Revista Colombiana de Estadística
Special Issue: Colombian and Brazil Graduate Programs
December 2025, volume 48, issue 3, pp. 269–299
http://doi.org/10.15446/rce.v48n3.122305

A Multilevel Nonparametric Bayesian Model

Un Modelo Bayesiano No Paramétrico Multinivel

Laura Camila Cruz De Paula[1,a], Juan Sosa[1,b]

[1]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

This work presents the development of a multilevel Bayesian nonparametric model that allows for the estimation of linear relationships in heterogeneous data sets, while simultaneously identifying clusters without the need to specify the number of groups in advance. The study includes the mathematical development of the model using the Chinese Restaurant Process and the implementation of algorithms for its fitting. The results obtained from real data show that the model performs well in both clustering data and characterizing linear relationships, achieving results comparable and even better to those obtained by traditional parametric methods.

***Keywords*:** Chinese Restaurant process; Clustering; Dirichlet process; Linear regression; Nonparametric Bayesian model.

Resumen

Este trabajo presenta el desarrollo de un modelo Bayesiano no paramétrico multinivel diseñado para estimar relaciones lineales en conjuntos de datos heterogéneos, mientras identifica simultáneamente conglomerados sin requerir la especificación previa del número de grupos. El estudio incluye la formulación matemática del modelo utilizando el Proceso del Restaurante Chino, así como la implementación de algoritmos para su ajuste. Los resultados obtenidos con datos reales muestran que el modelo tiene un buen desempeño tanto en la agrupación de los datos como en la caracterización de las relaciones lineales, alcanzando resultados comparables e incluso superiores a los obtenidos mediante métodos paramétricos tradicionales.

***Palabras clave*:** Agrupamiento; Modelo Bayesiano no paramétrico; Proceso de Dirichlet; Proceso del Restaurante Chino; Regresión lineal.

[a]M.Sc. E-mail: laccruzpa@unal.edu.co
[b]Ph.D. E-mail: jcsosam@unal.edu.co

1. Introduction

Nonparametric Bayesian modeling provides flexible inference for complex data by relaxing the assumption of a fixed functional form and a predetermined number of parameters. Here, *nonparametric* does not mean parameter free; rather, it allows infinite-dimensional parameter spaces, including random probability measures, families of distributions, mean functions, and feature allocations. We use the term *multilevel* to denote a hierarchical structure that represents clustered data. This flexibility is especially valuable in regression, where the goal is to relate a response to predictors through location and scale. In many applications, the error distribution and the regression parameters vary with predictors and cannot be adequately represented by finite-dimensional specifications (Quintana et al., 2022). Ignoring this variation can distort inference and lead to misleading conclusions.

These challenges motivate nonparametric Bayesian regression models that place nonparametric priors on the regression mean, on the residual distribution, or on both, thereby yielding fully nonparametric formulations (Müeller et al., 2018). A central example is Gaussian mixture regression (Müller et al., 1996), which jointly models the mean and the variance nonparametrically while treating predictors as fixed. This specification parallels the multilevel parametric Bayesian framework of Sosa & Aristizabal (2022), but it avoids constraints on parameter dimensionality, an advantage when latent clustering shapes the relationship between the response and the predictors.

Building on this literature, we develop a multilevel nonparametric Bayesian model that uses a Dirichlet process (DP) prior (Müller et al., 2015; Müeller et al., 2018; Jara, 2017; Xuan et al., 2019), with particular emphasis on its Chinese Restaurant Process (CRP) representation (Navarro & Perfors, 2023; Bouchard-Côté, 2011). To handle mixed predictors without partitioning the predictor space, the model deliberately avoids partition-based strategies, directly addressing limitations noted by Quintana et al. (2022).

This paper makes three contributions. First, it combines hierarchical modeling with DP–based nonparametric components to capture uncertainty in group structure without prespecifying the number of mixture components, allowing model complexity to adapt to the data. Second, it targets settings with pronounced heterogeneity and unknown latent structure, common with mixed data types, outliers, and multiple scales of variation, enabling unsupervised pattern discovery under weaker structural assumptions than parametric models. Third, it provides an expository development of the key building blocks, from Gaussian mixtures and DPs to Markov chain Monte Carlo (Gamerman & Lopes, 2006), together with a modular implementation (publicly available in a GitHub repository `https://github.com/Camilacruzdepaula/Bayesian_Non_Parametric_Model`) that facilitates extensions to more complex models. Applications to data from industry demonstrate the versatility and external validity of the approach. The model automatically learns latent groups and estimates group-specific regression relationships, providing a robust and flexible tool for evidence-based decision making and a competitive alternative to standard approaches.

The paper is organized as follows. Section 2 reviews background and applications of nonparametric Bayesian regression. Section 3 develops the theoretical foundations of the DP and related extensions. Section 4 presents the proposed multilevel model, its mathematical formulation, and the MCMC algorithms used for posterior inference. Section 5 assesses its performance on real data. Finally, Section 6 summarizes the main findings and outlines directions for future research.

2. Related Work

Nonparametric Bayesian statistics originated with the DP (DP) of Ferguson (1973) and has expanded rapidly over the past three decades, driven by advances in computation. Current research spans density estimation, clustering, regression, fixed-effects models, and survival analysis. See Müller et al. (2015); Müller & Mitra (2013); Dunson (2010) for comprehensive reviews.

In regression, we consider models of the form $y_i = f(x_i) + \varepsilon_i$, with $i = 1, \ldots, n$. Parametric Bayesian approaches impose distributional assumptions on ε_i and on the functional relationship between the response and the covariates, thereby restricting the form of f and the dimension of the parameter space. In contrast, (Müller & Mitra, 2013) identify three nonparametric strategies: modeling the residual distribution nonparametrically, modeling the mean function nonparametrically, and fully nonparametric regression.

