# Supplementary Materials for "Clustering blood donors via mixtures of product partition models with covariates" by Argiento R., Corradin R., Guglielmi A., and Lanzarone E.

## Web Appendix A: $\mathcal{D}_{A_j}$ is decreasing with the size of $A_j$

Let $A_j$ be an element of the partition of the sample labels. Since $\mathcal{D}_{A_j} = \sum_{i \in A_j} d(\mathbf{x}_i, \mathbf{c}_{A_j})$ and the Fréchet mean of order one is defined as

$$\mathbf{c}_{A_j} = \arg\min_{\mathbf{c} \in \mathbb{X}} \left\{ \sum_{i \in A_j} d(\mathbf{x}_i, \mathbf{c}) \right\},$$

it is easy to check that

$$\begin{aligned} \mathcal{D}_{A_j \cup \{\ell\}} &= \sum_{i \in A_j \cup \{\ell\}} d(\mathbf{x}_i, \mathbf{c}_{A_j \cup \{\ell\}}) = \sum_{i \in A_j} d(\mathbf{x}_i, \mathbf{c}_{A_j \cup \{\ell\}}) + d(\mathbf{x}_\ell, \mathbf{c}_{A_j \cup \{\ell\}}) \\ &\geq \sum_{i \in A_j} d(\mathbf{x}_i, \mathbf{c}_{A_j \cup \{\ell\}}) \geq \sum_{i \in A_j} d(\mathbf{x}_i, \mathbf{c}_{A_j}) = \mathcal{D}_{A_j}. \end{aligned} \tag{S1}$$

## Web Appendix B: supplementary figures

Let $t = \mathcal{D}_{A_j}$ and $t + \varepsilon = \mathcal{D}_{A_j \cup \{i\}}$, where $\varepsilon$ represents the increment of the average center-based distance when $\{x_i\}$ is assigned to cluster $A_j$. Figure S1 shows the ratio $g(t+\varepsilon)/g(t)$, for different values of $\varepsilon$ and similarity functions $g_A$, $g_B$ and $g_C$. See Section 3 in the manuscript.

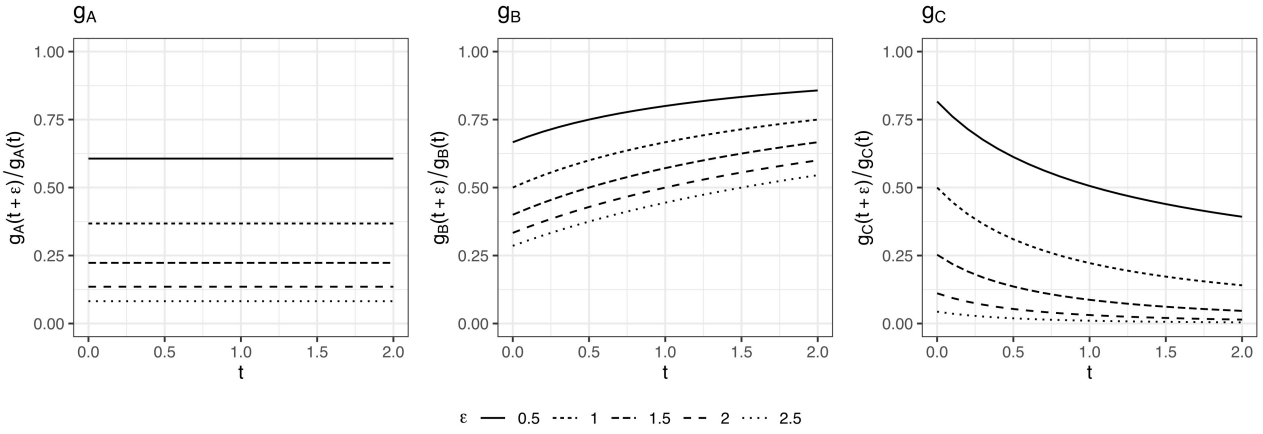


Figure S1: Ratios $g(t+\varepsilon)/g(t)$ as function of $t \in (0, 2)$, for different values of $\varepsilon \in \{0.5, 1, 1.5, 2, 2.5\}$. Left panel: $g_A(\cdot)$. Middle panel: $g_B(\cdot)$. Right panel: $g_C(\cdot)$.

Figures S2-S4 refer to the AVIS application. Figure S2 reports the mean and median trajectories of gap times for any recurrence $j = 1, \ldots, 20$. The average values decrease as $j$ increases, because, as the number of donations increases, the more *loyal* and regular the donor is.

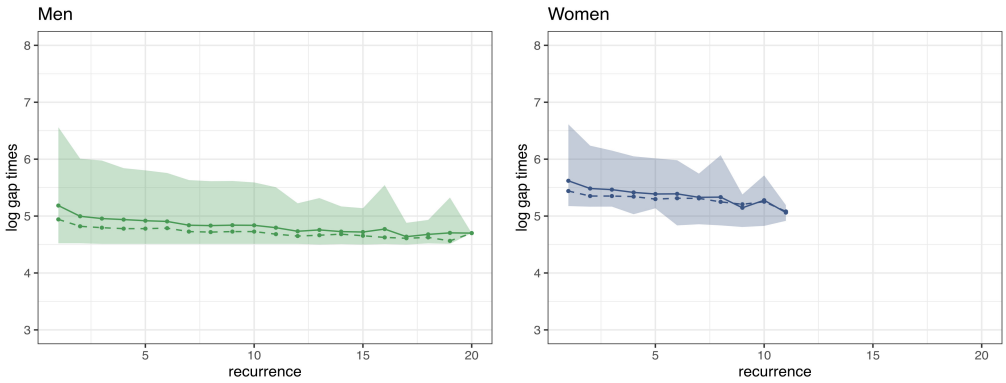


Figure S2: Sample mean (continuous line), median (dashed line), and 90% empirical quantile band of the recurrent gap times, reported on the log scale, for each value of $j \in \{1, \ldots, 20\}$, according to gender men (left) and women (right).

Figure S3 shows the traceplot of the entropy of the visited partitions while sampling from the posterior distribution via the MCMC strategy described in Web Appendix D. The associated effective sample size is 2 659.15 with 5 000 iterations as the final sample size. The figure corresponds to the optimal hyperparameters, i.e., to $\sigma = 0.15$ and $\lambda = 0.1$, but similar traceplots have been obtained for different values.

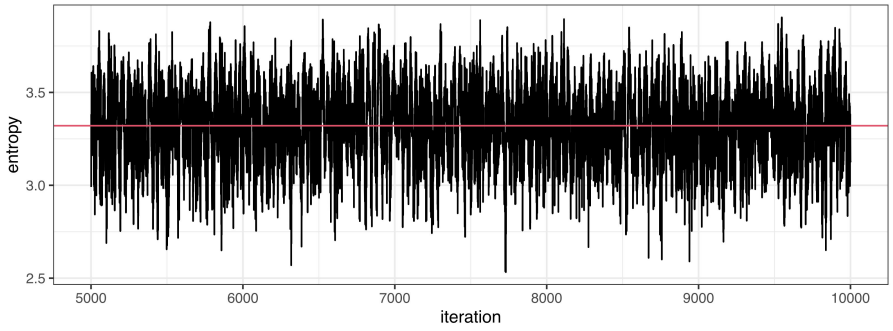


Figure S3: Traceplot of the entropy of the visited partition for the optimal hyperparameters, after burn-in iterations.

Figure S4 shows the regression coefficients for the only time-dependent covariate included in the study (age of the donor). These parameters are significantly different from zero for all donation occasions.

