specification of a parameter that regulates the inflation of the probability for a specific outcome which could not be modeled according to standard distributions.

Specifically, we propose a Bayesian random partition model (Lau and Green (2007)) in which the covariates are used jointly with the response to inform the clustering structure of the observations which has been a priori assumed to follow a Chinese Restaurant Process (Aldous (1985)). For a review about random partition models with covariates see Müller and Quintana (2010). Within each cluster we treat the WBC counts as independent and identically distributed (*iid*) random variables distributed according to a Zero-Inflated Poisson (ZIP) distribution with cluster specific parameters. In this way we assume the covariates to affect the response only through the clustering structure. This latter assumption can be relaxed to allow also the mean of the Poisson component to depend on the covariates. We call the resulting model Bayesian Nonparametric ZIP model (BNP-ZIP).

BNP-ZIP allows to associate different combinations of the covariates with different probabilities of having UTI (*i.e.* WBC$\geq$1) as well as with different levels of severity of UTI. The results of the study highlight the importance of the voiding class of symptoms for both the probability of being diagnosed with UTI and also its level of severity (which increases with the number of WBC in the urine). Differently, the urgency and stress incontinence symptoms have low probability of being associated with UTI when they appear alone or combined. We also believe that the predictive distributions which depend on the covariates may represent a useful tool for supporting the clinicians in the diagnostic process.

The rest of the work is organized as follows. In Section 2 we introduce and discuss the zero-inflated models, while in Section 3 we describe our nonparametric approach. Section 4 presents the analysis of the LUTS dataset. We conclude the paper with a discussion of the results in Section 5.

2. Models with zero-inflated (or deflated) distributions

Count data with out-of-pattern number of zeros are common in numerous real world applications. Modeling such data without accounting for the excess of zeros may lead to biased estimates of the parameters. The common approach to deal with this problem involves the use of mixture models in which a distribution over counts (*e.g.* Poisson distribution, Negative Binomial distribution, etc.) is mixed with a Dirac measure located in correspondence of the value 0. The most famous approaches include Hurdle models (Mullahy (1986)) and Zero-Inflated models (Lambert (1992)). The first type of models specifies a mixture of a point mass at zero and a zero truncated distribution for the non-zero observations. Differently, Zero-Inflated models mix a standard distribution with the Dirac measure and consequently model the inflation (or the deflation) of the probability of the zero outcomes.

In this work we focus prominently on Zero-Inflated Poisson (ZIP) distributions. ZIP models can be extended to incorporate covariate information through regressions using convenient link functions on both the mixing probability and the mean parameter of the Poisson distribution. In order to account for the heterogeneity of the patients, random effect models are also employed (Hall (2000),

Leann Long et al. (2015), Agarwal, Gelfand and Citron-Pousty (2002)). Random effects can either be assigned individually to each observation or to clusters of observations. The latter approach is more parsimonious in the number of parameters to be estimated but, when the clustering structure of the observations is not known a priori, it is often problematic to determine the number of clusters and their compositions in order to assign effectively the random effect avoiding problems of overfitting.

This motivates our proposed approach. Placing a prior distribution over the partition of the observations allows learning from the data the clustering structure and capturing patients heterogeneity within the data.

3. Bayesian nonparametric ZIP model (BNP-ZIP)

We present in this section a nonparametric model capable of dealing with observations having excess of zeros and accounting for clustering of individuals. We briefly show some properties of the model and we discuss Markov Chain Monte Carlo (MCMC) algorithms for posterior and predictive inference.

3.1. Random partition zero-inflated Poisson model

It is often of interest to model response variables within clusters. This allows us to account for possible patterns within the data as well as for highly dispersed observations and outliers. A common assumption is to consider the observations within each cluster as generated *iid* from a distribution having cluster-specific parameters. However, when the data are not naturally in clusters, it is also convenient to learn the clustering structure from the data. Thus, the strategy employed in this work consists in specifying a convenient prior over the partition of the patients and to fit independent models within each cluster. This modeling strategy belongs to the class of Random Partition Models (RPM, Lau and Green (2007)).

Let us introduce a convenient notation to deal with the partition of a set. The collection of sets $\rho_n = \{S_1, \ldots, S_k\}$ defines a partition of the set $N = \{1, \ldots, n\}$ if $\bigcup_{j=1}^{k} S_j = N$ and $S_j \bigcap S_{j'} = \emptyset$ for all j different from j'. We can consider N to be the set containing the labels of the observations while the sets in ρ_n denote clusters of observations with k being the number of clusters. The same partition can also be identified by using the cluster assignment vector $\boldsymbol{s} = (s_1, \ldots, s_n)$, whose components take value into the set of cluster labels, *i.e.* $\{1, \ldots, k\}$.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a collection of variables presenting an out-of-pattern number of zero observations. We assume the following joint model for the components of $\boldsymbol{y}$:

$$\boldsymbol{y} \mid \rho_n, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^* \sim \prod_{j=1}^{k} \prod_{i \in S_j} [(1 - \mu_j^*)\delta_0(y_i) + \mu_j^* \text{Poisson}(y_i \mid \lambda_j^*)], \tag{1}$$

where $\mu_j^* \in (0,1)$ and $\lambda_j^* \in (0,+\infty)$ are cluster-specific parameters, while $\delta_0(y_i)$ is the Dirac measure which places a unitary mass of probability in correspondence of $y_i = 0$.

Within each cluster, the model in (1) is a mixture between two distributions: the first one is a point mass located at 0 and the second one is a Poisson distribution with cluster-specific mean equal to λ_j^*. The model above implies that $\Pr(y_i = 0 \mid s_i, \mu_{s_i}^*, \lambda_{s_i}^*) = 1 - \mu_{s_i}^* + \mu_{s_i}^* \exp(-\lambda_{s_i}^*)$ and consequently that the probabilities of all other outcomes different from 0 follow a rescaled Poisson distribution. The role of the parameter μ_j^* is crucial since it determines the inflation level for the probability of the 0 outcome. Note that under a conventional Poisson distribution with mean λ we have $\Pr(y_i = 0 \mid \lambda) = \exp(-\lambda)$.

An alternative distribution on the counts may be employed in (1) instead of the Poisson. A common example is represented by the Negative Binomial, which having two parameters can account for over-dispersed observations within each cluster. A useful parameterization of the Negative Binomial that can be employed in this context is the one involving mean and dispersion parameters. Assuming these parameters together with the parameter controlling the zero-inflation to be cluster-specific allows writing an equivalent random partition Zero-Inflated NB (ZINB) model.