For nonparametric modeling of residuals, one assumes $\varepsilon_i \sim G$, where G is given a nonparametric prior, often DP-based or Pólya tree–based (Walker & Mallick, 1999; Hanson & Johnson, 2002; Müeller et al., 2018). Applications include modeling white blood cell counts in leukemia and abortion occurrence, where nonparametric priors on the error distribution yield fits comparable to parametric models while producing smooth posterior error distributions. Related developments include DP models for interval-censored data (Hanson & Johnson, 2004) and semiparametric logistic regression (Schörgendorfer et al., 2013). See Müeller et al. (2018) for further examples.

For nonparametric modeling of the mean, a common specification is a basis expansion of the form $f(x) = \sum_h d_h \varphi_h(x)$, with applications in fields such as astrophysics and signal processing (Barnes III et al., 2003; Clyde & George, 2000). These estimators are computationally competitive and robust to outliers. An alternative is the Gaussian process prior, which places a distribution over functions, $(f(x_1), \ldots, f(x_n)) \sim \mathsf{N}_n(0, \mathbf{S})$, with covariance matrix $\mathbf{S}$. See Williams & Rasmussen (2006) for further properties and applications.

Fully nonparametric regression can be achieved through the dependent Dirichlet process (DDP; MacEachern (1999)). A representative construction is $y_i \sim F = \sum_h w_h \, \mathsf{N}(m_h, \sigma^2)$, where the mixing distribution is assigned a DP prior and covariate dependence is introduced via $m_h(\mathbf{x}_i)$, often modeled using a Gaussian process or covariate-dependent stick-breaking weights (Xu et al., 2016). Applications include music segmentation, image processing, and finance (Foti & Williamson, 2013). Endogenous predictors can be incorporated by modeling $d_i = (y_i, \mathbf{x}_i)$ with

a multivariate covariate-dependent Gaussian mixture and analyzing the resulting conditional distributions (Müller et al., 1996). DDP formulations also permit joint inference on mean and variance, while avoiding multiplicative structures and ad hoc partitions for mixed discrete and continuous covariates (Quintana et al., 2022).

A related parametric approach with clustering in the response is proposed by Sosa & Aristizabal (2022), which requires a finite number of components. Building on Quintana et al. (2022), our work removes this restriction by modeling both the mean and the variance nonparametrically, avoiding partition-based fitting strategies for mixed data, and learning clusters without prespecifying their number. These advances motivate the multilevel nonparametric model developed in the remainder of the paper.

3. Bayesian Nonparametrics Foundations

This section develops the mathematical foundations of the Dirichlet process (DP) using the Gaussian mixture model as a canonical example in nonparametric Bayesian statistics. First, we show how the DP induces mixture models with an unknown, potentially unbounded number of components (Gelman et al., 2013). Next, we present the formal definition of the process and its two standard representations, the stick-breaking construction and the Chinese restaurant process (Navarro & Perfors, 2023; Bouchard-Côté, 2011).

3.1. Gaussian Mixture Models

Normal mixture models, or Gaussian mixture models (GMMs; see, for example, Bishop, 2006), are statistical models for data whose distribution is represented as a combination of multiple Gaussian components. Consider an independent sample of random vectors $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ and K Gaussian density functions:

$$p(\mathbf{y}_i \mid \{\pi_k\}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}) = \sum_{k=1}^{K} \pi_k \, \mathsf{N}_d(\mathbf{y}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,,$$

where π_k denotes the mixture weights, satisfying $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$, and $\mathsf{N}_d(\mathbf{y}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the d-dimensional Gaussian density with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. Additionally, introduce the latent variable $z_i \in \{1, 2, \ldots, K\}$ that indicates the component k to which observation $\mathbf{y}_i$ belongs, with allocation probabilities $p(z_i = k \mid \boldsymbol{\pi}) = \pi_k$. Figure 1 (taken from Kamper, 2013) presents the directed acyclic graph (DAG) of the model, illustrating not only the Gaussian mixture structure just described but also a joint prior distribution for $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\Lambda}_0$ denotes the vector of hyperparameters. Finally, the weights $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ are treated as random and assigned a prior distribution governed by a concentration parameter $\boldsymbol{\alpha}$.

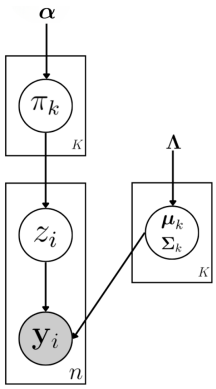


FIGURE 1: DAG for the GMM.

Consider $K > 2$. In this case, the probability vector is $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$. In this case, the Dirichlet distribution is used as a multivariate extension of the Beta distribution. Following Frigyik et al. (2010), the density of a random vector $\boldsymbol{\pi}$ following a Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is given by

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1},$$

with $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \in [0, 1]$, and $\alpha_0 = \sum_{k=1}^{K} \alpha_k$. As with the Beta distribution, the Dirichlet distribution can take a wide variety of shapes depending on the values of the parameters in $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$.

Once a prior distribution for the probability vector $\boldsymbol{\pi}$ has been specified, the distribution of the assignment vector $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, with $z_i \in \{1, 2, \ldots, K\}$, is given by

$$p(\mathbf{z} \mid \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{n_k}, \tag{1}$$

where n_k denotes the number of observations assigned to component k. As in the Beta case, the Dirichlet distribution is chosen for its flexibility and because it is conjugate to the Categorical (or Multinomial) distribution.