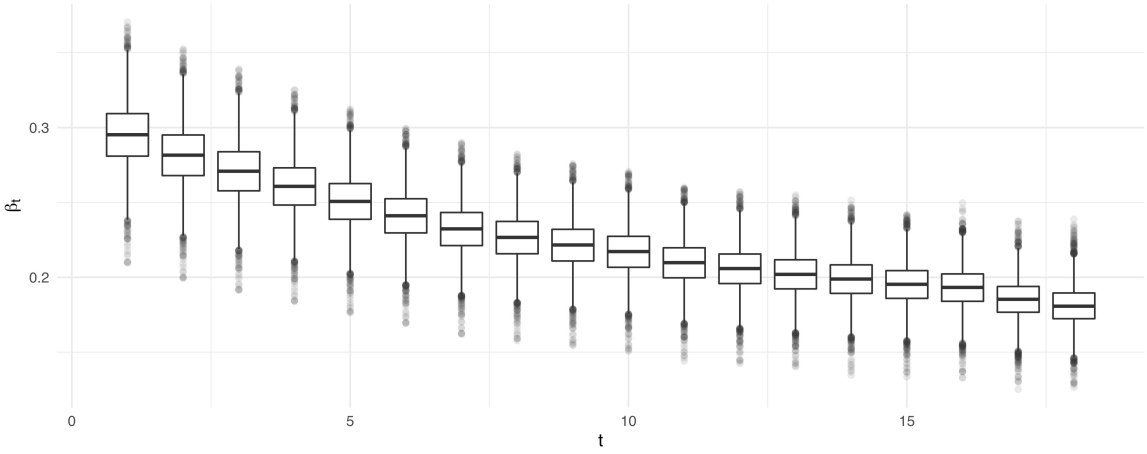


Figure S4: Regression coefficients for the only time-dependent covariate (age of the donor).

# Web Appendix C: Gibbs sampler for the Gaussian kernel with the linear regressor

In this section, we first sketch the Gibbs sampler Pólya urn scheme for the general model (2)-(4). Then, we focus on the particular case of the Gaussian kernel in the linear regression framework.

The joint law of data and all parameters, including the auxiliary variable $u$, is given by

$$\mathcal{L}\left(\{y_i\}_{i=1}^n, \rho_n, \boldsymbol{\theta}^*, u \mid \{\mathbf{x}_i\}_{i=1}^n\right) = \frac{D(u,n)}{M_g(\mathbf{x}_1,\ldots,\mathbf{x}_n)} \prod_{j=1}^{k_n} f(\mathbf{y}_j^* \mid \mathbf{x}_j^*, \theta_j^*) P_0(\theta_j^*) c(u, n_j) g(\mathbf{x}_j^*). \tag{S2}$$

The corresponding algorithm extends the augmented marginal Gibbs sampler for normalized completely random measures mixture models [Favaro and Teh, 2013]; see also Lijoi and Prünster [2010] for an earlier version of the algorithm. We repeatedly sample the full-conditionals below, where $\mathcal{L}(\cdot \mid -)$ indicates that we consider the law of the variable $\cdot$ given all remaining variables $-$ (including the data).

a) given the partition $\rho_n$, the auxiliary variable $u$ and data are independent and the full-conditional of $u$ does not include terms depending on covariates. So we have that $\mathcal{L}(u \mid -) \propto u^{n-1} \mathrm{e}^{-\Psi(u)} \prod_{j=1}^{k_n} c(u, n_j)$ with $\Psi(u) = \kappa \int_0^{+\infty} (1 - \mathrm{e}^{-us}) \rho(s) ds$; in the case of the NGG process, this full-conditional simplifies to

$$\mathcal{L}(\mathrm{d}u \mid -) \propto \frac{u^{n-1}}{(u+1)^{n-\sigma k_n}} \mathrm{e}^{-\frac{\kappa}{\sigma}((u+1)^\sigma - 1)} \mathbb{1}_{(0,+\infty)}(u) \mathrm{d}u.$$

In particular, to sample from this distribution, we suggest using a simple Metropolis-Hastings update with a Gaussian proposal kernel truncated in $(0, +\infty)$

b) For each $j = 1, \ldots, k_n$, we independently sample from

$$\mathcal{L}(\theta_j^* | -) \propto f(\mathbf{y}_j^* \mid \mathbf{x}_j^*, \theta_j^*) P_0(d\theta_j^*) = \prod_{i \in A_j} f(y_j^* \mid \mathbf{x}_j^*, \theta_j^*) P_0(d\theta_j^*) \tag{S3}$$

If $f$ and $P_0$ are conjugate, this step is straightforward. If not, we resort to a different sampling strategy such as, e.g., the algorithm in Web Appendix D.

c) Update the latent partition: the random partition $\rho_n$ is updated using a Gibbs sampling step where the cluster assignment of one item $Y_i$ is updated once at a time. We denote by $\rho_{n-1}^{(-i)}$ the partition of $n-1$ items where the $i$-th item has been removed and by $s_i = j$ the event that $Y_i$ is assigned to cluster $j$, where $j$ varies in $\left\{1, \ldots, k_{n-1}^{(-i)}, k_{n-1}^{(-i)} + 1\right\}$ and $k_{n-1}^{(-i)}$ is the number of clusters available in the partition without $i$. Note that $k_{n-1}^{(-i)} + 1$ is included to consider the case where the item forms a new cluster. Therefore, we have to sequentially sample from the following conditional distribution, for $i = 1, \ldots, n$,

$$\mathcal{L}\left(s_i = j \mid u, \mathbf{x}, \{y_i\}_{i=1}^n, \rho_{n-1}^{(-i)}\right) = \frac{\mathcal{L}\left(\{y_i\}_{i=1}^n \mid u, \mathbf{x}, \rho_{n-1}^{(-i)}, s_i = j\right) \mathcal{L}\left(s_i = j, \rho_{n-1}^{(-i)} \mid u, \mathbf{x}\right)}{\mathcal{L}\left(\{y_i\}_{i=1}^n \mid u, \mathbf{x}, \rho_{n-1}^{(-i)}\right) \mathcal{L}\left(\rho_{n-1}^{(-i)} \mid u, \mathbf{x}\right)}, \tag{S4}$$

where $j = 1, \ldots, k_{n-1}^{(-i)} + 1$, and $\mathbf{x} := \{\mathbf{x}_i\}_{i=1}^n$. Moreover, observe that, for any $l = 1, \ldots, k_{n-1}^{(-i)}$, the prior

on the partition can be written as:

$$\begin{aligned}
\mathcal{L}\left(\rho_{n-1}^{(-i)}, s_i = j \mid u, \mathbf{x}\right) &\propto D(u,n) \prod_{l=1}^{k_n} \left(c(u, n_l) g(\mathbf{x}_l^*)\right) \\
&\propto D(u,n) \prod_{l=1}^{k_{n-1}^{(-i)}} \left(c(u, n_l) g(\mathbf{x}_l^*)\right) c(u, n_j + 1) g(\mathbf{x}_j^* \cup \{\mathbf{x}_i\}) \\
&= D(u,n) \prod_{l=1}^{k_{n-1}^{(-i)}} \left(c(u, n_l) g(\mathbf{x}_l^*)\right) \frac{c(u, n_j + 1) g(\mathbf{x}_j^* \cup \{\mathbf{x}_i\})}{c(u, n_j) g(\mathbf{x}_j^*)} \\
&\propto \mathcal{L}\left(\rho_{n-1}^{(-i)} \mid u, \mathbf{x}\right) \frac{c(u, n_j + 1) g(\mathbf{x}_j^* \cup \{\mathbf{x}_i\})}{c(u, n_j) g(\mathbf{x}_j^*)}
\end{aligned}$$

while, since $g(\emptyset) = 1$, the conditional probability of assigning item $i$ to a new cluster is equal to

$$\mathcal{L}\left(\rho_{n-1}^{(-i)}, s_i = k_{n-1}^{(-i)} + 1 \mid u, \mathbf{x}\right) \propto \mathcal{L}\left(\rho_{n-1}^{(-i)} \mid u, \mathbf{x}\right) c(u, 1)$$

