Logit stick-breaking priors for partially exchangeable count data

Distribuzioni a priori stick-breaking logistiche per dati di conteggio parzialmente scambiabili

Tommaso Rigon

Abstract Recently, Rigon and Durante (2018) discussed a Bayesian nonparametric dependent mixture model, which is based on a predictor-dependent stick-breaking construction. They provided theoretical support and proposed a variety of algorithms for posterior inference, including a Gibbs sampler. Their results rely on a formal representation of the stick-breaking construction, which has an appealing interpretation in terms of continuation-ratio logistic regressions. In this paper we review the contribution of Rigon and Durante (2018), and we extend their proposal to the case of partial exchangeability with count data. As an illustration of this methodology, we analyze the number of epileptic seizures of a single patient in a clinical trial.

Abstract *Recentemente, Rigon e Durante (2018) hanno discusso un modello di mistura bayesiano nonparametrico basato su una costruzione di tipo stick-breaking e dipendente da covariate. Gli autori hanno fornito sostegno teorico e hanno introdotto vari algoritmi per condurre inferenza a posteriori, incluso un campionamento di tipo Gibbs. I loro risultati si basano su una rappresentazione formale della costruzione stick-breaking, la quale ha un'interessante interpretazione in termini di regressioni logistiche sequenziali. In questo contributo, viene sintetizzata la proposta di Rigon e Durante (2018), e viene estesa la loro proposta nel caso parzialmente scambiabile con dati di conteggio. Per illustrare le loro metodologie, vengono analizzati il numero di attacchi epilettici di un singolo paziente durante un test clinico.*

Key words: Continuation-ratio logistic regression, Partial exchangeability, Poisson mixture model.

Tommaso Rigon
Dip. di Scienze delle Decisioni, Università Bocconi, e-mail: tommaso.rigon@phd.unibocconi.it

1 Introduction

Let $Y_1,\dots,Y_n \in \mathbb{N}$ be a collection of count response variables, each corresponding to a qualitative covariate $x_i \in \{1,\dots,J\}$, for $i = 1,\dots,n$. The observations $y_1,\dots,y_n$ from $Y_1,\dots,Y_n$ can be naturally divided in J distinct groups, given the covariates $x_1,\dots,x_n$. Our goal is to flexibly model the conditional distributions $\text{pr}(Y = y \mid x = j) = p_j(y)$, for $j = 1,\dots,J$, under the assumption that each data point y_i is a conditionally independent draw from

$$(Y_i \mid x_i = j) \overset{\text{ind}}{\sim} p_j, \qquad i = 1,\dots,n, \tag{1}$$

where p_j denotes the probability mass function of the random variable $(Y_i \mid x_i = j)$. Within the Bayesian framework, assumption (1) is known as *partial exchangeability*, and model elicitation is completed by specifying a prior law Q_J for the vector of probability distributions: $(p_1,\dots,p_J) \sim Q_J$. Broadly speaking, the partial exchangeability assumption reflects an idea of homogeneity within the J subsets of observations but not across them. The prior measure Q_J governs dependence between groups, allowing borrowing of information across them. Maximal dependence, i.e. *exchangeability*, is attained if Q_J is such that $p_1 = \cdots = p_J$ almost surely, reflecting the prior belief that observations belong to the same latent population. Conversely, the case of full heterogeneity arises if each random probability distribution p_j is independent on $p_{j'}$ for any $j \neq j'$, implying that the distinct J groups share no information.

A common and flexible formulation for Q_J is given by mixture models of the form $p_j(y) = \int_\Theta K(y;\theta)\mathrm{d}P_j(\theta)$, where $K(y;\theta)$ denotes a known kernel function and $P_j(\theta)$ a random discrete mixing measure which is allowed to change across groups. In this paper we consider the class of predictor–dependent infinite mixture of Poisson distributions

$$p_j(y) = \int_\Theta \text{Pois}(y;\theta)\mathrm{d}P_j(\theta) = \sum_{h=1}^{+\infty} \pi_{hj}\text{Pois}(y;\theta_h), \qquad j = 1,\dots,J, \tag{2}$$

where $\pi_{hj} = \nu_{hj}\prod_{l=1}^{h-1}(1-\nu_{lj})$ are group–dependent mixing probabilities having a stick-breaking representation [11], and $\text{Pois}(y;\theta)$ denotes the probability mass function of a Poisson with mean θ. Additionally, we assume that the atoms θ_h in (2) are independent and identically distributed (iid) draws from a diffuse baseline measure P_0, that is, $\theta_h \sim P_0$ independently for $h = 1,\dots,+\infty$ and independently on the weights π_{hj}. As for the stick-breaking weights ν_{hj}, we let