Given the prior distribution of $\mathbf{z}$ in (1), the posterior distribution is obtained by analytically integrating the joint density $p(\mathcal{Y}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}, \mathbf{z}, \boldsymbol{\pi})$ with respect to the parameters $\{\boldsymbol{\mu}_k\}$, $\{\boldsymbol{\Sigma}_k\}$, and $\boldsymbol{\pi}$. Integrating out $\boldsymbol{\pi}$ induces dependence among all z_i (Murphy, 2012); consequently, the conditional posterior probability of z_i is given by

$$p\left(z_i = k \mid \mathbf{z}_{-i}, \mathcal{Y}, \boldsymbol{\alpha}, \boldsymbol{\Lambda}_0\right) \propto p\left(\mathbf{y}_i \mid \mathcal{Y}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\Lambda}_0\right) p\left(z_i = k \mid \mathbf{z}_{-i}, \boldsymbol{\alpha}\right), \tag{2}$$

where $\mathbf{z}_{-i}$ and $\mathcal{Y}_{-i}$ denote, respectively, the vector $\mathbf{z}$ and the collection of observations $\mathcal{Y}$ with the ith element removed.

The second term in Equation (2), corresponding to the prior contribution for z_i, is computed as in Kamper (2013). With $\alpha_k = \alpha/K$ so that $\alpha_0 = \sum_{k=1}^{K} \alpha_k = \alpha$, we obtain

$$p\left(z_i = k \mid \mathbf{z}_{-i}, \boldsymbol{\alpha}\right) = \frac{n_k - 1 + \alpha/K}{n - 1 + \alpha}, \tag{3}$$

where n_k is the number of observations in component k. The first term in (2), the data likelihood contribution, can be written as $p\left(\mathbf{y}_i \mid \mathcal{Y}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\Lambda}_0\right) = p\left(\mathbf{y}_i \mid \mathcal{Y}_{-i,k}, \boldsymbol{\Lambda}_0\right)$, where $\mathcal{Y}_{-i,k}$ is the set of observations in component k excluding $\mathbf{y}_i$. This probability satisfies $p\left(\mathbf{y}_i \mid \mathcal{Y}_{-i,k}, \boldsymbol{\Lambda}_0\right) = p\left(\mathcal{Y}_k \mid \boldsymbol{\Lambda}_0\right) / p\left(\mathcal{Y}_{-i,k} \mid \boldsymbol{\Lambda}_0\right)$, with

$$p\left(\mathcal{Y}_k \mid \boldsymbol{\Lambda}_0\right) = \int\int p\left(\mathcal{Y}_k \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) p\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \boldsymbol{\Lambda}_0\right) \, \mathrm{d}\boldsymbol{\mu}_k \, \mathrm{d}\boldsymbol{\Sigma}_k.$$

This marginal likelihood coincides with the posterior predictive distribution; see Kamper (2013) for details.

3.2. Nonparametric Bayesian Mixture Model

Another strategy for modeling mixtures is to use nonparametric Gaussian mixture models. In contrast to parametric mixtures, which require the number of components K to be specified a priori, nonparametric formulations learn the number of groups directly from the data. Figure 2 (taken from Kamper, 2013) presents the corresponding DAG, which closely resembles the parametric version but allows for a potentially infinite number of components. By allowing for an infinite number of components, the Dirichlet process (DP) naturally arises as an alternative prior for the allocation vector $\mathbf{z}$. This process, introduced by Ferguson (1973), is fundamentally a distribution over probability measures and has become one of the most widely used tools in nonparametric Bayesian statistics. In what follows, we develop the mathematical foundations of the process.

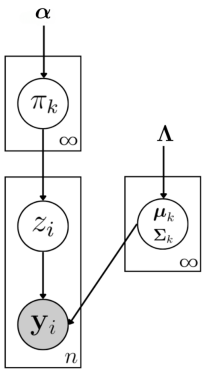


FIGURE 2: DAG for the nonparametric GMM.

3.2.1. Dirichlet Process

Let $(\Omega, \mathcal{A})$ be a measurable space, where Ω is the sample space and $\mathcal{A}$ is a σ-algebra of subsets of Ω. Let G be a random probability measure on this space, and suppose that G_0 is a probability measure on Ω and α is a positive real constant.

We say that G follows a DP on $(\Omega, \mathcal{A})$, denoted $G \sim \mathsf{DP}(\alpha, G_0)$, if for any finite measurable partition $\{A_1, \ldots, A_K\}$ of Ω, with $A_k \in \mathcal{A}$, the following holds:

$$\big(G(A_1), \ldots, G(A_K)\big) \sim \mathsf{Dir}\big(\alpha G_0(A_1), \ldots, \alpha G_0(A_K)\big), \tag{4}$$

where G_0 is known as the base distribution (or base measure) and α is referred to as the concentration (or mass) parameter (Ferguson, 1973). One way to interpret the DP is as a *distribution over distributions*, a perspective made explicit by the constructive *stick-breaking* representation.

3.2.2. Properties

The properties of the DP follow directly from those of the Dirichlet distribution (Teh, 2010). Its expectation is

$$\mathsf{E}\big(G(A_i)\big) = \frac{\alpha G_0(A_i)}{\sum_{j=1}^{K} \alpha G_0(A_j)} = G_0(A_i),$$

so the base distribution G_0 is the mean of the random distribution G. Similarly, the variance is

$$\mathsf{Var}\big(G(A_i)\big) = \frac{G_0(A_i)\big(1 - G_0(A_i)\big)}{\alpha + 1}.$$

Thus, the concentration parameter α controls the variability of G around G_0: larger values of α imply that realizations of G tend to be closer to G_0 for any measurable set A. However, this does not imply $G \to G_0$, since draws from a DP are almost surely discrete, even when G_0 is continuous.