3.2. Prior partition model

A RPM requires the specification of a prior distribution over ρ_n. A common choice is to use the distribution over partitions implied by the so called Chinese Restaurant Process (CRP, Aldous (1985))

$$p(\rho_n \mid \alpha) \propto \prod_{j=1}^{k} \alpha (n_j - 1)!, \tag{2}$$

where α is a positive scalar parameter and n_j is the cardinality of cluster S_j. The distribution above implies that also k is random taking value in $\{1, \ldots, n\}$. We clarify the role of α showing how to sample sequentially a partition from the CRP. Let us consider the point $i = 1$ and assign it to cluster $j = 1$ (*i.e.* S_1) with probability 1. CRP assumes that the probability for the observation $i = 2$ to be assigned either to a new cluster or to cluster $j = 1$ is proportional to α and the cardinality of S_1 (*i.e.* n_1, that in our example is equal to 1) respectively. The same procedure applies for all other points up to $i = n$. Thus, α determines the number of different clusters k.

3.3. Clustering with covariates information

When covariates are available, it can be convenient to modify the CRP prior for the partition of the observations in (2) in order to include clustering information contained within the covariates. This is equivalent to assume higher prior probability for two individuals having the same (or similar) covariate profile to co-cluster. The specification of a distribution over the partition of the

observations which could include covariates information has recently received remarkable attention in RPM literature and the variety of solutions have been discussed by Müller and Quintana (2010).

In this work we opt for specifying a model for the covariates in order to construct a covariate dependent model on the partition of the observations. This strategy is one of the most common in practice for its computational tractability and it has been introduced by Müller, Erkanli and West (1996). Extensions have been presented by Shahbaba and Neal (2009), Park and Dunson (2010), Molitor et al. (2010), Müller, Quintana and Rosner (2011), Hannah, Blei and Powell (2011). Let us consider a matrix of binary covariates $\boldsymbol{X}$ with n rows and D columns and denote with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})$ a generic row of $\boldsymbol{X}$. Similarly to $\boldsymbol{y}$, we assume clusters of rows of $\boldsymbol{X}$ to be generated by the same distribution. We use $\boldsymbol{\zeta}_j^* = (\zeta_{j1}^*, \ldots, \zeta_{jD}^*)$ to denote the cluster-specific parameters for the model of the covariates and we write

$$\boldsymbol{X} \mid \rho_n, \boldsymbol{Z}^* \sim \prod_{j=1}^{k} \prod_{i \in S_j} \prod_{d=1}^{D} \text{Bernoulli}(x_{id} \mid \zeta_{jd}^*), \tag{3}$$

where $\boldsymbol{Z}^* = (\boldsymbol{\zeta}_1^*, \ldots, \boldsymbol{\zeta}_k^*)$.

The formulation proposed above allows to modify (2) writing the conditional probability of the partition given the covariates, which is

$$p(\rho_n \mid \alpha, \boldsymbol{X}, \boldsymbol{Z}^*) \propto \prod_{j=1}^{k} \alpha (n_j - 1)! \prod_{i \in S_j} \prod_{d=1}^{D} \text{Bernoulli}(x_{id} \mid \zeta_{jd}^*), \tag{4}$$

and we adopt the latter to be the prior over the random partition of the observation. The second part in the distribution above represents the likelihood of the covariates within cluster S_j which takes larger values in clusters having *similar* covariates. This corrects the probability of the partition implied by the CRP favoring clusters containing homogeneous covariate patterns.

An advantage of the proposed model on the partition of the observations is the flexibility with respect to the covariate type. In (4), modifying the model on the covariates with other suitable distributions allows the user to include in the partition information from different (or mixed) covariate types. On the other hand, the main disadvantage of this formulation arises when a large number of covariates is included in the model. In this situation, the clustering information contained in the covariates tends to dominate the partition which becomes insensitive to the clustering patterns contained in the outcome. A possible solution to this problem has been presented by Wade et al. (2014).

3.4. Joint probability model

We call the resulting model Bayesian Nonparametric ZIP model (BNP-ZIP), which can be summarized by the following joint probability model

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{X}, \rho_n, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*, \boldsymbol{Z}^*, \alpha) = \\ p(\boldsymbol{y} \mid \rho_n, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) p(\boldsymbol{X} \mid \rho_n, \boldsymbol{Z}^*) p(\rho_n \mid \alpha) p(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*, \boldsymbol{Z}^*) p(\alpha), \end{aligned} \tag{5}$$

where $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ are independent of $\boldsymbol{Z}^*$. From (5) we can derive $p(\boldsymbol{y}, \rho_n, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*, \boldsymbol{Z}^*, \alpha \mid \boldsymbol{X})$, which gives an RPM with a covariate dependent partition.

An important aspect of the proposed formulation is that the joint model in (5), when $p(\rho_n \mid \alpha)$ is as in (2), corresponds to the joint model under a Dirichlet Process Mixture (DPM, Lo (1984)) model in which a Dirichlet Process (DP, Ferguson (1973); Antoniak (1974)) prior is specified for the parameters of the response and the covariates. Specifically, the joint model in (5) can be rewritten as the following hierarchical model:

$$\begin{aligned}
y_i, \mid \mu_i, \lambda_i &\sim (1-\mu_i)\delta_0(y_i) + \mu_i \text{Poisson}(y_i \mid \lambda_i) \\
\boldsymbol{x}_i \mid \boldsymbol{\zeta}_i &\sim \prod_{d=1}^{D} \text{Bernoulli}(x_{id} \mid \zeta_{id}) \\
(\mu_i, \lambda_i, \boldsymbol{\zeta}_i) \mid G &\sim G = \sum_{j=1}^{\infty} v_j^* \prod_{l<j} (1-v_l^*) \delta_{(\mu_j^*, \lambda_j^*, \boldsymbol{\zeta}_j^*)} \\
v_j^* \mid \alpha &\sim \text{Beta}(v_j^* \mid 1, \alpha) \\
(\mu_j^*, \lambda_j^*, \boldsymbol{\zeta}_j^*) &\sim p(\mu_j^*, \lambda_j^*, \boldsymbol{\zeta}_j^*) = p(\mu_j^*, \lambda_j^*) p(\boldsymbol{\zeta}_j^*) \\
\alpha &\sim p(\alpha).
\end{aligned} \tag{6}$$

The random quantity G in the model above has been constructed using the so called stick-breaking procedure and it has been proved by Sethuraman (1994) to be DP distributed. Details about the relationship between CRP and DP may be found in Blackwell and MacQueen (1973) while the connection between RPM and DPM models is presented in Quintana and Iglesias (2003). The equivalence between the BNP-ZIP and a DPM model is very useful when performing posterior inference.