$$\text{logit}(\nu_{hj}) = \alpha_{hj}, \qquad \text{with} \quad \alpha_h = (\alpha_{h1},\dots,\alpha_{hJ})^\intercal \overset{\text{iid}}{\sim} \text{N}_J(\mu_\alpha,\Sigma_\alpha), \tag{3}$$

independently for every $h = 1,\dots,+\infty$. Specification of equations (2)-(3) can be regarded as a particular instance of the more general logit stick-breaking process (LSBP) of [8, 9], in which the covariate space is finite dimensional and with a Pois-

son kernel. As such, it inherits all the theoretical and computational properties of LSBP processes, some of which are reviewed in this manuscript.

Let us first consider an equivalent formulation of the logit stick-breaking Poisson mixture model of equations (2)-(3). Leveraging standard hierarchical representations of mixture models, the independent samples $y_1, \dots, y_n$ can be obtained equivalently from the random variable

$$\begin{aligned} (Y_i \mid G_i = h) &\sim \text{Pois}(\theta_h), \\ \text{pr}(G_i = h \mid x_i = j) = \pi_{hj} &= \nu_{hj} \prod_{l=1}^{h-1} (1 - \nu_{lj}), \end{aligned} \tag{4}$$

for every unit $i = 1, \dots, n$, where $G_i \in \{1, 2, \dots, +\infty\}$ is a categorical random variable denoting the mixture component associated to the i-th unit. Each indicator G_i has probability mass function $p(G_i \mid x_i = j)$ which can be written, after some algebraic manipulation, as

$$p(G_i \mid x_i = j) = \prod_{h=1}^{+\infty} \pi_{hj}^{\mathbb{1}(G_i = h)} = \prod_{h=1}^{+\infty} \nu_{hj}^{\mathbb{1}(G_i = h)} (1 - \nu_{hj})^{\mathbb{1}(G_i > h)}, \tag{5}$$

for any $j = 1, \dots, J$. Equation (5) suggests an appealing interpretation of the stick-breaking weights ν_{hj} as the allocation probabilities to component h, conditionally on the event of surviving to the previous $1, \dots, h-1$ components, precisely

$$\nu_{hj} = \text{pr}(G_i = h \mid G_i > h - 1, x_i = j), \tag{6}$$

for each $h = 1, \dots, +\infty$ and $j = 1, \dots, J$. This result, together with the prior formulation of equation (3), allows to interpret the stick-breaking construction (4) in terms of continuation–ratio logistic regressions [12]. This connection with the literature on sequential inference for categorical data is common to all the stick-breaking priors [e.g. 1, 8–10] and provides substantial benefits. Indeed, this characterization implies a simple sequential generative process for each membership indicator G_i and facilitates the implementation of a Gibbs sampler for posterior inference.

We briefly recall here the generative mechanism underlying equations (4), as described in [9], for the j-th group of observations. In the first step of the sequential process, each unit of the j-th group is either assigned to the first component $G_i = 1$ with probability ν_{1j} or to one of the subsequents with probability $1 - \nu_{1j}$. If $G_i = 1$ the process stops, otherwise we draw another binary indicator, with probability ν_{2j}, to decide whether $G_i = 2$ or $G_i > 2$. The following steps proceed in a similar manner. Thus, we can reformulate each $\mathbb{1}(G_i = h) = \zeta_{ih}$, that is, the assignment indicator of each unit to the h-th component, in terms of binary sequential choices

$$\zeta_{ih} = z_{ih} \prod_{l=1}^{h-1} (1 - z_{il}), \quad (z_{ih} \mid x_i = j) \sim \text{Bern}(\nu_{hj}), \tag{7}$$

for each $h = 1, \ldots, +\infty$ and $j = 1, \ldots, J$, where z_{ih} is a Bernoulli random variable representing the h-th sequential decision.

2 Theoretical properties

Let $(P_1, \ldots, P_J)$ denote the vector of dependent random probability measures on $\mathbb{R}^+$ induced by the LSBP in equation (2). Thus, each random probability measure P_j can be represented as

$$P_j(\cdot) = \sum_{h=1}^{+\infty} \pi_{hj} \delta_{\theta_h}(\cdot), \qquad j = 1, \ldots, J. \tag{8}$$

The stick-breaking representation of the π_{hj} implies that the random weights π_{hj} sum to 1 almost surely. Although this result is straightforward to derive, it should not taken for granted because of the analogy with [11], which leverages on peculiar characteristics of the Dirichlet process. This property is formalized in Proposition 1.