3.2.3. Posterior Distribution

Consider $\theta_1, \ldots, \theta_n$ as an independent sample from G, where $G \sim \mathsf{DP}(\alpha, G_0)$. Let $A_1, \ldots, A_K$ be a finite measurable partition of the sample space, and define $n_j = \#\{i : \theta_i \in A_j\}$ as the number of observations falling in set A_j. The probability of θ_i given G can be written as

$$p(\theta_i \mid G) = \sum_{j=1}^{K} G(A_j)\, 1_{\{\theta_i \in A_j\}},$$

where $1_{\{\theta_i \in A_j\}}$ is the indicator function that equals 1 if $\theta_i \in A_j$ and 0 otherwise. The corresponding likelihood function is

$$p(\boldsymbol{\theta} \mid G) = \prod_{i=1}^{n} \sum_{j=1}^{K} G(A_j)\, 1_{\{\theta_i \in A_j\}}, \tag{5}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. Using (5), the posterior distribution of G given the observed values $\theta_1, \ldots, \theta_n$ is

$$p(G \mid \boldsymbol{\theta}) \propto p(\boldsymbol{\theta} \mid G)\, p(G) \propto \prod_{j=1}^{K} G(A_j)^{\, n_j + \alpha G_0(A_j) - 1}. \tag{6}$$

Equivalently, the indicator $1_{\{\theta_i \in A_j\}}$ may be denoted by δ_{ij}, with $\delta_{ij} = 1$ if $\theta_i \in A_j$ and $\delta_{ij} = 0$ otherwise.

The distribution obtained in (6) corresponds to a Dirichlet distribution. Therefore, the posterior distribution $p(G \mid \boldsymbol{\theta})$ is

$$\big(G(A_1), G(A_2), \ldots, G(A_K)\big) \mid \boldsymbol{\theta} \sim \mathsf{Dir}\big(n_1 + \alpha G_0(A_1), \ldots, n_K + \alpha G_0(A_K)\big). \quad (7)$$

That is, the posterior distribution $p(G \mid \boldsymbol{\theta})$ is again a DP, and can be written as

$$G \mid \boldsymbol{\theta} \sim \mathsf{DP}\left(\alpha + n,\ \frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right), \quad (8)$$

where δ_{θ_i} denotes a point mass at θ_i. Note that the updated base distribution is a weighted average of the prior base distribution G_0 and the empirical distribution $\frac{1}{n}\sum_{i=1}^{n} \delta_{\theta_i}$, with respective weights α and n.

3.2.4. Posterior Predictive Distribution

Consider again the random sample $\theta_1, \theta_2, \ldots, \theta_n$ drawn from G, where $G \sim \mathsf{DP}(\alpha, G_0)$. Suppose we wish to compute the distribution of θ_{n+1} given $\theta_1, \ldots, \theta_n$. Following Teh (2010), since $\theta_{n+1} \mid G, \theta_1, \ldots, \theta_n \sim G$, for any measurable set $A \subset \Omega$ we have

$$p\big(\theta_{n+1} \in A \mid \theta_1, \ldots, \theta_n\big) = \frac{\alpha\ G_0(A)}{\alpha + n} + \frac{\sum_{i=1}^{n} \delta_{\theta_i}(A)}{\alpha + n}.$$

Integrating out G yields

$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n \sim \frac{\alpha\ G_0}{\alpha + n} + \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}. \quad (9)$$

Thus, the predictive distribution of θ_{n+1} given $\theta_1, \ldots, \theta_n$ coincides with the updated base distribution of the posterior DP in (8).

3.2.5. Pólya Urn Model

The sequence of posterior predictive distributions for $\theta_1, \theta_2, \ldots$ is known as the Pólya urn model, in reference to the classical urn scheme that motivates its name (Theodoridis, 2020). The model describes filling an urn with balls of different colors. Initially, the urn is empty and each value in Ω corresponds to a distinct color. For the first draw, a color is sampled from the base distribution, i.e., $\theta_1 \sim G_0$, and a ball of that color is placed in the urn. For the second draw, there are two possibilities: a new color is sampled from G_0 with probability $\frac{\alpha}{\alpha+1}$, or the color of the existing ball is selected with probability $\frac{1}{\alpha+1}$. More generally, at step $n+1$ a new color $\theta_{n+1} \sim G_0$ is drawn with probability $\frac{\alpha}{\alpha+n}$, or one of the n existing balls is chosen uniformly at random and its color is copied with probability $\frac{n}{\alpha+n}$.

From the above construction, as n increases, the probability $\frac{n}{\alpha+n}$ of selecting an existing color from the urn grows relative to the probability $\frac{\alpha}{\alpha+n}$ of drawing a new color from G_0. It is also important to note that the probability of generating

a sequence of colors $\theta_1, \ldots, \theta_n$ is equal to the probability of generating the same multiset of colors in any other order, that is,

$$p(\theta_1, \ldots, \theta_n) = p(\theta_{\pi(1)}, \ldots, \theta_{\pi(n)}),$$

where $\pi \in \mathfrak{S}_n$ denotes a permutation of $\{1, 2, \ldots, n\}$. This exchangeability property allows the Pólya urn scheme to establish the existence of the DP via the following argument (Blackwell & MacQueen, 1973). For $n \geq 1$, the joint distribution of $\theta_1, \ldots, \theta_n$ can be written as

$$p(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}).$$

Since the sequence is infinitely exchangeable, de Finetti's theorem (Hoff, 2009) implies that there exists a random probability measure G such that, for any n,

$$p(\theta_1, \ldots, \theta_n) = \int \prod_{i=1}^{n} G(\theta_i) \, \mathrm{d}p(G),$$

that is, the sequence $\theta_1, \theta_2, \ldots$ can be viewed as conditionally i.i.d. draws from G.

The prior on the random probability measure G is $\mathsf{DP}(\alpha, G_0)$, which provides an explicit distribution over G and formalizes its existence (Teh, 2010). Moreover, this construction induces a clustering structure (Orhan, 2012), where observations are grouped according to shared values (colors), and drawing a particular color corresponds to selecting a cluster with a certain probability. This perspective motivates the Chinese Restaurant Process (CRP; see, for example, Navarro & Perfors, 2023 and Bouchard-Côté, 2011), a partition-based representation closely related to the DP.