3.5. Prior specification

The model described above is completed by specifying the hyperprior distributions for the parameters. We assume independent prior distributions for the cluster specific parameters

$$\begin{aligned}
\mu_j^* &\sim \text{Beta}(\mu_j^* \mid a_\mu, b_\mu) \\
\lambda_j^* &\sim \text{Gamma}(\lambda_j^* \mid a_\lambda, b_\lambda) \\
\boldsymbol{\zeta_j^*} &\sim \prod_{d=1}^{D} \text{Beta}(\zeta_{jd}^* \mid a_\zeta, b_\zeta).
\end{aligned}$$

We also assume a prior distribution for the parameter α of the distribution of ρ_n. This parameter takes value into the set of positive real numbers, thus we employ a Gamma prior distribution with parameters a_α and b_α. However a prior distribution over subsets of its real support can also be employed when in (6) the random distribution G is replaced with its truncated version (up to K

mixture components)

$$G_K = \sum_{j=1}^{K} v_j^* \prod_{l<j} (1 - v_j^*) \delta_{(\mu_j^*, \lambda_j^*, \zeta_j^*)},$$

for computational reasons. A detailed discussion about the approximation of G with G_K has been presented in Ishwaran and James (2002). When G_K is employed Ohlssen, Sharples and Spiegelhalter (2007) discuss the choice of a Uniform prior distribution for α.

3.6. Posterior inference and MCMC

The model described above is a joint RPM model on the response and the covariates. The connection between the proposed RPM and the DPM model highlighted above is convenient since it allows using available efficient Markov Chain Monte Carlo (MCMC) algorithms developed for DPM models for sampling from the posterior distributions. A review of these algorithms is found in Neal (2000).

In a Gibbs fashion, the posterior inference can be divided in three main stages. In the first stage we resample ρ_n from its full conditional, whereas in the second one we resample the cluster specific parameters of the response and the covariates from their full conditional distributions and finally we resample α from its full conditional distribution. The first stage can be performed using the Algorithm 8 in Neal (2000), or alternatively through the Blocked Gibbs sampler proposed by Ishwaran and James (2001). The cluster specific parameters are resampled independently across clusters. A Metropolis-within-Gibbs step can be designed for the parameters of the response, while known full conditional distributions are available for the parameters of the covariate model. The resampling of α can be performed using a Metropolis-within-Gibbs step. Alternatively, imposing a Gamma prior on α leads to tractable full conditional distribution as discussed in Escobar and West (1995).

Posterior inference for BNP-ZIP can be also performed using `WinBUGS` (Lunn et al. (2000)), `JAGS` (Plummer et al. (2003)) or `Stan` (Carpenter, Gelman and Hoffman (2015)) softwares for Bayesian inference. `JAGS` code is provided in the appendix. All these softwares implement a truncated version of the DPM model to perform inference (details about the truncated approach to DPM models are presented in Ishwaran and James (2001)).

Posterior predictive inference is a key aspect of BNP-ZIP. Using the distribution in (2) for the partition allows the model to grow in complexity when new observations arise adding clusters to the partition. Furthermore, enriching (2) with the information of the covariates, as showed in (3), encourages observations with similar covariates to be assigned to the same cluster and hence to predict similar responses. In a standard statistical problem the response of a new individual is unknown and needs to be evaluated, while the covariates are available. Denoting with $\tilde{\boldsymbol{x}}$ and $\tilde{y}$ respectively the covariates and the response

for a new individual, the predictive distribution $p(\tilde{y} \mid \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{x}})$ can be evaluated within the MCMC scheme assigning the new individual to a cluster given the available information ($\tilde{\boldsymbol{x}}$ included) and sampling from the distribution in (1) using the parameters $\mu^*_{\tilde{s}}$ and $\lambda^*_{\tilde{s}}$, where $\tilde{s}$ is the cluster allocation for the new observation with covariates $\tilde{\boldsymbol{x}}$. Note that if $\tilde{s}$ indicates a new cluster the two parameters are sampled from their prior distributions. Details of this procedure are presented in Müller, Quintana and Rosner (2011).

4. Data Analysis: Lower urinary tract symptoms

In this section we present the analysis of the LUTS data using the BNP-ZIP model. After a detailed presentation of the data, we describe the results in terms of clustering of the patients and of predictive inference. We also highlight the medical implications of the results.

4.1. Data

In this study we consider $n = 1424$ patients at the first visit attendance at the *Lower Urinary Tract Service Clinic* (Whittington Hospital, London, UK). All patients are female over 18 years of age. For each of them the result of the microanalysis of a sample of urine has been recorded in terms of the WBC count. Presence of WBC in the urine (regardless of the quantity) indicates the presence of Urinary Tract Infection (Kupelian et al. (2013)). It is worth noticing that a large number of WBC is also the sign of a high degree of inflammation and thus can be somehow treated as an indicator of the severity of the infection. The empirical distribution of WBC count is strongly positively skewed: this is due to the fact that over 50% of the counts is equal to 0. Moreover the WBC counts different from 0 are highly dispersed, ranging from 1 to 3840.

For each of the patients a profile of LUTS has been recorded. Each profile contains information about four different types of symptoms: urgency symptoms, pain symptoms, stress incontinence symptoms and voiding symptoms. We recorded the profiles by binary vectors with 4 components, each taking value equal to 1 when the correspondent category of symptoms is activated and zero otherwise. On average patients have between 2 and 3 categories activates, and there are 66 patients that do not show any symptom. 161 patients suffer from all four categories of symptoms.

4.2. Prior settings

For the analysis of the data described above we set the hyperparameters $a_\mu = b_\mu = a_\zeta = b_\zeta = 1$, implying minimal prior information. Also the hyperparameters a_λ and b_λ are set equal to 1. We adopt the blocked Gibbs sampler to sample from the full conditional distribution of the partition, approximating the complexity of the model up to a certain number of possible occupied clusters. We consider $K = 70$ as maximum number of clusters and we also set the

hyperparameters a_α and b_α equal to 1. The truncation of the Dirichlet process has been discussed by several authors. Following the results in Ishwaran and James (2002), the adopted truncation level leads to negligible approximation error (given the levels of α explored by the Gibbs sampler). A different practical approach to determine K has been discussed by Ohlssen, Sharples and Spiegelhalter (2007), who employ also a Uniform prior for α on the set $(0, 10)$. This allows to set a priori the largest possible approximation error.

We initialize the MCMC chain taking random starting points from the prior distributions. We save 20000 samples after a burnin period of 10000 interactions. The convergence of the MCMC chain to the posterior distribution has been assessed by trace plots and computing sample autocorrelations and effective sample sizes.

4.3. Results

In this section we present the results obtained fitting the BNP-ZIP on the LUTS dataset. We recall that the objective is to identify the categories of symptoms most associated with infection, *i.e.* with a count of WBC larger than 0. Furthermore, we want to assess which category of LUTS indicate a high level of WBC, which is then related to the severity of the UTI.