Proposition 1 (Rigon and Durante (2018)). *Let $(P_1, \ldots, P_J)$ be a vector of random probability measures defined as in* (8) *and with stick-breaking weights defined as in* (3)*. Then, $\sum_{h=1}^{+\infty} \pi_{hj} = 1$ almost surely for any $j = 1, \ldots, J$.*

Proposition 2 provides some insights about the first two moments of the random vector $(P_1, \ldots, P_J)$.

Proposition 2 (Rigon and Durante (2018)). *Let $(P_1, \ldots, P_J)$ be a vector of random probability measures defined as in* (8)*, with stick-breaking weights defined as in* (3)*. Then, for any measurable set B, and for any $j = 1, \ldots, J$ and $j' = 1 \ldots, J$, it holds*

$$\begin{aligned} \mathrm{E}\{P_j(B)\} &= P_0(B), \\ \mathrm{cov}\{P_j(B), P_{j'}(B)\} &= P_0(B)(1 - P_0(B)) \frac{\mathrm{E}(\nu_{1j}\nu_{1j'})}{\mathrm{E}(\nu_{1j}) + \mathrm{E}(\nu_{1j'}) - \mathrm{E}(\nu_{1j}\nu_{1j'})}. \end{aligned}$$

The expectation of $P_j(\cdot)$ coincides with the base measure $P_0(\cdot)$, which can be therefore interpreted as the prior guess for the mixing measure for any $j = 1, \ldots, J$. Also, the variance of the random probability $P_j(B)$ can be recovered from the above covariance by letting $j = j'$. Unfortunately, the expectations in Proposition 2 are not available in closed form, although they can be easily computed numerically.

As noted by [9], the prior covariance between pairs of random probabilities is governed by the hyperparameters in specification (3) and it is always positive. From a modeling standpoint, this suggests that full heterogeneity among groups—using the terminology of Section 1—can be approximated for some suitable choice of the hyperparameters but it cannot be attained completely. A similar reasoning holds also for maximal dependence among groups which, again, arises only as a limiting case.

3 Posterior inference via Gibbs sampling

In this section we adapt the Gibbs sampler of [9] to the proposed infinite mixture model of Poisson distributions. Our approach exploits representation (4) and the continuation–ratio characterization of the logit stick-breaking prior. By conditioning on the latent indicators $G_1,\dots,G_n$, the model reduces to a set of standard conjugate updates—one for each mixture component—as long as the prior distribution of the atoms is

$$\theta_h \sim \text{Gamma}(a_\theta, b_\theta), \qquad h = 1,\dots,+\infty.$$

Moreover, exploiting the sequential representation, posterior inference for the stick-breaking parameters α_h in (3) proceeds as in a Bayesian logistic regression in which the latent binary indicators z_{ih} in (7) play the role of the response variables, precisely

$$(z_{ih} \mid x_i) \sim \text{Bern}\left(\{1+\exp(-\psi(x_i)^\intercal \alpha_h)\}^{-1}\right), \tag{9}$$

for each $i = 1,\dots,n$ and $h = 1,\dots,+\infty$, where $\psi(x_i) = \{\mathbb{1}(x_i = 1),\dots,\mathbb{1}(x_i = J)\}^\intercal$, and with $\mathbb{1}(\cdot)$ denoting the indicator function. To perform conjugate inference also for α_h, we adapt a recent Pólya-Gamma data augmentation scheme for logistic regression [7] to our statistical model, which relies on the following integral identity

$$\frac{e^{z_{ih}\psi(x_i)^\intercal\alpha_h}}{1+e^{\psi(x_i)^\intercal\alpha_h}} = \frac{1}{2}\int_{\mathbb{R}^+} f(\omega_{ih}) \exp\left\{(z_{ih}-0.5)\psi(x_i)^\intercal\alpha_h - \omega_{ih}(\psi(x_i)^\intercal\alpha_h)^2/2\right\}\mathrm{d}\omega_{ih},$$

for each $i = 1,\dots,n$ and $h = 1,\dots,+\infty$, where $f(\omega_{ih})$ denotes the density function of a Pólya-gamma random variable $\text{PG}(1,0)$. Thus, the updating of α_h for any $h = 1,\dots,+\infty$ can be easily accomplished noticing that—given the Pólya-gamma random variables ω_{ih}—the contributions to the log-likelihood are quadratic in α_h and hence conjugate under the Gaussian priors (3). Moreover, the conditional density

$$f(\omega_{ih} \mid \alpha_h) = \frac{\exp[-0.5\{\psi(x_i)^\intercal\alpha_h\}^2\omega_{ih}]f(\omega_{ih})}{[\cosh\{0.5\psi(x_i)^\intercal\alpha_h\}]^{-1}},$$

defined for every $i = 1,\dots,n$ and $h = 1,\dots,+\infty$, is still a Pólya-Gamma random variable—and therefore conjugate—with updated parameters $f(\omega_{ih} \mid \alpha_h) \sim \text{PG}(1, \psi(x_i)^\intercal\alpha_h)$. This scheme allows posterior inference under a classical Bayesian linear regression.