3.3. GEM Distribution

This distribution, introduced by Ewens (1990) and named after Griffiths, Engen, and McCloskey, formalizes the stick-breaking construction, which serves as a constructive definition of the Dirichlet Process. Specifically, it generates the weights from a common Beta distribution by recursively partitioning a unit interval as follows:

$$\begin{aligned}
V_1 &\sim \mathsf{Beta}\,(1, a)\,, & \pi_1 &= V_1, \\
V_2 &\sim \mathsf{Beta}\,(1, a)\,, & \pi_2 &= (1 - \pi_1)\, V_2, \\
&\vdots & & \\
V_k &\sim \mathsf{Beta}\,(1, a)\,, & \pi_k &= \Big(1 - \sum_{i=1}^{k-1} \pi_i\Big) V_k.
\end{aligned}$$

The resulting sequence $(\pi_1, \pi_2, \ldots)$ lives in the infinite-dimensional simplex of $\mathbb{R}^\infty$ and satisfies $\sum_{k=1}^{\infty} \pi_k = 1$ almost surely.

3.4. Sampling a DP from the GEM Distribution

Consider again $G \sim \mathsf{DP}(\alpha, G_0)$. This process can also be represented via the stick-breaking construction as

$$G = \sum_{k=1}^{\infty} \pi_k \, \delta_{\phi_k},$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots) \sim \mathsf{GEM}(\alpha)$, $\phi_k \overset{\text{ind}}{\sim} G_0$, and δ_a denotes a point mass at a. To simulate from this representation, one first generates the sequence of weights $\{\pi_k\}$ using the $\mathsf{GEM}(\alpha)$ distribution, ensuring that $\sum_{k=1}^{\infty} \pi_k = 1$, and then, for each π_k, draws an atom ϕ_k independently from G_0 to define the corresponding point mass δ_{ϕ_k}.

Starting from the definition in (4), suppose for concreteness that the sample space is $\Omega = \mathbb{R}$. Then, for any measurable set $A \subset \mathbb{R}$,

$$G(A) = \sum_{k} \pi_k \, \delta_{\phi_k}(A) = \sum_{k:\phi_k \in A} \pi_k,$$

where $\delta_{\phi_k}(A)$ is the indicator that $\phi_k \in A$. If A is fixed and we draw $G \sim \mathsf{DP}(\alpha, G_0)$, each realization produces a different set of weights $\{\pi_k\}$ contributing to A. The fact that $G(A)$ is random for every measurable set A implies that the DP is a stochastic process.

The graphical model of a nonparametric GMM using the stick-breaking representation is shown in Figure 3. We begin by allowing a countably infinite collection of components $k = 1, 2, \ldots$ and sampling the corresponding weights π_k from the $\mathsf{GEM}(\alpha)$ distribution. In practice, a finite truncation level K is used as a computational convenience (rather than as part of the theoretical model), and this truncation determines the effective number of groups. Each group k, with weight π_k, is assigned a mean vector $\boldsymbol{\mu}_k$ drawn from the base distribution G_0, taken here to be a multivariate Normal distribution with parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$. This yields K mixture components with assignment probabilities π_k and means $\boldsymbol{\mu}_k$, from which observations y_i are generated for $i = 1, \ldots, n$.

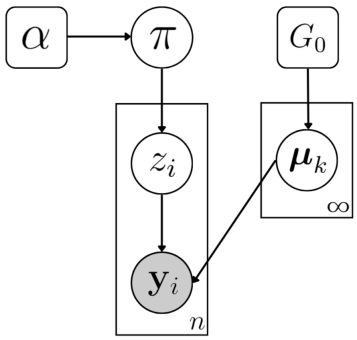


Figure 3: DAG for the nonparametric GMM.

3.5. Chinese Restaurant Process

Another interpretation of the DP, directly derived from the Pólya urn model and the posterior predictive distribution of θ_{n+1} in (9), is the Chinese Restaurant Process (CRP; see, for example, Navarro & Perfors (2023) and Bouchard-Côté, 2011). The name of the process comes from a metaphor that provides an intuitive description of how it induces a distribution over all partitions of the set $\{1, 2, 3, \ldots\}$.

Imagine a restaurant with a countably infinite number of numbered tables. Following Blei (2007), as customers arrive, they choose a table according to the following procedure:

1. The first customer sits at the first empty table.
2. The second customer chooses an empty table with probability $\frac{\alpha}{\alpha+1}$, or sits at the table occupied by the first customer with probability $\frac{1}{\alpha+1}$.
3. If the second customer sits at the occupied table, then the third customer chooses an empty table with probability $\frac{\alpha}{\alpha+2}$, or sits at the occupied table with probability $\frac{2}{\alpha+2}$.
4. In general, customer $n+1$ chooses an empty table with probability $\frac{\alpha}{\alpha+n}$, or sits at table k with probability $\frac{n_k}{\alpha+n}$, where n_k is the number of customers already seated at table k.

The probability of the assignment vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$ is determined by the predictive distribution in (9), where G_0 represents the distribution for new (previously empty) tables, so that $t^* \sim G_0$. However, since G_0 does not affect the partition structure encoded by $\boldsymbol{z}$, the CRP is defined solely in terms of the concentration parameter α. Thus, the probability of obtaining a specific configuration of assignments $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$ with concentration parameter α, denoted $\boldsymbol{z} \sim \mathsf{CRP}(\alpha)$, is given by

$$p(\boldsymbol{z} \mid \alpha, n) = \frac{\Gamma(\alpha) \prod_{t=1}^{T} \Gamma(n_t)}{\Gamma(n+\alpha)} \alpha^T, \tag{10}$$

where T is the number of occupied tables (clusters) and n_t is the number of customers seated at table t.