4.3.1. Clustering output

The starting point of our analysis consists in investigating the posterior distribution of the partition of the observations, *i.e.* $p(\rho_n \mid \boldsymbol{y}, \boldsymbol{X})$.

In order to investigate the composition of the clusters in terms of patients we compute the posterior probability for all pairs of observations to be assigned to the same cluster. These probabilities can be computed using the samples from $p(\rho_n \mid \boldsymbol{y}, \boldsymbol{X})$ of the MCMC algorithm. With the aim of highlighting the patterns that lead to the clustering structure, we plot the probabilities of co-clustering ordering the observations according to different criteria. Figure 1 shows the probabilities of co-clustering ordering the patients for increasing values of WBC (left panel) and grouping the observations in terms of observed combinations of the covariate profiles (right panel).

In the left panel, blocks of observations with large probability of co-clustering are clearly visible along the diagonal of the plot. These blocks correspond to groups of observations having similar responses. Evidently, these do not mix with other groups, indicating quite distinct clusters of patients. An interesting exception is represented by the first block that represents the patients with response equal to 0. In our dataset the patients with WBC equal to 0 are 717. This block mixes with the blocks on the top right corner, which are those with the largest value of WBC, underlying the difficulty of the diagnostic process for UTI: similar patients (in terms of symptoms) may have a severe infection (using the number of WBC for evaluating the severity of UTI) or no UTI.

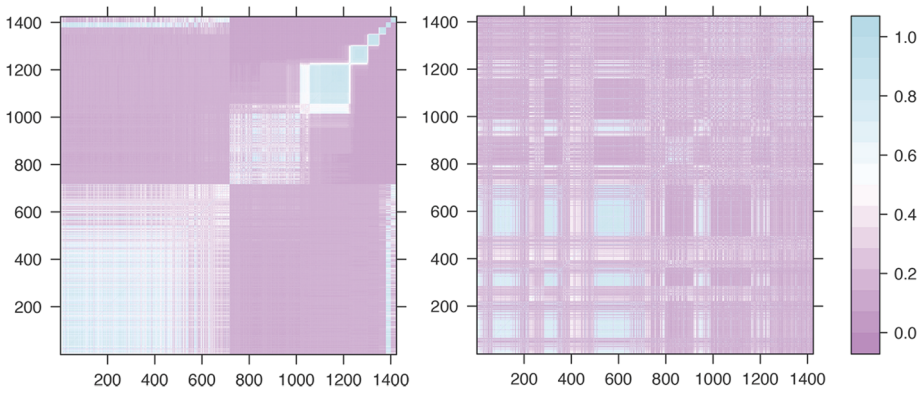

FIG 1. *Levelplots of the probabilities of co-clustering of the patients ordered by increasing value of the response (left panel) and combinations of activated covariates (right panel).*

TABLE 1

Combinations of the covariates. The columns From and To identify the positions of the groups of patients sharing the same combination of covariates in Figure 1 (right panel)

Index	Urgency	Pain	Incontinence	Voiding	From	To
1	0	0	0	0	1	66
2	1	0	0	0	67	220
3	0	1	0	0	221	285
4	0	0	1	0	286	362
5	0	0	0	1	363	394
6	1	1	0	0	395	494
7	1	0	1	0	495	712
8	1	0	0	1	713	785
9	0	1	1	0	786	797
10	0	1	0	1	798	917
11	0	0	1	1	918	936
12	1	1	1	0	937	990
13	1	1	0	1	991	1159
14	1	0	1	1	1160	1246
15	0	1	1	1	1247	1263
16	1	1	1	1	1264	1424

The right panel in Figure 1 displays the co-clustering probabilities rearranging the patients by different combinations of the covariates. Specifically, each covariate profile is composed by four binary indicators which imply 16 different combinations of covariates (all observed in the dataset).

Table 1 indexes the different combinations of the covariates following the order in which they appear in the right panel of Figure 1. Moreover, it gives the exact positions of the groups of patients characterized by the same covariate profile on the levelplot. Looking at the areas with high probability of co-clustering in the right panel of Figure 1 we notice that patients having the first seven combinations of covariates tend to co-cluster with the patients presenting the same symptoms and also with some of the patients having only one category activated. From the left panel of the same figure we know that ordering the pa-

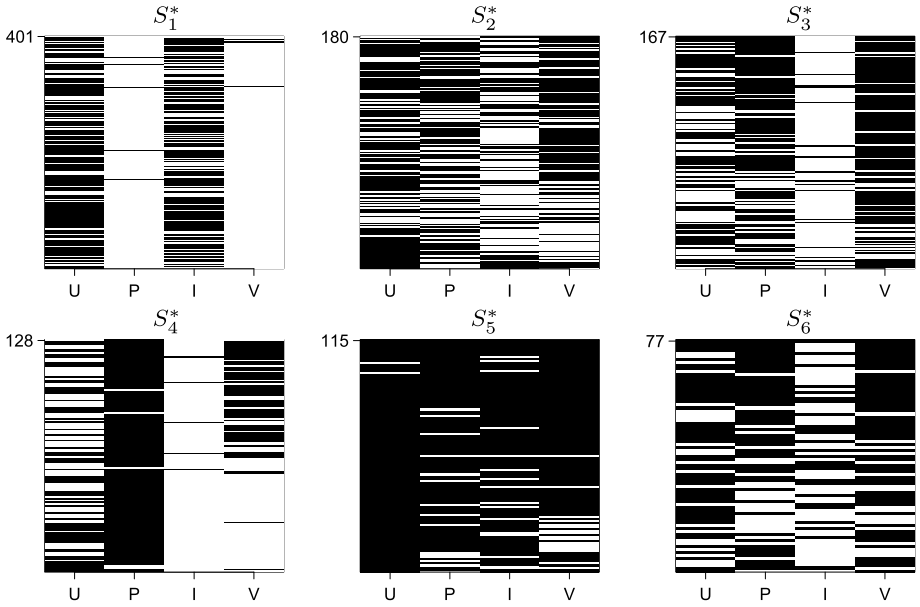

FIG 2. *Symptom indicators (black) for the six largest clusters in* ρ_n^*, i.e. *the partition estimated minimizing the Binder loss function. For each panel (corresponding to a cluster), the horizontal axis is related to the symptoms, whereas vertical axis shows the patients for each cluster. The number in the top-left corner of each panel corresponds to the cluster size.*

tients based on the value of WBC highlights distinct clusters. Therefore, finding high co-clustering probabilities for these indexes of symptoms implies that these are likely to indicate specific mixture components. On the other hand, indexes from 8 to 16 show less evident clustering structure (with some exception for example for index 12). This indicates that the symptom configurations coded with these indexes may belong to different mixture components which are also connected with different values of the WBC.