Before providing a detailed derivation of the Gibbs sampler, we first describe a truncated version of the vector of random probability measure $(P_1,\dots,P_J)$, which can be regarded as an approximation of the infinite process. In line with [8–10], we develop a Gibbs sampler based on this finite representation, which has key computational benefits. We induce the truncation by letting $v_{Hj} = 1$ for some integer $H > 1$ and any $j = 1,\dots,J$, which guarantees that $\sum_{h=1}^{H}\pi_{hj} = 1$ almost surely. According to Theorem 1 in [9], the discrepancy between the two processes is exponentially decreasing in H, and therefore the number of components has not to be very large in practice to accurately approximate the infinite representation. Refer to [9] for a

Algorithm 1: Steps of the Gibbs sampler

begin

[1] Assign each unit $i = 1, \ldots, n$ to a mixture component $h = 1, \ldots, H$;
for *i from* 1 *to n* **do**
Sample $G_i \in (1, \ldots, H)$ from the categorical variable with probabilities

$$\text{pr}(G_i = h \mid -) = \frac{\pi_{hx_i}\text{Pois}(y_i; \theta_h)}{\sum_{q=1}^{H} \pi_{qx_i}\text{Pois}(y_i; \theta_q)},$$

for every $h = 1, \ldots, H$.

[2] Update the parameters α_h, $h = 1, \ldots, H-1$;
for *h from* 1 *to* $H-1$ **do**
for *every i such that* $G_i > h-1$ **do**
Sample the Pólya-Gamma data ω_{ih} from $(\omega_{ih} \mid -) \sim \text{PG}(1, \psi(x_i)^\intercal \alpha_h)$.
Given the Pólya-Gamma data, update α_h from the full conditional

$$(\alpha_h \mid -) \sim \text{N}_J(\mu_{\alpha_h}, \Sigma_{\alpha_h}),$$

having $\mu_{\alpha_h} = \Sigma_{\alpha_h}\{\Psi_h^\intercal \kappa_h + \Sigma_\alpha^{-1}\mu_\alpha\}$, $\Sigma_{\alpha_h} = \{\Psi_h^\intercal \Omega_h \Psi_h + \Sigma_\alpha^{-1}\}^{-1}$, $\Omega_h = \text{diag}(\omega_{i1}, \ldots, \omega_{i\bar{n}_h})$ and $\kappa_h = (z_{i1} - 0.5, \ldots, z_{i\bar{n}_h} - 0.5)^\intercal$, with $z_{ih} = 1$ if $G_i = h$ and $z_{ih} = 0$ if $G_i > h$.

[3] Update the kernel parameters θ_h, $h = 1, \ldots, H$, in (2), leveraging standard results;
for *h from* 1 *to H* **do**
Sample the parameters θ_h from the full conditional

$$(\theta_h \mid -) \sim \text{Gamma}\left(a_\theta + \sum_{i:G_i=h} y_i,\ b_\theta + \sum_{i=1}^{n} \mathbb{1}(G_i = h)\right).$$

more formal treatment. As a historical remark, the idea of truncating discrete nonparametric priors was firstly given by [6] and later developed by [3]. Theorem 1 in [9] is somehow the analogue of these results, for a class of models beyond exchangeability.

Let Ψ_h denote the $\bar{n}_h \times J$ predictor matrix in (9) having row entries $\psi(x_i)^\intercal$, for only those statistical units i such that and $G_i > h-1$. The Gibbs sampler for the truncated representation of model (2) alternates between the full conjugate updating steps in Algorithm 1.

4 Illustration

As an illustration of the proposed methodology, we apply the LSBP Poisson mixture model to the `seizure` dataset, which was already analyzed in [13] and is available in the `flexmix` R package [2]. Data are extracted from a clinical trial conducted

at the British Columbia's Children's Hospital, aiming to assess the effect of intravenous gamma globulin in reducing the daily frequency of epileptic seizures. Our dataset consists of daily myoclonic seizure counts (`seizures`) for a single subject, comprising a series of $n = 140$ days. After 27 days of baseline observation (`Treatment: No`), the subject received monthly infusions of intravenous gamma globulin (`Treatment: Yes`). The relative frequency of counts are shown in the upper plots of Figure 1, where the two groups—days with treatment and days without treatment—are compared.