The contribution of the likelihood in (S4) can be written as

$$\begin{aligned}
\mathcal{L}\left(\{y_i\}_{i=1}^n \mid u, \mathbf{x}, \rho_{n-1}^{(-i)}, s_i = j\right) &= \prod_{l=1, l\neq j}^{k_{n-1}^{(-i)}} m(\mathbf{y}_l^*) m\left(\mathbf{y}_j^* \cup \{y_i\}\right) \frac{m(\mathbf{y}_j^*)}{m(\mathbf{y}_j^*)} \\
&= \mathcal{L}\left(\{y_\ell\}_{\ell \neq i} \mid \rho_{n-1}^{(-i)}\right) \frac{m(\mathbf{y}_j^* \cup \{y_i\})}{m(\mathbf{y}_j^*)},
\end{aligned}$$

where $m(\emptyset) = 1$ in the case of a new cluster. Therefore, (S4) becomes

$$\Pr\left(s_i = j \mid u, \mathbf{x}, \{y_i\}_{i=1}^n, \rho_{n-1}^{(-i)}\right) \propto \frac{m(\mathbf{y}_j^* \cup \{y_i\})}{m(\mathbf{y}_j^*)} \frac{c(u, n_j + 1) g(\mathbf{x}_j^* \cup \{\mathbf{x}_i\})}{c(u, n_j) g(\mathbf{x}_j^*)}, \quad j = 1, \ldots, k_{n-1}^{(-i)} \tag{S5}$$

and, similarly,

$$\Pr\left(s_i = k_{n-1}^{(-i)} + 1 \mid u, \mathbf{x}, \{y_i\}_{i=1}^n, \rho_{n-1}^{(-i)}\right) \propto m(y_i) c(u, 1). \tag{S6}$$

Each $s_i$ is sequentially assigned according to the law defined by (S5) and (S6). In particular, $m(\boldsymbol{y}^*)$ is the marginal density of the parametric Bayesian model defined in (S3) computed over the data in the $j$-th cluster $A_j$, namely

$$m(\mathbf{y}_j) := \int_\Theta f(\mathbf{y}_j^* \mid \mathbf{x}_j^*, \theta_j^*) P_0(\mathrm{d}\theta_j^*)$$

This density is available analytically in case $f$ and $P_0$ are conjugate. If not, we need to modify this step; see Web Appendix D, step [7].

We now specialize the previous Gibbs sampler Pólya urn scheme to the case when the sampling model is a linear regression, i.e. when $f(\mathbf{y}_j \mid \mathbf{x}_j^*, \boldsymbol{\theta}_j^*) = \prod_{i \in A_j} \phi(y_i; \mathbf{x}_i^\intercal \boldsymbol{\beta}_j^*, \sigma_j^{2*})$. In this case, the cluster-specific parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ consists of the vector of regression coefficients $\boldsymbol{\beta}$ and the residual variance $\sigma^2$. We assume that the base distribution $P_0$ is conjugate to the mixture kernel, that is

$$\theta = \left(\boldsymbol{\beta}, \sigma^2\right) \sim P_0(d\boldsymbol{\beta}, d\sigma^2) = \mathrm{N}_p(d\boldsymbol{\beta}; \boldsymbol{\mu}_0, \sigma^2 B_0) \times \mathrm{IG}(d\sigma^2; a_0, b_0).$$

Thanks to the choice of a conjugate $P_0$, the algorithm basically consists of a marginal MCMC sampler for nonparametric mixture models in its simplest form [Neal, 2000]. Indeed, in the conjugate case, the vector of cluster parameters $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_{k_n}^*)$ can be efficiently marginalized out from the joint distribution (S2)

obtaining

$$\mathcal{L}\left(\{y_i\}_{i=1}^n, \rho_n, u \mid \{\mathbf{x}_i\}_{i=1}^n\right) \propto D(u,n) \prod_{j=1}^{k_n} m(\mathbf{y}_j) c(u, n_j) g(\mathbf{x}_j^*) \tag{S7}$$

The Gibbs algorithms proceed by sequentially sampling from the following full-conditionals.

a-reg) Update $u$: Since the full-conditional of the mixing parameter $u$, given $\rho_n$, does not include terms that depend on covariates. The full conditional $\mathcal{L}(u \mid -)$ correspond to the general one provided in point a) of the general algorithm.

b-reg) Update the group-specific parameters: each $\theta_j^* = \left(\boldsymbol{\beta}_j^*, \sigma_j^{2*}\right)$, for $j = 1, \ldots, k_n$ is updated within each cluster according to the usual parametric update in the conjugate case with normal likelihood and normal-inverse gamma prior distribution. In particular, we have that, for each $j = 1, \ldots, k_n$, the cluster-specific parameters can be sampled independently from the following distributions:

$$\boldsymbol{\beta}_j^* \mid \sigma_j^{2*}, - \overset{\text{ind}}{\sim} \mathrm{N}_p\left(\boldsymbol{\mu}_j, \sigma_j^{2*} B_j\right)$$

where $B_j = \left(B_0^{-1} + \sum_{i \in A_j} \mathbf{x}_i \mathbf{x}_i^\intercal\right)^{-1}$, $\boldsymbol{\mu}_j = B_j \left(B_0^{-1}\boldsymbol{\mu}_0 + \sum_{i \in A_j} y_i \mathbf{x}_i\right)$ and $\sigma_j^{2*} \mid - \sim \mathrm{IG}\left(a_j, b_j\right)$, with $a_j = a_0 + \dfrac{n_j}{2}$ and $b_j = b_0 + \dfrac{1}{2}\left(\boldsymbol{\beta}_0^\intercal B_0^{-1} \boldsymbol{\beta}_0 + \sum_{i \in A_j} y_i^2 - \boldsymbol{\mu}_j^\intercal B_j^{-1} \boldsymbol{\mu}_j\right)$. Here $n_j := |A_j|$ is the size of cluster $A_j$ in the partition.

c-reg We have to specialize formulas (S5) and (S6). In particular, we have to compute $m(\mathbf{y}_j^*)$, which in this case, thanks to the conjugacy, is available analytically and equals a multivariate t-density.

# Web Appendix D: Gibbs sampler for the blood donations application

In this section, we describe a Gibbs sampler for the posterior of model (9)-(13). The state of the Markov chain is

$$\left\{\{\eta_{it}\}_{t=1}^{m_i+1}, i = 1, .., n; \{Y_{i,n_i+1}^{cens}\}_{i=1}^n; \boldsymbol{\beta}_0; \{\boldsymbol{\beta}_t\}_{t=1}^J; \{\tau_i^2\}_{i=1}^{p_2}; \rho_n; \left\{\left(\alpha_j, \psi_j, \sigma_j^2\right)^\intercal\right\}_{j=1}^{k_n}\right\}.$$

The full-conditionals are outlined below: we provide the details of the computation only when the conditional posterior distribution is not straightforward. As before, $\mathcal{L}(\cdot \mid -)$ indicates that we consider the law of the variable $\cdot$ given all remaining variables $-$ (including the data).

[1] Update the latent variables $\eta_{i,t}$'s: each $\eta_{i,t}$, conditionally on $s_i = j$, is independently sampled from

$$\mathcal{L}(\eta_{i,t} \mid -) \propto \exp\left\{-\frac{1}{2\sigma_j^2}\left(y_{i,t} - \left(\alpha_j + \boldsymbol{\beta}_0^\intercal \mathbf{x}_i + \boldsymbol{\beta}_t^\intercal \mathbf{x}_{i,t} + \psi_j \eta_{i,t}\right)\right)^2 - \frac{1}{2}\eta_{i,t}^2\right\} \mathbb{I}\left(\eta_{i,t} > 0\right)$$

which turns out to be a truncated normal, namely

$$\eta_{i,t} \mid - \sim \mathrm{TN}_{[0,\infty)}\left(\frac{\psi_j}{\sigma_j^2 + \psi_j^2}\left(y_{i,t} - \left(\alpha_j + \boldsymbol{\beta}_0^\intercal \mathbf{x}_i + \boldsymbol{\beta}_t^\intercal \mathbf{x}_{i,t}\right)\right), \frac{\sigma_j^2}{\sigma_j^2 + \psi_j^2}\right)$$

independently for each $t = 1, \ldots, m_i + 1$ and $i = 1, \ldots, n$ such that $s_i = j$.