We further characterize the clusters in terms of symptoms considering a point estimate of ρ_n, say $\rho_n^* = \{S_1^*, \ldots, S_k^*\}$, and controlling which symptoms are activated for the different sets of ρ_n^*. We estimate ρ_n^* minimizing the Binder loss function (Binder (1978)), using $p(\rho_n \mid \boldsymbol{y}, \boldsymbol{X})$. This can be done using the `R` package `mcclust` (https://cran.r-project.org/web/packages/mcclust/).

In Figure 2 we display the composition (in terms of symptoms) of the six largest clusters, which contain 75% of the patients. Each panel corresponds to a cluster: each row of the plot corresponds to a patient while the x-axis represents the four symptoms. Black cells indicate activated symptoms. In most of the panels in Figure 2 a pattern is evident. For example cluster S_1^* (top-left panel), which corresponds to the largest estimated cluster, contains mainly patients with urgency symptoms and stress incontinence symptoms. Other examples are S_4^* (bottom-left panel), which contains mainly patients with the pain symptom activated, or S_5^* which shows patients with all symptoms activated. Recalling

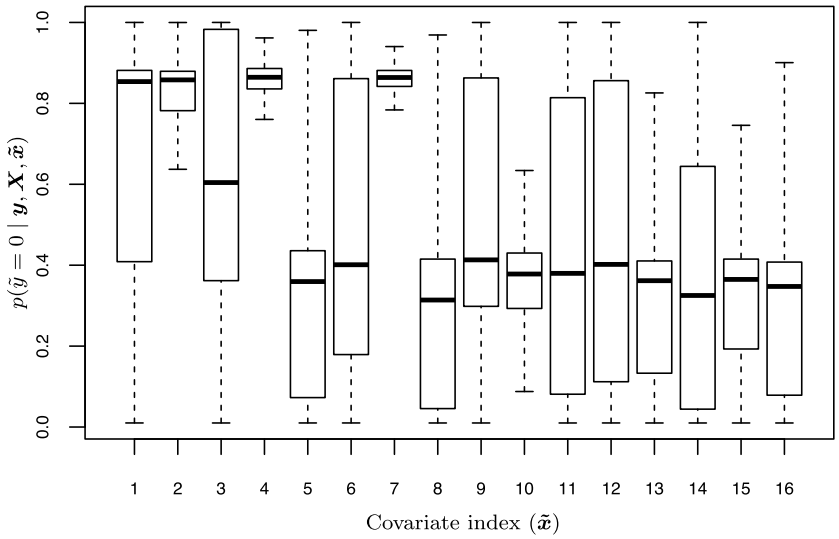

FIG 3. *Posterior predictive distirbution of the probability of WBC equal to 0, given the covariate indexes in Table 1. The black line in each box represents the median.*

that each cluster is associated with similar covariates as well as with response values generated by the same distribution, finding a pattern in the covariates implies that particular symptoms, or combination of symptoms, are predictive of similar response levels.

In order to have a better understanding of how different combinations of symptoms relate with the response, in particular for those symptoms which have high uncertainty about the clustering assignment, we explore the predictive distribution of the response conditioning on symptoms combinations.

4.3.2. Predictive inference

Treating the symptom profiles as random in order to incorporate the covariate information in the partition of the observations has remarkable advantages in practice when the objective is to predict the level of WBC (the response), given the symptom profile $\tilde{\boldsymbol{x}}$. The BNP-ZIP will tend to assign the new patient to the cluster characterized by similar/equal symptoms combination, and thus predict a value of the response similar to the response of the patients in that cluster. This practical advantage has been widely discussed in the literature by Müller, Erkanli and West (1996), Müller, Quintana and Rosner (2011), Park and Dunson (2010), Hannah, Blei and Powell (2011) (among the others) and in the review papers by Müller and Quintana (2010) and Cruz-Marcelo et al. (2013).

We analyze the predictive distribution $p(\tilde{y} \mid \boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{X})$ in order to gain some understanding about the relationship between the different covariates combinations and the presence and severity of UTI. In Figure 3, we plot the posterior

predictive distribution of y, $p(\tilde{y} = 0 \mid \boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{x}})$, for $\tilde{\boldsymbol{x}}$ equal to the different combinations of the covariates indexed according to Table 1. This is equivalent to the predictive distribution of not having UTI. This figure shows that the covariates with index 2,4 and 7 have posterior median probability of WBC equal to 0 close to 0.9 and with small dispersion. Moreover, Figure 2 seems to suggest that these covariate indexes often co-cluster (see top-left panel relative to S_1^*). Also covariate index 1 has a similar median, but with larger dispersion. The first three combinations of the covariates highlight that the categories of urgency and stress incontinence (or their combination) are associated with low probability of UTI, while index 1 corresponds to the configuration without symptoms. Other combinations present similar and very high median probability of having UTI. Interestingly, the profiles presenting voiding category activated have low medians for the probability of WBC equal to 0 and small dispersion. This is evident especially for index 10 and 13, which seem to often belong to the same cluster (see top and bottom right panels referring to S_3^* and S_6^* in Figure 2). Also pain symptoms seem connected with infection, although the respective distributions are right skewed or very dispersed (see box plots relative to the covariates indexed as 3, 6, 9 and 12).

In order to study the relation between the categories of symptoms and the severity of UTI, we compute the distribution of the third quartile of the predictive distribution for all the combinations of the covariates. While clinicians commonly agree that high levels of WBC are connected with complicated infections, the third quartile of the distribution of the WBC does not have *per se* a clinical interpretation. In fact, the choice of the third quartile has only a statistical interpretation. The distributions of these quantities for all symptoms combinations are displayed in Figure 4. The distributions displayed are often right skewed with very long tail. The median of all distributions is smaller than 20. The largest median is associated with profile 14, which has also the second longest tail. Profiles 14 and 8 are characterized by the voiding category activated together with the urgency and stress incontinence categories, which confirms the results about the probability of having UTI. This suggests that not only voiding category indicates high probability of UTI, but also it indicates severe infection (when combined with urgency and stress incontinence problems).

We have performed a similar analysis using a ZINB within-cluster likelihood, which we call Bayesian Nonparametric ZINB model (BNP-ZINB). This has been done to check whether the Poisson assumption within each cluster could be too restrictive. We compare the BNP-ZIP with BNP-ZINB using the Brier score function as described in Section 4.3.3. The results of this comparison show that BNP-ZIP produces more accurate prediction for the presence of infection, while it is comparable to the BNP-ZINB for predicting high values of WBC.