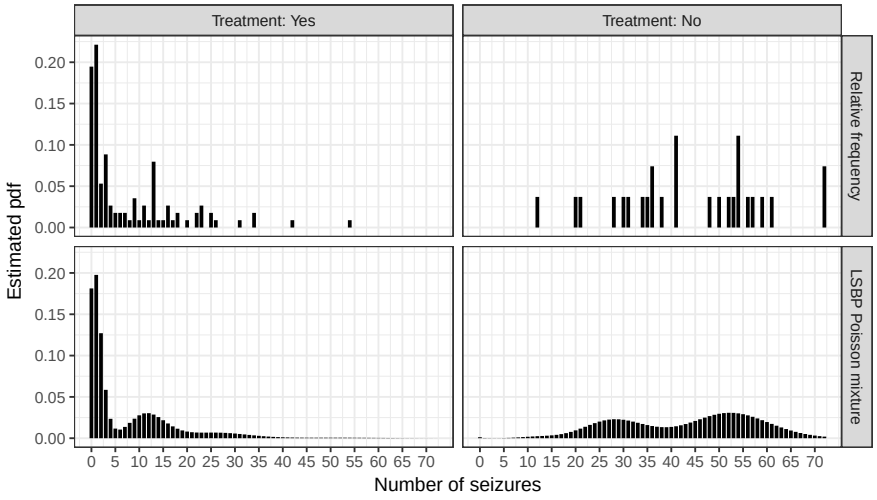

Fig. 1 Upper plots: for the two groups of observations (`Treatment: Yes` and `Treatment: No`), the relative frequencies of the daily number of myoclonic seizure counts are reported. Lower plots: for both groups of observations the (MCMC) posterior expectation of the probability mass function arising from the LSBP model is reported.

As evidenced by the raw frequencies displayed in Figure 1—which seem to present a multimodal structure—and consistent with the discussion in [13], a simple parametric formulation might be overly restrictive for the data at our disposal, thus motivating flexible representations. Additionally, regardless the effectiveness of the treatment, some form of dependence structure among observations from the two groups is expected, since they all refer to the same subject.

Consistent with these considerations, we model the `seizures` counts using the flexible mixture of Poissons described in Section 1. The number of groups is $J = 2$. As prior hyperparameters for the stick-breaking weights in (3), we set $\mu_\alpha = (0,0)$ and $\Sigma_\alpha = \text{diag}(1000, 1000)$, expressing the prior belief of a moderate amount of dependence among groups. As for the kernels parameters, we set $a_\theta = b_\theta = 0.05$, inducing a prior centered on 1 with a relatively large variance. Finally, we truncated the infinite mixture model choosing a conservative upper bound $H = 20$ for the number of mixture components. Although other hyperparameters settings

are certainly possible, this would require a careful sensitivity analysis, which is beyond the scope of this paper. The Gibbs sampler in Section 3 was run for 50000 iterations, discarding the first 5000 draws as a burn-in period. The traceplots showed a satisfactory mixing and no evidence against convergence.

In the lower plots of Figure 1 we report the MCMC approximation of the posterior expectation for the probability mass function under the proposed LSBP Poisson mixture model, for the two groups. From this simple posterior check, it is apparent that our model is able to capture the main tendencies of the data. In particular, the proposed mixture model effectively resembles the multimodal behavior of the data.

References

[1] Dunson, D. B. and Park, J. H. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307–323.

[2] Grün, B. and Leisch, F. (2008). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* **28**, 1–35.

[3] Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

[4] MacEachern, S. N. (1999). Dependent nonparametric processes. In *Proceedings of the Bayesian Section*, Alexandria, VA: American Statistical Association, pp. 50–55.

[5] MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Department of Statistics, Ohio State University.

[6] Muliere, P. and Tardella, L. (1998) Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics* **26**, 283–297.

[7] Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association* **108**, 1339–1349.

[8] Ren, L., Du, L., Carin, L. and Dunson, D. B. (2011). Logistic stick-breaking process. *Journal of Machine Learning Research* **12**, 203–239.

[9] Rigon, T. and Durante, D. (2018). Tractable Bayesian density regression via logit stick-breaking priors. *arXiv:1701.02969*.

[10] Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**, 145–178.

[11] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

[12] Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics & Data Analysis* **11**, 275–295.

[13] Wang, P., Puterman, M. L. and Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics* **52**, 381–400.