Below we present the probabilistic formulation of the nonparametric GMM in Figure 2, using the CRP as the prior distribution for component assignments. The Normal–Inverse-Wishart distribution with hyperparameters $\boldsymbol{\Lambda}_0 = (\boldsymbol{\mu}_0, \kappa_0, \upsilon_0, \mathbf{S}_0)$ is used as G_0, and, as in the parametric GMM, the full conditional posterior for the component label z_i is

$$p\big(z_i = k \mid \mathbf{z}_{-i}, \mathcal{Y}, \alpha, \boldsymbol{\Lambda}_0\big) \propto p\big(z_i = k \mid \mathbf{z}_{-i}, \alpha\big)\, p\big(\mathbf{y}_i \mid \mathcal{Y}_{-i}, z_i = k, \mathbf{z}_{-i}, \boldsymbol{\Lambda}_0\big), \tag{11}$$

where $\mathcal{Y}_{-i}$ denotes the set of observations excluding $\mathbf{y}_i$.

The prior term $p(z_i = k \mid \mathbf{z}_{-i}, \alpha)$ is obtained analogously to (3), but now assuming $\boldsymbol{z} \sim \mathsf{CRP}(\alpha)$, so that $p(\boldsymbol{z} \mid \alpha)$ is given by (10). Its computation separates into two cases:

- $p\big(z_i = k \mid \mathbf{z}_{-i}, \alpha\big)$, where k is an existing component:

$$p\big(z_i = k \mid \mathbf{z}_{-i}, \alpha\big) = \frac{n_k - 1}{n - 1 + \alpha} = \frac{n_{-i;k}}{n - 1 + \alpha},$$

  where $n_{-i;k} = n_k - 1$ is the number of observations assigned to component k excluding i.

- $p\big(z_i = k \mid \mathbf{z}_{-i}, \alpha\big)$, where k is a new component:

$$p\big(z_i = k \mid \mathbf{z}_{-i}, \alpha\big) = \frac{\alpha}{n - 1 + \alpha}.$$

Combining these cases, we obtain

$$p\big(z_i = k \mid \mathbf{z}_{-i}, \alpha\big) = \begin{cases} \dfrac{n_{-i;k}}{n + \alpha - 1}, & \text{if } k \text{ is an existing component,} \\ \dfrac{\alpha}{n + \alpha - 1}, & \text{if } k \text{ is a new component.} \end{cases} \tag{12}$$

Note that (12) can be derived from (3) by taking the limit $K \to \infty$. For the first case, it suffices to apply this limit directly; for a detailed derivation of the second case, see Kamper (2013).

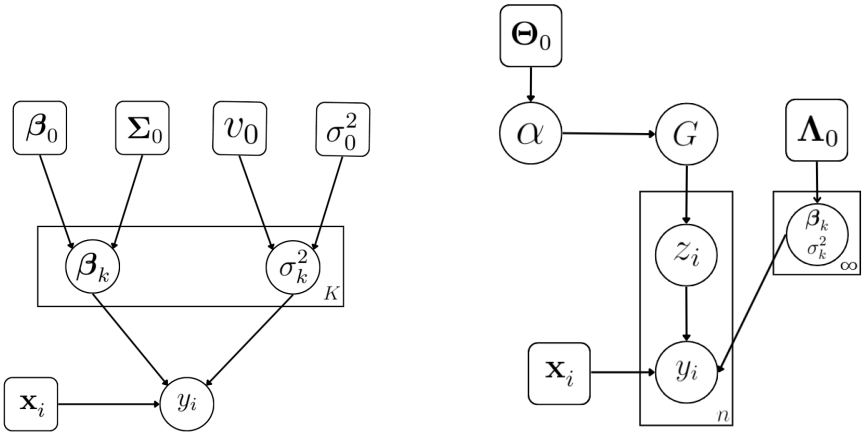
4. A Nonparametric Bayesian Multilevel Model

An appealing approach in regression analysis is to study the relationship between dependent and independent variables through fixed effects while simultaneously grouping observations with similar characteristics. This idea is implemented via the parametric multilevel Bayesian model described in Sosa & Aristizabal (2022).

Let y_i, with $i = 1, \ldots, n$, be a conditional independent sample, and let $\mathbf{x}_i$ denote the covariate vector associated with individual i. The model is specified as

$$y_i \mid \mathbf{x}_i, \{\boldsymbol{\beta}_k\}, \{\sigma_k^2\} \overset{\text{ind}}{\sim} \sum_{k=1}^{K} \omega_k \, \mathsf{N}\big(\mathbf{x}_i^\top \boldsymbol{\beta}_k, \sigma_k^2\big),$$

where K is a positive integer indicating the number of mixture components (groups). The component-specific regression coefficients are $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, and the corresponding variances are $\sigma_1^2, \ldots, \sigma_K^2$. The probability that observation i belongs to component k is $\omega_k = \mathrm{P}(z_i = k)$, where z_i is a latent variable indicating the allocation of observation i to component k. The associated DAG is shown in Figure 4(a).



(a) Parametric model. (b) Nonparametric model.

FIGURE 4: DAGs for the parametric and nonparametric Bayesian multilevel model.

4.1. Model Specification

Unlike the parametric approach, the nonparametric formulation does not require the number of groups to be specified in advance; instead, it allows for a countably infinite collection of components:

$$y_i \mid \mathbf{x}_i, \{\boldsymbol{\beta}_k\}, \{\sigma_k^2\} \overset{\text{ind}}{\sim} \sum_{k=1}^{\infty} \omega_k \, \mathsf{N}\big(\mathbf{x}_i^\top \boldsymbol{\beta}_k, \sigma_k^2\big).$$

The clustering structure is induced by a Dirichlet process (DP), with the Chinese Restaurant Process (CRP) specifying the prior on the allocation vector $\boldsymbol{z}$, as detailed in Section 3.5. Under this specification, the model can be written as

$$\begin{aligned} y_i \mid z_i = k, \mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k^2 &\overset{\text{ind}}{\sim} \mathsf{N}\big(\mathbf{x}_i^\top \boldsymbol{\beta}_k, \sigma_k^2\big), \\ (\boldsymbol{\beta}_k, \sigma_k^2) \mid F &\sim F, \\ \boldsymbol{z} &\sim \mathsf{CRP}(\alpha), \end{aligned}$$

where $k = 1, 2, \ldots$, α is the concentration parameter, and F denotes the joint distribution of $(\boldsymbol{\beta}_k, \sigma_k^2)$. Figure 4(b) shows the corresponding DAG, which also includes a prior distribution for α with hyperparameter vector $\boldsymbol{\Theta}_0$. The hyperparameter vector associated with the joint prior for $(\boldsymbol{\beta}_k, \sigma_k^2)$ is denoted by $\boldsymbol{\Lambda}_0$.