The results of the analysis are of considerable clinical importance. Most clinicians assume that pain is the primary symptom indicative of urinary infection. In fact, some doctors will not consider the diagnosis of UTI in the absence of pain. Thus, the findings of this work suggest that the treatment of the infection should reverse this situation. Regrettable, urologists assume that the voiding symptoms are caused by a structural obstruction of the urethra and

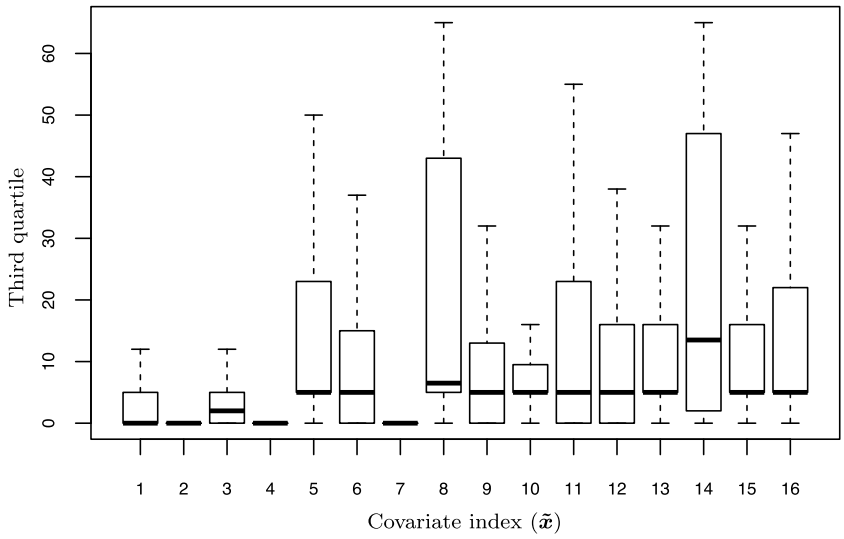

FIG 4. *Distribution of the third quartile of the predictive distribution of WBC, given the covariate indexes in Table 1. The black line in each box represents the median.*

treat affected women by stretching the urethra. This procedure is unlikely to help infection and carries the risk of causing urinary incontinence. Instead, it is more likely that the voiding symptoms arise because of the inflammation induced by swelling of the urethra induced by the infection which in turn causes a relative obstruction to the urinary outflow.

4.3.3. Comparison with related methods

We evaluate the performance of the proposed method comparing it with a Bayesian ZIP model (see for example Neelon, O'Malley and Normand (2010)) and with a DPM of Poisson distributions equivalent to the BNP-ZIP except for the absence of zero-inflating parameters in the likelihood. In order to perform the comparison we divide the the entire data set into a training set (which contains 80% of the records) and a test set, both maintaining the same proportion of covariate types as the whole data set. After fitting the model on the training set we evaluate predictive performance using the test set. We use the distribution of the Brier score (Brier (1950)) which is calculated as

$$\text{Brier}_{(q)} = \frac{1}{m}\sum_{i=1}^{m}\left(f_i^{(q)} - y_i^{(q)}\right)^2$$

where $y_i^{(q)}$ is equal to 1 if $y_i > q$ and 0 otherwise, $f_i^{(q)}$ is the predictive probability under the model to observe a response larger than q and m is

the dimension of the test set. Small values of the Brier score function indicate good predictions. We consider $q = 0, 10, 45$, the latter two being the third quartile and mean of the WBC. The results show that the nonparametric methods outperform evidently the parametric ZIP. Instead, between the BNP-ZIP and DPM of Poisson distributions the differences are less evident (especially for the discretization level equal to 10), but in favor of the proposed method. The same conclusions can be achieved also comparing the models in terms of Deviance Information Criterion (Spiegelhalter et al. (2002)). Although the predictive performances are similar, the main difference between a DPM of Poisson distributions and the BNP-ZIP is in the cluster composition and consequently their interpretation, which we reckon more natural and connected to traditional ZIP regression models. In fact, in our model clusters with the same combination of covariates can accommodate both the excess of zeros and the non-zero counts. On the other hand, the DPM creates clusters with the mean of the Poisson very close to zero to accommodate the excess number of zeros, but it also yields extra clusters if for same combination of symptoms a significant number of high counts are observed. As a result, our model leads to a more parsimonious representation of the clustering structure.

Traditional methods for the analysis of WBC counts using the symptoms as predictors include Classification And Regression Trees (CART, Breiman et al. (1984)) and random forests (Breiman (2001)). These are likelihood-free methods which partition progressively the covariate space according to some decision rule in order to reduce the variability of the associated response variable within each partition set. We compare the partition obtained through these methods with the one estimated by the proposed technique. Both CART and random forests highlight the importance of voiding symptoms. For random forests this has been evaluated using the decrease in residual sum of squares in a cluster (or node) achievable splitting on a certain variable. The importance of voiding symptoms in the analysis with BNP-ZIP has been underlined and it seems evident looking at the division between the groups of indexes in Figure 3 and 4.

4.3.4. Sensitivity analysis

The BNP-ZIP requires the specification of four pairs of hyperparameters, namely (a_α, b_α), (a_ζ, b_ζ), (a_μ, b_μ) and (a_λ, b_λ). We check the sensitivity of our model to different choices of the hyperparameters, focusing on the effects on cluster compositions. We propose two different checks. The first one consists of computing the absolute values of the difference (entry-wise) of the co-clustering probability matrices obtained with the values of the hyperparameters in Section 4.2 (used as reference values) and under alternative scenarios. We summarize the distribution of the entries of the upper-triangular matrix containing the absolute valued differences of the co-clustering probabilities using 95% credible intervals. The second method consists of estimating the mode of the number of clusters (ordered by size) which contains 95% of the patients under different choices of the hyperparameters.

TABLE 2

Results of the sensitivity analysis for different choices of hyperparameters. Upper bound *refers to the upper bound of the 95% credible intervals of distribution of the absolute values of the differences of the co-clustering probabilities.* Mode *indicates the mode of the distribution of the number of clusters (ordered by size) which contain 95% of the patients.*

Scenario	$E(k)$	$V(k)$	Upper bound	Mode
Reference	7.84	44.57	-	10
(i) $a_\alpha = 1, b_\alpha = 5$	2.51	3.59	0.0575	10
(ii) $a_\alpha = 3, b_\alpha = 1$	19.02	95.27	0.1465	14
(iii) $a_\alpha = 5, b_\alpha = 5$	7.84	13.87	0.0690	10
(iv) $a_\alpha = 3, b_\alpha = 2$	10.84	34.18	0.1975	15

TABLE 3

Results of the sensitivity analysis for different choices of hyperparameters. Upper bound *refers to 95% credible intervals of distribution of the absolute values of the differences of the co-clustering probabilities.* Mode *indicates the mode of the distribution of the number of clusters (ordered by size) which contain 95% of the patients.*