[2] Update the censored values: the censored observations are independently sampled from

$$Y_{i,m_i+1}^{cens} \mid - \sim \mathrm{TN}_{[y_{i,m_i+1}, +\infty)}\left(\alpha_j + \boldsymbol{\beta}_0^\intercal \mathbf{x}_i + \boldsymbol{\beta}_t^\intercal \mathbf{x}_{i,t} + \psi_j \eta_{i,t}, \sigma_j^2\right)$$

for $i = 1, \ldots, n$. Here $y_{i,m_i}$ is the last observed gap time (in the log scale) for any $i$, and $y_{i,m_i+1} =$

$\log(\tau_i - (e^{y_{i,1}} + \cdots + e^{y_{i,m_i}}))$ is the amount of time, in the log scale, between the censoring time and the time of the last observed event.

[3] Update the common regression coefficients: by conjugacy, the full-conditional is the multivariate $p_1$-dimensional Gaussian, with mean $\widetilde{\boldsymbol{\beta}}_0$ and variance-covariance matrix $\widetilde{\Sigma}_0$, where

$$\widetilde{\Sigma}_0 = \left( \Sigma_0^{-1} + \sum_{j=1}^{k_n} \frac{1}{\sigma_j^2} \left( \sum_{i \in A_j} (m_i + 1) \mathbf{x}_i \mathbf{x}_i^\intercal \right) \right)^{-1}$$

and

$$\widetilde{\boldsymbol{\beta}}_0 = \widetilde{\Sigma}_0 \left( \sum_{j=1}^{k_n} \sum_{i \in A_j} \sum_{t=1}^{m_i+1} \frac{y_{i,t} - (\alpha_j + \boldsymbol{\beta}_t^\intercal \mathbf{x}_{i,t} + \psi_j \eta_{i,t})}{\sigma_j^2} \mathbf{x}_i \right).$$

[4] Update the time-gap specific regression coefficients: each parameter vector $\boldsymbol{\beta}_t$ is sampled independently from the multivariate $p_2$-dimensional Gaussian, with mean $\widetilde{\boldsymbol{\beta}}_t$ and variance-covariance matrix $\widetilde{\Sigma}_t$, where

$$\widetilde{\Sigma}_t = \left( \Xi_0^{-1} + \sum_{j=1}^{k_n} \sum_{i \in A_j : m_i + 1 \geq t} \frac{1}{\sigma_j^2} \mathbf{x}_{i,t} \mathbf{x}_{i,t}^\intercal \right)^{-1}$$

and

$$\widetilde{\boldsymbol{\beta}}_t = \widetilde{\Sigma}_t \left( \sum_{j=1}^{k_n} \sum_{i \in A_j : m_i + 1 \geq t} \frac{y_{i,t} - (\alpha_l + \mathbf{x}_i^\intercal \boldsymbol{\beta}_0 + \psi_l \eta_{i,t})}{\sigma_l^2} \mathbf{x}_{i,t} \right),$$

for $t = 1, \ldots, J$. Here $\Xi_0$ is the diagonal matrix $\text{diag}(\xi_1^2, \ldots, \xi_{p_2}^2)$.

[5] Update the dispersion parameter $\xi_1^2, \ldots, \xi_{p_2}^2$: each parameter is independently sampled from

$$\xi_m^2 \mid - \sim \text{IG}\left(\nu_m, \gamma_m\right),$$

with $\nu_m = \nu_0 + \dfrac{J}{2}$ and $\gamma_m = \gamma_0 + \dfrac{1}{2} \sum_{t=1}^{J} \beta_{tm}^2$, for $m = 1, \ldots, p_2$.

[6] Update the cluster-specific parameters: the likelihood for data in cluster $A_j$ that is used to build the joint distribution for $(\alpha_j, \psi_j, \sigma_j^2)$ is proportional to

$$\mathcal{L}(\mathbf{y}_j^* \mid -) \propto \prod_{i \in A_j} \prod_{t=1}^{m_i+1} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left( -\frac{1}{2\sigma_j^2} \left( \hat{y}_{i,t} - (\alpha_j + \psi_j \eta_{i,t}) \right)^2 \right)$$

with $\hat{y}_{i,t} = y_{i,t} - \boldsymbol{\beta}_t^\intercal \mathbf{x}_{i,t} - \boldsymbol{\beta}_0^\intercal \mathbf{x}_i$. It is straightforward to check that this is the likelihood of a regression model where the regression parameters are $\alpha_j, \psi_j$ and the residual variance is $\sigma_j^2$, with prior $P_0$ as specified in (13), that is the usual conjugate prior. Therefore, the full-conditionals these parameters are:

$$\sigma_j^2 \mid - \sim \text{IG}(\widetilde{a}_j, \widetilde{b}_j)$$
$$(\alpha_j, \psi_j)^\intercal \mid \sigma_j^2, - \sim \text{N}_2(\widetilde{\boldsymbol{\theta}}_0, \sigma_j^2 \widetilde{K}_j)$$

where

$$\widetilde{K}_j = \left( \sum_{i \in A_j} \sum_{t=1}^{m_i+1} \boldsymbol{\zeta}_{i,t} \boldsymbol{\zeta}_{i,t}^{\intercal} + \mathrm{diag}(\kappa_0^{-1}, \kappa_1^{-1}) \right)^{-1}$$

$$\widetilde{\boldsymbol{\theta}}_0 = \widetilde{K}_j \left( \sum_{i \in A_j} \sum_{t=1}^{m_i+1} \hat{y}_{i,t} \boldsymbol{\zeta}_{i,t} + \mathrm{diag}(\kappa_0^{-1}, \kappa_1^{-1}) \boldsymbol{\theta}_0 \right)$$

$$\widetilde{a}_j = a + \frac{n_j}{2}$$

$$\widetilde{b}_j = b + \frac{1}{2} \left( \sum_{i \in A_j} \sum_{t=1}^{m_i+1} \hat{y}_{i,t}^2 + \boldsymbol{\theta}_0^{\intercal} \mathrm{diag}(\kappa_0^{-1}, \kappa_1^{-1}) \boldsymbol{\theta}_0 - \widetilde{\boldsymbol{\theta}}_0^{\intercal} \widetilde{K}_j^{-1} \widetilde{\boldsymbol{\theta}}_0 \right)$$

and $m_j = \sum_{i \in A_j} (m_i + 1)$, $\zeta_{i,t} = (1, \eta_{i,t})^{\intercal}$.