4.2. Prior Distribution

The prior distribution for $(\boldsymbol{\beta}_k, \sigma_k^2)$ is specified as

$$\boldsymbol{\beta}_k \mid \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0 \sim \mathsf{N}_{p+1}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0), \qquad \sigma_k^2 \mid \upsilon_0, \sigma_0^2 \sim \mathsf{IG}\left(\frac{\upsilon_0}{2}, \frac{\upsilon_0 \sigma_0^2}{2}\right), \tag{13}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$ are the mean vector and covariance matrix of the Normal prior for the regression coefficients, and v_0 and σ_0^2 are the degrees of freedom and scale parameter of the prior for σ_k^2. The concentration parameter α is assigned a Gamma prior, $\alpha \sim \mathsf{Gamma}(a, b)$, where a and b are the shape and scale hyperparameters.

The model formulated in this section has a multilevel structure because it incorporates several hierarchical layers that reflect the clustered nature of the data. First, each observation is associated with a latent group characterized by its own parameters $(\boldsymbol{\beta}_k, \sigma_k^2)$, which induces between-group variability. Second, the latent group assignments $\boldsymbol{z}$ arise from a distribution governed by a DP, represented here through a $\mathsf{CRP}(\alpha)$ with concentration parameter α, providing a second hierarchical level. Finally, α itself depends on the hyperparameters (a, b) via its Gamma prior, forming a third level. This multilevel construction allows the model to capture both heterogeneity across groups and uncertainty in the distribution of their parameters.

4.3. Joint Distribution

The joint distribution of the model is given below, using the CRP representation of the DP to characterize the mixture components. This distribution describes the complete probabilistic structure of the model and serves as the basis for Bayesian inference and for implementing the MCMC algorithm.

Considering the probability of $\boldsymbol{z}$ in (10), we obtain

$$\begin{aligned} p(\mathbf{y}, \{\boldsymbol{\beta}_k\}, \{\sigma_k^2\}, \boldsymbol{z}, \alpha) = \prod_{i=1}^{n} \mathsf{N}\big(y_i \mid \mathbf{x}_i^\top \boldsymbol{\beta}_{z_i}, \sigma_{z_i}^2\big) \times \prod_{k=1}^{K^*} \mathsf{N}\big(\boldsymbol{\beta}_k \mid \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0\big) \\ \times \prod_{k=1}^{K^*} \mathsf{IG}\big(\sigma_k^2 \mid v_0, \sigma_0^2\big) \times \frac{\Gamma(\alpha) \prod_{k=1}^{K^*} \Gamma(n_k)\, \alpha^{K^*}}{\Gamma(n+\alpha)} \\ \times \mathsf{Gamma}(\alpha \mid a, b), \end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function, $\sum_{k=1}^{K^*} n_k = n$, and K^* is the number of non-empty components.

4.4. Full Conditional Distributions

From the joint distribution, the following full conditional distributions are obtained:

1. Conditional distribution of $\boldsymbol{\beta}_k$:

$$p(\{\boldsymbol{\beta}_k\} \mid \mathbf{y}, \boldsymbol{z}, \{\sigma_k^2\}) \propto \prod_{i=1}^{n} \mathsf{N}\big(y_i \mid \mathbf{x}_i^\top \boldsymbol{\beta}_{z_i}, \sigma_{z_i}^2\big) \prod_{k=1}^{K^*} \mathsf{N}\big(\boldsymbol{\beta}_k \mid \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0\big).$$

It suffices to derive the conditional distribution of $\boldsymbol{\beta}_k$ for a single component k:

$$p(\boldsymbol{\beta}_k \mid \mathbf{y}_k, \sigma_k^2) \propto \prod_{i=1}^{n_k} \mathsf{N}\big(y_{ik} \mid \mathbf{x}_{ik}^\top \boldsymbol{\beta}_k, \sigma_k^2\big)\, \mathsf{N}\big(\boldsymbol{\beta}_k \mid \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0\big),$$

where y_{ik} is observation i assigned to component k, $\mathbf{x}_{ik}$ is its covariate vector, n_k is the number of observations in component k, and σ_k^2 is the corresponding variance. This is the standard conditional distribution of $\boldsymbol{\beta}_k$ in a Bayesian linear regression model. Therefore,

$$\boldsymbol{\beta}_k \mid \mathbf{y}_k, \sigma_k^2 \sim \mathsf{N}\left(\left(\boldsymbol{\Sigma}_0^{-1} + \tfrac{\mathbf{X}_k^\top \mathbf{X}_k}{\sigma_k^2}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \tfrac{\mathbf{X}_k^\top \mathbf{y}_k}{\sigma_k^2}\right), \left(\boldsymbol{\Sigma}_0^{-1} + \tfrac{\mathbf{X}_k^\top \mathbf{X}_k}{\sigma_k^2}\right)^{-1}\right),$$

where $\mathbf{y}_k$ is the vector of responses in component k and $\mathbf{X}_k$ is the associated design matrix.