Scenario	Upper bound	Mode
Reference	-	10
$(a_\zeta = 0.5, b_\zeta = 0.5)$, $(a_\mu = 1, b_\mu = 1)$, $(a_\lambda = 1, b_\lambda = 1)$	0.0510	10
$(a_\zeta = 1.5, b_\zeta = 1.5)$, $(a_\mu = 1, b_\mu = 1)$, $(a_\lambda = 1, b_\lambda = 1)$	0.0380	10
$(a_\zeta = 1, b_\zeta = 1)$, $(a_\mu = 1, b_\mu = 1)$, $(a_\lambda = 1, b_\lambda = 0.1)$	0.1485	13
$(a_\zeta = 1, b_\zeta = 1)$, $(a_\mu = 1, b_\mu = 1)$, $(a_\lambda = 0.1, b_\lambda = 0.1)$	0.0875	12
$(a_\zeta = 1, b_\zeta = 1)$, $(a_\mu = 0.5, b_\mu = 0.5)$, $(a_\lambda = 1, b_\lambda = 1)$	0.0610	10
$(a_\zeta = 1, b_\zeta = 1)$, $(a_\mu = 1.5, b_\mu = 1.5)$, $(a_\lambda = 1, b_\lambda = 1)$	0.0535	10

We start considering the sensitivity of the proposed model to a_α and b_α, keeping the reference values for the other hyperparameters. We set different scenarios in order to have different values of prior expectation and variance of the number of clusters, *i.e.* $E(k)$ and $V(k)$ (formulae for approximating these quantities are presented by Jara, García-Zattera and Lesaffre (2007)). The reference choice, $a_\alpha = b_\alpha = 1$, leads, for $n = 1424$, to $E(k) \approx 7.84$ and $V(k) \approx 44.57$, which we reckon to be a good trade-off between prior mean and prior variance (*e.g.* compare $E(k)$ with the prior standard deviation of k). The scenarios considered are presented in Table 2.

The results presented in Table 2 show that the proposed model is robust to the choices of hyperparameters in scenario (i) and (iii) compared to the reference scenario. Differently, scenarios (ii) and (iv) are less robust. This suggests that increasing the prior expectation of the number of clusters impact the posterior inference in particular when this variations does not correspond to a proportional increase in the variance of k.

We use the same strategy to assess the sensitivity of the model to the hyperparameters for the distributions of the cluster specific parameters. In addition to the choice adopted in this work, *i.e.* $a_\zeta = b_\zeta = a_\mu = b_\mu = a_\lambda = b_\lambda = 1$ we consider the scenarios in Table 3. The hyperparameters a_α and b_α are set equal to 1 in all scenarios above. Results for all scenarios (including the reference one) are in the same table. These show that under the stated criteria the BNP-ZIP is robust to different values of a_ζ, b_ζ, a_μ and b_μ. Differently, the clustering composition is slightly affected by the choice of a_λ, b_λ. In fact, the distribution of the differences of co-clustering probabilities shifts to higher values following higher

prior variances for λ_j^*. In the same way also the mode of the number of clusters containing the 95% of the patients increase.

5. Discussion

The present work proposes an approach for the study of Lower Urinary Tract Symptoms (LUTS) and their relation with Urinary Tract Infection (UTI). LUTS comprise a group of symptoms that can indicate a variety of diseases, however they are frequently associated with UTI. The latter is identified through the presence of White Blood Cells (WBC) in the urine. Moreover, large WBC counts can also be connected with the severity of the infection and the degree of inflammation. Finally, UTI can become chronic if treatments for acute infections are not delivered promptly. For these reasons it is valuable to gain insight into the relationship between LUTS and UTI and to provide the clinicians with a tool capable of supporting the diagnostic process.

To this end, we propose a model for a dataset of patients affected by LUTS and for which both the symptoms profiles and WBC counts have been provided. More than half of the patients present WBC counts equal to 0, forcing a modeling strategy that could take this into account. Thus, we propose a Zero-Inflated Poisson model for the WBC with cluster-specific parameters. We employ a prior distribution on the possible partitions of the observations that includes also clustering information within the covariates. Covariate information is incorporated by modeling the covariates as random and deriving the distribution of the partition given the covariates.

The proposed model strategy, called BNP-ZIP, builds on existing literatures about covariates dependent random partition models and Zero-Inflated (or deflated) distributions. BNP-ZIP allows estimating the probability of having UTI and the level of UTI (measured in number of WBC in the urine) given the patients' symptoms. Thus, it identifies the combinations of covariates related with the largest probability of having UTI as well as those connected with the largest counts of WBC. Furthermore, the covariate dependent partition can model the over dispersion in the data by including a larger number of clusters leading to robust estimates. BNP-ZIP can be specified also as a Dirichlet Process Mixture of the response variable and the covariates jointly. This property, which has already been widely used in the Bayesian literature, simplifies posterior computations, allowing also the use of convenient MCMC samplers for numerical approximations.

The results show the importance of the urgency and stress incontinence symptoms. Patients with these symptoms activated are often clustered together and have probability close to 0.9 to have WBC equal to 0, which is equivalent to the absence of the infection. On the other hand voiding symptoms are highly related with a large probability of having UTI. Furthermore, a large number of WBC is predicted for the combinations of covariates including voiding category together with the urgency and stress incontinence symptoms, which underlines the importance of voiding symptoms in evaluating UTI. In fact, this has strong

clinical impact since in clinical practice pain symptoms are generally considered related with infection and voiding symptoms are instead treated as consequences of structural obstruction of the urinary tract. In this sense, the estimated predictive distributions may offer an interesting tool for clinicians to support diagnosis.

Appendix

We provide below the `JAGS` code for BNP-ZIP.

```
model{
  C <- 10000 # Zero-trick (see Neelon et al, 2010)

for(i in 1:N) {
    z[i] <- step(y[i] - 1) # Indicator function for y>0

    lambda[i] <- lamj[g[i]] # Parameter for Poisson distribution
    p.y[i] <- muj[g[i]] # Parameter for zero-inflation

##### ZIP model #####
    pz.y[i] <- p.y[i]*(1 - exp(-lambda[i])) # Probability of y>0

    ll[i] <- (1 - z[i]) * log(1 - pz.y[i]) + z[i] * (log(pz.y[i]) + y[i] *
         log(lambda[i]) - lambda[i] - loggam(y[i] + 1) -
         log(1 - exp(-lambda[i]))) # Log-likelihood

    phs[i] <- -ll[i] + C  # Zero-trick

    zeros[i] ~ dpois(phs[i]) # ''zero'' is a vector with n 0 components

##### Covariate model #####
    for(p in 1:P) {
      x[i,p] ~ dbern(phi[g[i],p])
    }

    g[i] ~ dcat(psi[]) # distribution over the cluster assignment
}


##### Within-cluster priors ######
    for(clus in 1:K) {
      muj[clus] ~ dbeta(1,1)I(0.01,0.99)
      lamj[clus] ~ dgamma(1,1)
        for(p in 1:P) {
          phi[clus,p] ~ dbeta(1,1)I(0.01,0.99)
        }
    }

##### Dirichlet Process Prior #####
  alpha ~ dgamma(1,1)
  for(clus in 1:(K - 1)) {
    V[clus] ~ dbeta(1,alpha)
  }
  psi[1] <- V[1]                                    # Stick breaking
  for(clus in 2:(K - 1)) {
```

```
    psi[clus] <- V[clus] * (1 - V[clus-1]) * psi[clus-1] / V[clus-1]
  }
  psi[K] <- 1 - sum(psi[1:(K - 1)])
}
```