[7] Update the latent partition of the data: we adapt Algorithm 8 in Neal [2000] to consider the non-conjugacy of the kernel density and $P_0$ and to take into account the predictive structure of the PPMx-mixt prior. In the non-conjugate case, the full conditionals of the cluster allocations depend on a vector of cluster-specific parameters. The latter vector must be augmented by considering $R$ new auxiliary variables $\left\{\alpha_r, \psi_r, \sigma_r^2\right\}$ sampled from the prior $P_0$, representing potential new clusters. This augmentation step is implemented to improve the mixing by adopting the re-use strategy in Favaro and Teh [2013]. In particular, the probability of assigning the $i$-th subject to cluster $j$, $j = 1, 2, \ldots, k_n^{-i}$, similarly as in (S5), here becomes

$$\begin{aligned} \Pr\left(s_i = j \mid -\right) \propto & \frac{c(u, n_j + 1) g\left(\mathbf{x}_j^* \cup \{\mathbf{x}_i\}\right)}{c(u, n_j) g\left(\mathbf{x}_j^*\right)} \\ & \times \prod_{t=1}^{m_i+1} \phi\left(y_{it}; \alpha_j + \mathbf{x}_{it}^{\intercal} \boldsymbol{\beta}_t + \mathbf{x}_i^{\intercal} \boldsymbol{\beta}_0 + \psi_j \eta_{it}, \sigma_j^2\right) \end{aligned} \tag{S8}$$

where $\rho_{n-1}^{-i}$ has the same definition as in the previous section. Analogously, observing that $g(\emptyset) = 1$, the probability of allocating the subject to one of the new $R$ clusters is, for $j = k_n^{-i} + 1, ..., k_n^{-i} + R$

$$\Pr\left(s_i = j \mid -\right) \propto \frac{1}{R} c(u, 1) g\left(\mathbf{x}_i\right) \prod_{t=1}^{m_i+1} \phi\left(y_{it}; \alpha_j + \mathbf{x}_{it}^{\intercal} \boldsymbol{\beta}_t + \mathbf{x}_i^{\intercal} \boldsymbol{\beta}_0 + \psi_j \eta_{it}, \sigma_j^2\right) \tag{S9}$$

Specifically, under the NGG assumption, $\dfrac{c(u, n_j + 1)}{c(u, n_j)} = \dfrac{n_j - \sigma}{(1 + u)}$ if $n_j > 0$, and $\kappa (u+1)^{\sigma}$ for $n_j = 0$.

[8] Update the latent parameter $u$: given the partition $\rho_n$, the auxiliary variable $u$ and data are independent, so that $\mathcal{L}(u \mid -) \propto u^{n-1} \mathrm{e}^{-\Psi(u)} \prod_{j=1}^{k_n} c(u, n_j)$ with $\Psi(u) = \kappa \int_0^{+\infty} (1 - \mathrm{e}^{-us}) \rho(s) ds$; in the case of the NGG process, this full-conditional simplifies to

$$\mathcal{L}(\mathrm{d}u \mid -) \propto \frac{u^{n-1}}{(u+1)^{n - \sigma k_n}} \mathrm{e}^{-\frac{\kappa}{\sigma}((u+1)^{\sigma} - 1)} \mathbb{1}_{(0, +\infty)}(u) \mathrm{d}u.$$

In particular, to sample from this distribution, we use a simple Metropolis-Hastings update with a Gaussian proposal kernel truncated in $(0, +\infty)$.

# Web Appendix E: robustness with respect to the similarity function

We provide a simulation study to compare the effect of the three similarity functions on posterior distribution in a linear regression context. We have simulated $n = 200$ observations $(y_i, 1, x_{i1}, \ldots, x_{i4})$ for $i = 1, \ldots, n$. The last two covariates are binary, while the first two are continuous. We generated data independently from three

groups, with sizes 75, 75 and 50, respectively, as follows

$$(x_{i1}, x_{i2}) \overset{\text{iid}}{\sim} \text{N}_2(\boldsymbol{\mu}_j, 0.5\mathbb{I}_2), \quad x_{i3}, x_{i4} \overset{\text{iid}}{\sim} \text{Bern}(q_j), \quad Y_i \sim \text{N}(\mathbf{x}_i^T \boldsymbol{\beta}_j^0, 0.5), \quad j = 1, 2, 3.$$

Specifically,

1) for the first group we set $\boldsymbol{\mu}_1 = (-3, 3)$, $q_1 = 0.1$, and $\boldsymbol{\beta}_1^0 = (1, 5, 2, 1, 0)$;

2) for the second group, we have $\boldsymbol{\mu}_2 = (0, 0)$, $q_2 = 0.5$, and $\boldsymbol{\beta}_2^0 = (4, 2, -2, 1, -1)$;

3) for the third group, we have $\boldsymbol{\mu}_3 = (3, 3)$, $q_3 = 0.9$, and $\boldsymbol{\beta}_3^0 = (-1, -5, -2, -1, 1)$.

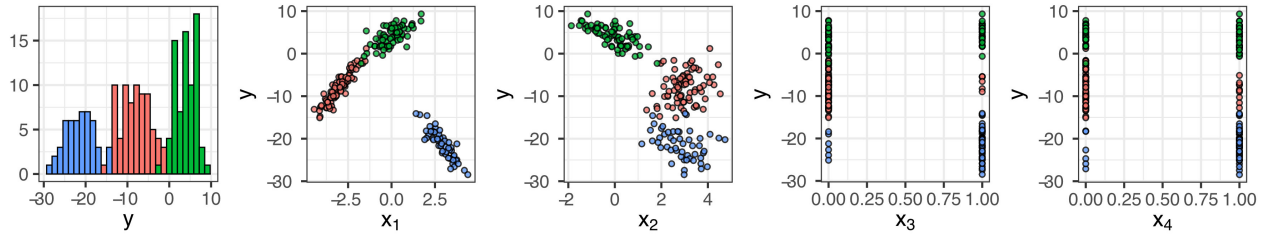


Figure S5: Simulated data. Left to right: histogram of the response variable $y_i$, scatterplots of $x_{i1}$ and $x_{i2}$ (continuous covariates) and scatterplots of $x_{i3}$ and $x_{i4}$ (discrete covariates) versus the response variable. Different colors represent the three different groups from which the data have been generated.

Figure S5 shows the simulated dataset. Three separate groups are clear for this figure, looking at the responses and the covariates. We have fitted model (2)-(4), with intensity given by a NGG process, when

$$f(\mathbf{y}_j^* \mid \mathbf{x}_j^*, \boldsymbol{\theta}_j^*) = \prod_{i \in A_j} \phi(y_i; \mathbf{x}_i^\intercal \boldsymbol{\beta}_j, \sigma_j^2),$$

and $\phi(y_i; \mathbf{x}_i^\intercal \boldsymbol{\beta}_j, \sigma_j^2)$ is the univariate Gaussian density with mean $\mathbf{x}_i^\intercal \boldsymbol{\beta}_j$ and variance $\sigma_j^2$. We include the whole vectors of $\mathbf{x}_i$ in the similarity and assume the cohesion function of the NGG process with $\kappa = 0.3$, $\sigma = 0.2$, so that the prior number of clusters without covariate effect ($g \equiv 1$) has mean equal to 5.9 and variance equal to 7.7. Moreover, the base measure $P_0$ on $\mathbb{R}^p \times \mathbb{R}_+$ is $\text{N}_p\left(\mathbf{0}, \sigma^2/\kappa_0 \mathbb{I}_{5\times 5}\right) \times \text{IG}(a, b)$ with $\kappa_0 = 0.01$, $(a, b) = (2, 1)$, and $\text{IG}(a, b)$ denotes the inverse-gamma distribution with mean $b/(a-1)$.

We run the algorithm described in Web Appendix C to obtain 5 000 final iterations, after a burnin of 10 000. We compute the posterior estimated partition as the one $\rho_n$ minimizing the posterior expectation of the variation of information loss function with equal misclassification costs [Wade and Ghahramani, 2018, Rastelli and Friel, 2018]. When classifying the datapoints according to these cluster estimates, we found that misclassification rates are 1%, 2.5% and 8.5% for $g_C$, $g_A$ ($\alpha = 1$), and $g \equiv 1$, respectively, when $\lambda = 0.5$.

We have also computed posterior predictive distributions as explained at the end of Section 2 in the manuscript. Figure S6 reports the posterior predictive distribution for a *new* individual, with the same covariate vector as the first subject in the sample. It is clear from the figure that the prediction is more precise under $g_A$ and $g_C$. In fact, when we do not include covariate information in the prior for the random partition, the posterior predictive density is not able to distinguish to which of the three groups the item belongs and the three peaks have approximately the same height. In contrast, when $g_A$ or $g_C$ are included in the prior, the posterior predictive density exhibits one main peak, so that covariate information helps, in this case, in selecting the right group the observation should be assigned to. This is also confirmed by the misclassification rates we have reported above. Figures S7 and S8 report the cluster estimate under $g_C$ and $g \equiv 1$ (no covariates in the prior), respectively. It is clear also from these plots that the inclusion of covariates in the prior specification improves the clustering performance of the model.