This uses a trick to code the Zero-Inflated Poisson model which have been employed in the `WinBUGS` code of Neelon, O'Malley and Normand (2010) (available at http://people.musc.edu/~brn200/winbugs/).

Abbreviations

TABLE 4
List of abbreviations.

Acronym	Full Name
BNP-ZINB	Bayesian Nonparametric Zero-Inflated Negative Binomial
BNP-ZIP	Bayesian Nonparametric Zero-Inflated Poisson
CART	Classification And Regression Trees
CRP	Chinese Restaurant Process
DP	Dirichlet Process
DPM	Dirichlet Process Mixture
LUTS	Lower Urinary Tract Symptoms
MCMC	Markov Chain Monte Carlo
RPM	Random Partition Model
UTI	Urinary Tract Infection
WBC	White Blood Cells
ZINB	Zero-Inflated Negative Binomial
ZIP	Zero-Inflated Poisson

Acknowledgements

We acknowledge the associate editor and two anonymous referees for their comments and suggestions.

References

AGARWAL, D. K., GELFAND, A. E. and CITRON-POUSTY, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics* **9** 341–355. MR1951713

ALDOUS, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983* 1–198. Springer, Berlin, Heidelberg. MR0883646

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* 1152–1174. MR0365969

BINDER, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65** 31–38. MR0501592

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics* 353–355. MR0362614

Breiman, L. (2001). Random forests. *Machine Learning* **45** 5–32.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and regression trees*. CRC press. MR0726392

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78** 1–3.

Carpenter, B., Gelman, A. and Hoffman, M. (2015). Stan: a probabilistic programming language. *Forthcoming*.

Cruz-Marcelo, A., Rosner, G. L., Müller, P. and Stewart, C. F. (2013). Effect on Prediction When Modeling Covariates in Bayesian Nonparametric Models. *Journal of Statistical Theory and Practice* **7** 204–218. MR3196596

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90** 577–588. MR1340510

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 209–230. MR0350949

Gill, K., Horsley, H., Kupelian, A. S., Baio, G., De Iorio, M., Sathiananamoorthy, S., Khasriya, R., Rohn, J. L., Wildman, S. S. and Malone-Lee, J. (2015). Urinary ATP as an indicator of infection and inflammation of the urinary tract in patients with lower urinary tract symptoms. *BMC Urology* **15** 7.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56** 1030–1039. MR1815581

Hannah, L. A., Blei, D. M. and Powell, W. B. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research* **1** 1–33. MR2819022

Irwin, D. E., Milsom, I., Hunskaar, S., Reilly, K., Kopp, Z., Herschorn, S., Coyne, K., Kelleher, C., Hampel, C., Artibani, W. and Abrams, P. (2006). Population-based survey of urinary incontinence, overactive bladder, and other lower urinary tract symptoms in five countries: results of the EPIC study. *European Urology* **50** 1306–1315.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**. MR1952729

Ishwaran, H. and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures. *Journal of Computational and Graphical Statistics* **11**. MR1938445

Jara, A., García-Zattera, M. J. and Lesaffre, E. (2007). A Dirichlet process mixture model for the analysis of correlated binary responses. *Computational Statistics & Data Analysis* **51** 5402–5415. MR2370880

Khasriya, R., Khan, S., Lunawat, R., Bishara, S., Bignal, J., Malone-Lee, M., Ishii, H., O'Connor, D., Kelsey, M. and Malone-Lee, J. (2010). The inadequacy of urinary dipstick and microscopy as surrogate mark-

ers of urinary tract infection in urological outpatients with lower urinary tract symptoms without acute frequency and dysuria. *The Journal of Urology* **183** 1843–1847.

Kupelian, A. S., Horsley, H., Khasriya, R., Amussah, R. T., Badiani, R., Courtney, A. M., Chandhyoke, N. S., Riaz, U., Savlani, K., Moledina, M., Montes, S., O'Connor, D., Visavadia, R., Kelsey, M., Rohn, J. L. and Malone-Lee, J. (2013). Discrediting microscopic pyuria and leucocyte esterase as diagnostic surrogates for infection in patients with lower urinary tract symptoms: results from a clinical and laboratory evaluation. *BJU International* **112** 231–238.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16** 526–558. MR2351079

Leann Long, D., Preisser, J. S., Herring, A. H. and Golin, C. E. (2015). A marginalized zero-inflated Poisson regression model with random effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* **12** 351–357. MR0733519

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10** 325–337.

Molitor, J., Papathomas, M., Jerrett, M. and Richardson, S. (2010). Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics* **11** 484–498.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33** 341–365. MR0867980

Müller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79. MR1399156

Müller, P. and Quintana, F. (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference* **140** 2801–2808. MR2651966

Müller, P., Quintana, F. and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics* **20**. MR2816548

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9** 249–265. MR1823804

Neelon, B. H., O'Malley, A. J. and Normand, S.-L. T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling* **10** 421–439. MR2797247

Ohlssen, D. I., Sharples, L. D. and Spiegelhalter, D. J. (2007). Flexible random-effects models using Bayesian semi-parametric models: appli-

cations to institutional comparisons. *Statistics in Medicine* **26** 2088–2112. MR2364293

Park, J.-H. and Dunson, D. B. (2010). Bayesian generalized product partition model. *Statistica Sinica* **20** 1203–1226. MR2730180

Plummer, M. et al. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* **124** 125. Technische Universit at Wien.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65** 557–574. MR1983764

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* 639–650. MR1309433

Shahbaba, B. and Neal, R. (2009). Nonlinear models using Dirichlet process mixtures. *The Journal of Machine Learning Research* **10** 1829–1850. MR2540778

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 583–639. MR1979380

Wade, S., Dunson, D. B., Petrone, S. and Trippa, L. (2014). Improving prediction from Dirichlet process mixtures via enrichment. *The Journal of Machine Learning Research* **15** 1041–1071. MR3195338