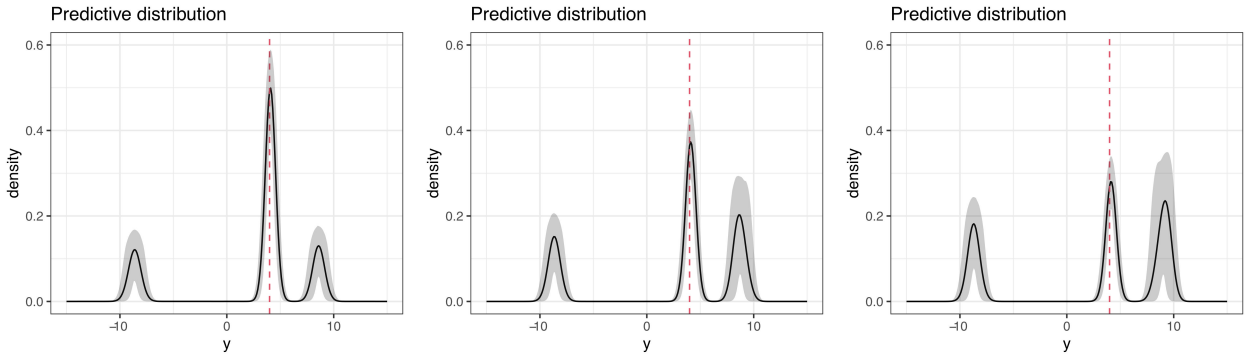


Figure S6: Posterior predictive distribution of $Y_1$, i.e., for a *new* individual with the same covariate values as the first individual in the sample, under $g_C$ (left), $g_A$ (center) and $g \equiv 1$ (right). The dashed red vertical lines denote the true value. Black lines denote the posterior mean of the density, the shaded areas denote 90% credibility band, based on the MCMC sample.

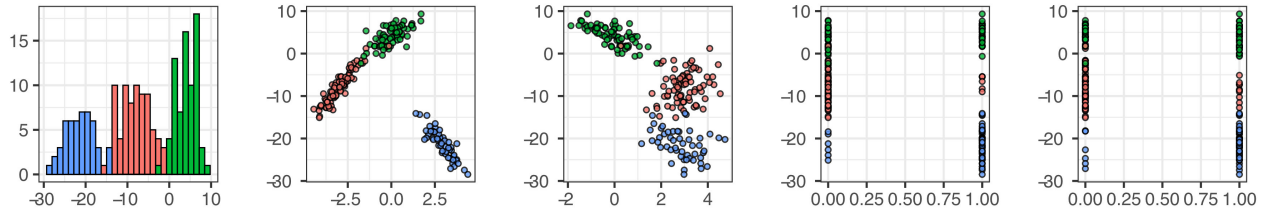

Figure S7: Simulated data. Left to right: histogram of the response variable $y_i$, scatterplots of $x_{i1}$ and $x_{i2}$ (continuous covariates) and scatterplots of $x_{i3}$ and $x_{i4}$ (discrete covariates) versus the response variable. Different colors represent the different clusters estimated with the PPMx-mixt model and $g_C$ is the similarity function in the prior.

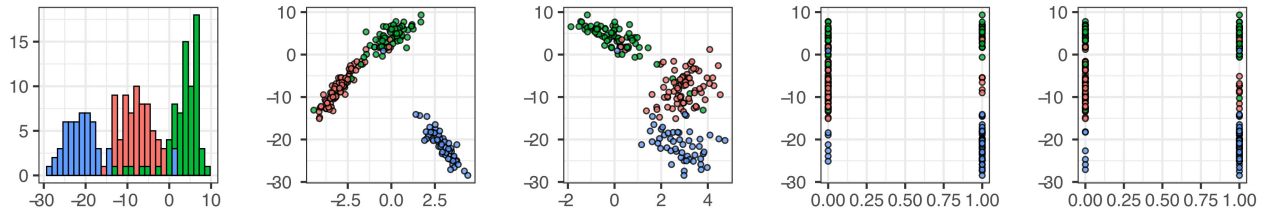

Figure S8: Simulated data. Left to right: histogram of the response variable $y_i$, scatterplots of $x_{i1}$ and $x_{i2}$ (continuous covariates) and scatterplots of $x_{i3}$ and $x_{i4}$ (discrete covariates) versus the response variable. Different colors represent the different clusters estimated with the PPM prior without covariate in the prior ($g \equiv 1$).

# Web Appendix F: Comparison to alternative models

We fit model (2)-(4), with intensity given by an NGG process, in the regression context, to the same simulated dataset as in Müller et al. [2011], Section 5.2. The simulation "truth" consists of 12 different distributions, corresponding to different covariate settings (see Figure 1 of that paper). The original dataset contains 1,000 data with three covariates $\mathbf{x}_i := (x_{i1}, x_{i2}, x_{i3})$ for each item $i$, where $x_{i1} \in \{-1, 0, 1\}$ and $x_{i2}$ and $x_{i3}$ are binary. We adopt the conditional distribution of data in cluster $A_j$ (see (2)) as

$$f(\mathbf{y}_j^* \mid \mathbf{x}_j^*, \boldsymbol{\theta}_j^*) = \prod_{i \in A_j} \phi(y_i; \mathbf{x}_i^\intercal \boldsymbol{\beta}_j, \sigma_j^2),$$

where $\phi(y_i; \mathbf{x}_i^\intercal \boldsymbol{\beta}_j, \sigma_j^2)$ is the univariate Gaussian density with mean $\mathbf{x}_i^\intercal \boldsymbol{\beta}_j$ and variance $\sigma_j^2$. The linear term includes the intercept. We assume the prior as in Section 2, with similarity function $g_C$ and distance $d(\mathbf{x}_1, \mathbf{x}_2)$ as described in Section 3. The prior $P_0$ of the cluster-specific parameters is a normal-inverse-gamma distribution, with

$$(\beta_j, \sigma_j^2) \sim \mathrm{N}_4(\boldsymbol{\beta}_0, \sigma_j^2/\kappa_0 \mathbb{I}_{4\times 4}) \times \mathrm{IG}(\sigma_j^2; a_0, b_0).$$

To replicate tests in Müller et al. [2011] and Bianchini et al. [2020] a total of $M = 100$ datasets of size 200 were generated by randomly subsampling 200 out of the 1,000 available observations. We compute the root MSE (over the 100 datasets) for estimating $\mathbb{E}(Y \mid x_1; x_2; x_3)$ for each of the 12 covariate combinations [see details in Müller et al., 2011]. We fix $P_0$ via the empirical Bayes approach using the responses' overall sample mean and variance. Table S1 displays root MSE under our model when the hyperparameters in the cohesion function are fixed as $\sigma = 0.2$, $\kappa = 0.001$; this corresponds assuming that, without covariate effect (i.e. when $g \equiv 1$), the prior number of clusters has mean equal to 3 and variance equal to 5.8. Parameter $\lambda$ in $g_C$ has been fixed to 0.041, with $\varepsilon^*$ in Section 3 equal 0.1; the table reports also the last two columns in Table 7 in Bianchini et al. [2020]. We consider two distinct specifications of the PPMx-mixt model, one where the distance is defined as a weighted sum of the Mahalanobis and the Hamming distances (PPMx-mixt$^a$), and a second one where the Mahalanobis distance has been previously transformed through $q : [0, +\infty) \to [0, 1)$ with $q(x) = \arctan(x)$ (PPMx-mixt$^b$). This second alternative has been tested since the magnitude of the Mahalanobis distances, ranging between 0 and $+\infty$, might overwhelm the normalized Hamming distances. The function $q(x)$ maps the Mahalanobis distance in (8) into a finite interval.

| $x_1$ | $x_2$ | $x_3$ | PPMx-mixt$^a$ | PPMx-mixt$^b$ | DPP | PPMx |
|---|---|---|---|---|---|---|
| -1 | 0 | 0 | 14.0 | 14.6 | 6.1 | 7.9 |
| 0 | 0 | 0 | 2.8 | 3.2 | 6.7 | 3.9 |
| 1 | 0 | 0 | 3.1 | 2.8 | 7.2 | 2.8 |
| -1 | 1 | 0 | 13.9 | 14.8 | 6.5 | 5.4 |
| 0 | 1 | 0 | 3.3 | 2.6 | 6.5 | 4.6 |
| 1 | 1 | 0 | 8.7 | 8.1 | 6.8 | 4.0 |
| -1 | 0 | 1 | 5.9 | 6.0 | 6.8 | 6.1 |
| 0 | 0 | 1 | 2.2 | 2.1 | 6.1 | 4.2 |
| 1 | 0 | 1 | 2.2 | 2.2 | 5.7 | 4.5 |
| -1 | 1 | 1 | 4.8 | 4.8 | 5.9 | 9.5 |
| 0 | 1 | 1 | 2.6 | 2.7 | 6.6 | 8.3 |
| 1 | 1 | 1 | 2.4 | 2.3 | 5.8 | 6.2 |
| **avg** | | | **5.5** | **5.5** | **6.4** | **5.6** |

Table S1: Root MSE for estimating $\mathbb{E}(Y \mid x_1, x_2, x_3)$ for 12 combinations of covariates $(x_1, x_2, x_3)$ and $PPM_x$ and DPP as competing models of reference; the last two columns are those in Table 7 of Bianchini et al., 2020. Two specifications of the PPMx-mixt model, with PPMx-mixt$^b$ corresponding to the model where the Mahalanobis distance has been previously transformed through the arctan function.

The PPMx-mixt shows performance comparable to the competitors DPP and PPMx in Bianchini et al. [2020] and Müller et al. [2011], respectively. Note that our model gives more variability among different combinations of covariates.

# Web Appendix G: Additional comments and comparison with alternative models for the AVIS application

In this section, we provide additional insight into the application as well as a comparison with alternative model-based clustering procedures for the AVIS application.

Figure S9 shows the barplots of normalized recurrences grouped by cluster, which are defined as:

$$\frac{\text{number of donors in the cluster who perform exactly } k \text{ recurrences}}{\text{total number of donors in the cluster}},$$

for $k = 1, \ldots,$ the maximum number of recurrences in the cluster. Figure S10 shows the histograms of empirical yearly rates of donation per cluster, i.e., the total number of donations divided by the number of years under observation for each donor. The two figures highlight the differences in terms of these two variables among the five clusters. Figure S11 shows the posterior predictive density functions at the first gap time for different clusters in the optimal partitions. In particular, the densities are evaluated for *new* donors with covariates equal to the empirical mean (in case of a continuous covariate) or mode (for a discrete covariate). The posterior predictive densities are plotted for *new* female and male donors in dashed and continuous lines, respectively. The plot shows that for the first cluster, denoted in red, which corresponds to the largest cluster, the posterior predictive densities have larger dispersion. Further, the last two clusters, blue and purple, have similar density functions, with the latter slightly shifted to higher values of the log-gap times.

In the first comparison we study how the inclusion of $g_C$ in the prior affects posterior predictive inference, we consider a cross-validation approach where we subsample 50 different training subsets containing 90% of the donors. We compute the associated posterior for each training subset and use the remaining donors as the testing set for prediction. We adopt the same model specification given in Section 4.2 of the manuscript with $\lambda = 0.1$ and $\sigma = 0.15$. Posterior predictive inference has been computed as explained at the end of Section 2 of the manuscript. We also compare with the case $g \equiv 1$. or each training subsample, we ran the Gibbs sampler for 3 000 iterations, of which 2 000 were discarded as burn-in iterations. The root mean square error (rMSE) between observed and predicted log-gap times decreases from 2.242 – obtained with $g \equiv 1$ – to 1.825 obtained under the PPMx-mixt model with $g_C$.

We have also compared the cluster estimates discussed in the manuscript, with optimal hyperparameters ($\lambda = 0.1$, $\sigma = 0.15$) for $g_C$, with competitor models: $(i)$ $g \equiv 1$ and $\sigma = 0.15$ (no effect of covariates in the prior), $(ii)$ $\lambda = 0.1$ and $\sigma = 0.001$ for $g_C$, i.e., cohesion function corresponding to the Dirichlet process and $(iii)$ the original PPMx in Müller et al. [2011]. See the , Tables S2 $(i)$, S3 $(ii)$ and S4 $(iii)$ of this supplemental document. Specifically, these tables show the joint empirical frequencies of individuals in the two estimated partitions. From the comparison, it is clear that our best model ($\lambda = 0.1$, $\sigma = 0.15$) mitigates the rich-get-richer property of other PPMx-mixt specifications. Table S5 shows empirical summaries of the covariates (included in the prior) within each estimated cluster under the PPMx prior. Tables S4 and S5 point out that subjects in Clusters 1 and 2 under the PPMx make Cluster 1 under our model. See also the cluster-specific covariate description of Table S5. Similarly, Cluster 5 under our model mainly corresponds to individuals in Cluster 8 under the PPMx prior. Cluster 4 in Table S5 is similar to Cluster 6 under our model. As a final remark, note that, under the original PPMx, the number of estimated clusters is larger than ours, which implies, in general, that it is more difficult to interpret them. As a consequence, the estimated clusters under the original PPMx prior are not more clearly interpretable in terms of covariates than ours.

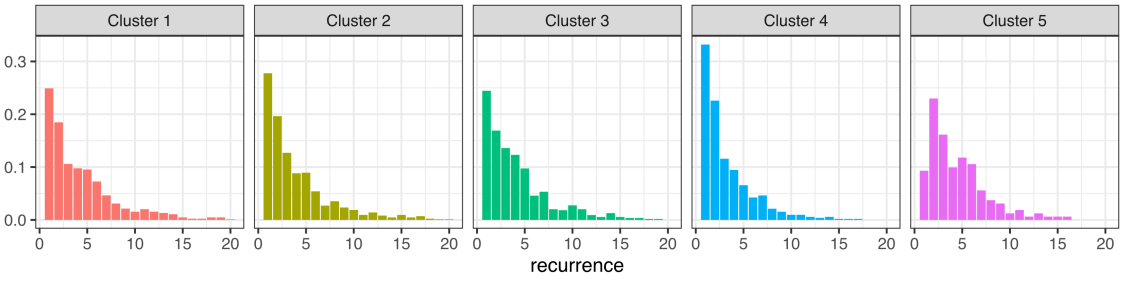


Figure S9: Empirical distributions of the donation recurrences per cluster.

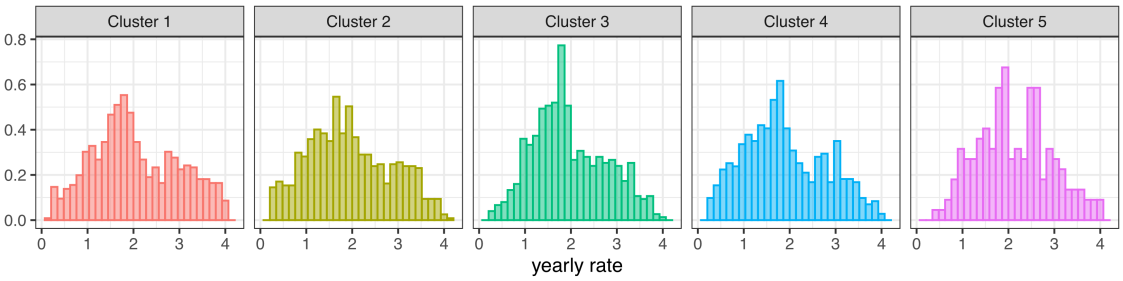


Figure S10: Histograms of yearly rates of donation.

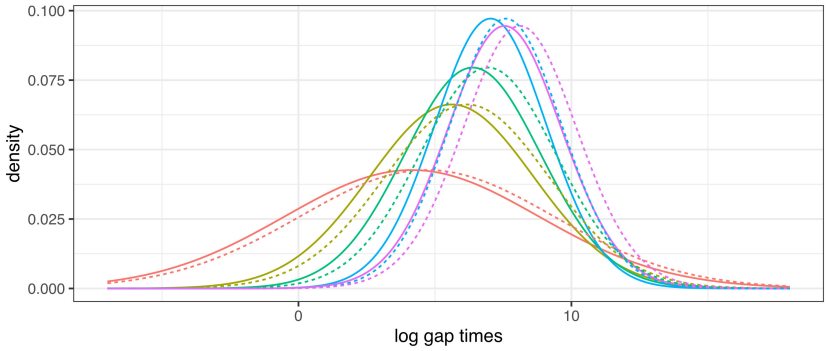


Figure S11: Cluster-specific posterior predictive distribution for the first recurrent donation. Different colors correspond to different cluster. Continuous lines refer to male donors, dashed lines to female donors.

| | | $g \equiv 1,\ \lambda = 0,\ \sigma = 0.15$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | |
| $\lambda = 0.1,\ \sigma = 0.15$ | Cluster 1 | 839 | 0 | 0 | 0 | 0 | 839 |
| | Cluster 2 | 284 | 486 | 80 | 0 | 0 | 850 |
| | Cluster 3 | 0 | 188 | 348 | 8 | 0 | 544 |
| | Cluster 4 | 0 | 0 | 53 | 443 | 22 | 518 |
| | Cluster 5 | 0 | 0 | 0 | 30 | 131 | 161 |
| | | 1123 | 674 | 481 | 481 | 153 | |

Table S2: Comparison of the cluster estimates under our model with optimal hyperparameters (by rows) against the model with $g \equiv 1$ and $\sigma = 0.15$ (by columns).

| | | $\lambda = 0.1,\ \sigma = 0.001$ | | | |
|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | Cluster 3 | |
| $\lambda = 0.1,\ \sigma = 0.15$ | Cluster 1 | 766 | 73 | 0 | 839 |
| | Cluster 2 | 19 | 696 | 135 | 850 |
| | Cluster 3 | 0 | 211 | 333 | 544 |
| | Cluster 4 | 0 | 0 | 518 | 518 |
| | Cluster 5 | 0 | 0 | 161 | 161 |
| | | 785 | 980 | 1147 | |

Table S3: Comparison of the cluster estimates under our model with optimal hyperparameters (by rows) against our model with $g_C$, $\lambda = 0.1$ and $\sigma = 0.001$ (by columns).

| | | PPMx | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Cl. 1 | Cl. 2 | Cl. 3 | Cl. 4 | Cl. 5 | Cl. 6 | Cl. 7 | Cl. 8 | |
| $\lambda = 0.1$, $\sigma = 0.15$ | Cluster 1 | 398 | 441 | 0 | 0 | 0 | 0 | 0 | 0 | 839 |
| | Cluster 2 | 303 | 0 | 342 | 198 | 4 | 3 | 0 | 0 | 850 |
| | Cluster 3 | 0 | 0 | 26 | 240 | 267 | 11 | 0 | 0 | 544 |
| | Cluster 4 | 0 | 0 | 0 | 10 | 21 | 456 | 15 | 16 | 518 |
| | Cluster 5 | 3 | 0 | 0 | 0 | 0 | 10 | 42 | 106 | 161 |
| | | 704 | 441 | 368 | 448 | 292 | 480 | 57 | 122 | |

Table S4: Comparison of the cluster estimates under our model with optimal hyperparameters (by rows) against the PPMx (by columns).

| | Size | Age | BMI | Gender | Blood Type | | | | RH | Smoke | | No. Donations | log gap-time |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Female | A | B | AB | 0 | + | yes | | mean (sd) | mean (sd) |
| Cl. 1 | 704 | 39.65 | 24.31 | 31.11% | 39.77% | 12.93% | 4.12% | 43.18% | 90.20% | 35.51% | M | 5.04 (4.01) | 4.94 (0.47) |
| | | | | | | | | | | | F | 2.53 (1.85) | 5.52 (0.39) |
| Cl. 2 | 441 | 51.29 | 24.38 | 42.63% | 40.36% | 12.70% | 2.04% | 44.90% | 88.44% | 35.37% | M | 4.89 (4.06) | 4.92 (0.50) |
| | | | | | | | | | | | F | 3.16 (2.36) | 5.42 (0.41) |
| Cl. 3 | 638 | 33.24 | 24.36 | 27.72% | 38.59% | 12.23% | 1.63% | 47.55% | 87.77% | 35.87% | M | 5.14 (4.04) | 4.93 (0.47) |
| | | | | | | | | | | | F | 2.70 (1.76) | 5.52 (0.38) |
| Cl. 4 | 448 | 29.46 | 23.80 | 33.04% | 35.71% | 10.27% | 3.79% | 50.22% | 83.71% | 29.46% | M | 2.70 (1.76) | 5.52 (0.38) |
| | | | | | | | | | | | F | 2.18 (1.59) | 5.61 (0.48) |
| Cl. 5 | 292 | 26.89 | 23.60 | 33.90% | 33.22% | 12.33% | 4.45% | 50.00% | 89.38% | 31.51% | M | 4.90 (3.80) | 4.98 (0.48) |
| | | | | | | | | | | | F | 2.61 (1.73) | 5.50 (0.35) |
| Cl. 6 | 480 | 23.01 | 23.08 | 29.38% | 39.79% | 12.08% | 4.79% | 43.33% | 85.42% | 29.79% | M | 3.65 (3.09) | 5.05 (0.54) |
| | | | | | | | | | | | F | 2.43 (1.60) | 5.57 (0.41) |
| Cl. 7 | 57 | 20.25 | 23.84 | 1.75% | 28.07% | 15.79% | 17.54% | 38.60% | 71.93% | 35.09% | M | 2.88 (2.22) | 5.24 (0.66) |
| | | | | | | | | | | | F | 2.00 (0.00) | 5.99 (0.91) |
| Cl. 8 | 122 | 20.49 | 23.42 | 13.11% | 37.70% | 14.75% | 5.74% | 41.80% | 74.59% | 22.13% | M | 5.16 (3.29) | 4.92 (0.41) |
| | | | | | | | | | | | F | 4.06 (2.59) | 5.38 (0.25) |
| All | 2912 | 33.83 | 23.93 | 31.39% | 38.11% | 12.33% | 3.91% | 45.64% | 86.74% | 32.69% | M | 4.55 (3.76) | 4.97 (0.51) |
| | | | | | | | | | | | F | 2.64 (1.91) | 5.50 (0.41) |

Table S5: Empirical summaries of the covariates within each estimated cluster for the blood donation application under the PPMx.

# References


S. Favaro and Y. W. Teh. MCMC for normalized random measure mixture models. *Stat. Sci.*, 28:335–359, 2013.

A. Lijoi and I. Prünster. Models beyond the dirichlet process. In Müller Hjort, Holmes and Walker, editors, *Bayesian nonparametrics*, page 3. Cambridge, 2010.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.*, 9: 249–265, 2000.

S. Wade and Z. Ghahramani. Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). *Bayesian Anal.*, 13(2):559 – 626, 2018.

R. Rastelli and N. Friel. Optimal Bayesian estimators for latent variable cluster models. *Stat. Comput.*, 28(6): 1169–1186, Nov 2018.

P. Müller, F. A. Quintana, and G. A. Rosner. A product partition model with regression on covariates. *J. Comput. Graph. Stat*, 20:260–278, 2011.

I. Bianchini, A. Guglielmi, and F. A. Quintana. Determinantal Point Process Mixtures Via Spectral Density Approach. *Bayesian Anal.*, 15(1):187 – 214, 2020.