2. Conditional distribution of σ_k^2:

$$p(\{\sigma_k^2\} \mid \mathbf{y}, \boldsymbol{z}, \{\boldsymbol{\beta}_k\}) \propto \prod_{i=1}^{n} \mathsf{N}\big(y_i \mid \mathbf{x}_i^\top \boldsymbol{\beta}_{z_i}, \sigma_{z_i}^2\big) \prod_{k=1}^{K^*} \mathsf{IG}\big(\sigma_k^2 \mid v_0, \sigma_0^2\big).$$

As in the previous case, it is enough to consider a single component k:

$$p(\sigma_k^2 \mid \mathbf{y}_k, \boldsymbol{\beta}_k) \propto \prod_{i=1}^{n_k} \mathsf{N}\big(y_{ik} \mid \mathbf{x}_{ik}^\top \boldsymbol{\beta}_k, \sigma_k^2\big)\, \mathsf{IG}\big(\sigma_k^2 \mid v_0, \sigma_0^2\big).$$

This is the usual conditional distribution of σ_k^2 in a Bayesian linear regression model. Hence,

$$\sigma_k^2 \mid \mathbf{y}_k, \boldsymbol{\beta}_k \sim \mathsf{IG}\left(\frac{v_0 + n_k}{2}, \frac{v_0\sigma_0^2 + (\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_k)^\top(\mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_k)}{2}\right).$$

3. Conditional distribution of z_i: Analogously to the posterior distribution of the component labels in the nonparametric GMM in (11), we have

$$p\big(z_i = k \mid \mathbf{y}, \boldsymbol{z}_{-i}, \alpha\big) \propto p\big(y_i \mid z_i = k, \mathbf{y}_{-i}, \boldsymbol{z}_{-i}\big)\, p\big(z_i = k \mid \boldsymbol{z}_{-i}, \alpha\big). \tag{14}$$

The term $p(z_i = k \mid \boldsymbol{z}_{-i}, \alpha)$ is given by the CRP prior in (12), so that

$$p(z_i = k \mid \boldsymbol{z}_{-i}, \alpha) = \begin{cases} \dfrac{n_{-i;k}}{n + \alpha - 1}, & \text{if } k \text{ is an existing component,} \\ \dfrac{\alpha}{n + \alpha - 1}, & \text{if } k \text{ is a new component,} \end{cases}$$

where $n_{-i;k}$ is the number of observations in component k excluding i.

4. Conditional distribution of α: Following the approach of West (1992), the conditional distribution of α satisfies

$$\begin{aligned} p(\alpha \mid \boldsymbol{z}) &\propto \frac{\Gamma(\alpha)\, \prod_{k=1}^{K^*} \Gamma(n_k)\, \alpha^{K^*}}{\Gamma(n + \alpha)} \times \frac{b^a}{\Gamma(a)}\, \alpha^{a-1} e^{-b\alpha} \\ &\propto \frac{\Gamma(\alpha)\, \alpha^{K^*}}{\Gamma(n + \alpha)}\, \alpha^{a-1} e^{-b\alpha}. \end{aligned}$$

Using properties of the Gamma function, for $\alpha > 0$ we have

$$\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} = \frac{\Gamma(\alpha+1)}{\alpha} \frac{\alpha+n}{\Gamma(\alpha+n+1)} = \frac{\Gamma(\alpha+1)\,\Gamma(n)}{\Gamma(\alpha+n+1)} \frac{\alpha+n}{\alpha\,\Gamma(n)}.$$

Since $\frac{\Gamma(\alpha+1)\,\Gamma(n)}{\Gamma(\alpha+n+1)} = B(\alpha+1, n)$, where $B(\cdot,\cdot)$ is the Beta function, it follows that

$$\begin{aligned} p(\alpha \mid \boldsymbol{z}) &\propto \frac{(\alpha+n)\,B(\alpha+1,n)\,\alpha^{K^*}}{\alpha\,\Gamma(n)}\,\alpha^{a-1}e^{-b\alpha} \\ &\propto (\alpha+n)\,\alpha^{K^*+a-2}e^{-b\alpha}\left(\int_0^1 x^{\alpha}(1-x)^{n-1}\,dx\right). \end{aligned}$$

Thus $p(\alpha \mid \boldsymbol{z})$ can be viewed as the marginal of the joint distribution of (α, x), where

$$p(\alpha, x \mid \boldsymbol{z}) \propto (\alpha+n)\,\alpha^{K^*+a-2}e^{-b\alpha}\,x^{\alpha}(1-x)^{n-1},$$

with $\alpha > 0$ and $0 < x < 1$. The conditional distribution of α is then sampled via MCMC by working with the posterior of (α, x).

- Posterior distribution of α:

$$\begin{aligned} p(\alpha \mid \boldsymbol{z}, x) &\propto (\alpha+n)\,\alpha^{K^*+a-2}e^{-b\alpha}x^{\alpha} \\ &\propto \alpha^{K^*+a-1}e^{-\alpha(b-\log x)} + n\,\alpha^{K^*+a-2}e^{-\alpha(b-\log x)}. \end{aligned}$$

  Hence, $p(\alpha \mid \boldsymbol{z}, x)$ is a mixture of two Gamma distributions with common rate $b - \log x$ and shapes $K^* + a$ and $K^* + a - 1$. Up to proportionality, the two mixture components are

$$w_1 \propto \frac{\Gamma(K^*+a)}{(b-\log x)^{K^*+a}}, \qquad w_2 \propto \frac{n\,\Gamma(K^*+a-1)}{(b-\log x)^{K^*+a-1}},$$

  so that, after normalization, $\alpha \mid \boldsymbol{z}, x$ can be sampled from

$$\mathsf{Gamma}(K^*+a,\, b-\log x) \qquad \text{or} \qquad \mathsf{Gamma}(K^*+a-1,\, b-\log x),$$

  with mixture weights proportional to w_1 and w_2. The ratio of the unnormalized weights is

$$\frac{w_1}{w_2} = \frac{(K^*+a-1)}{n\,(b-\log x)}.$$

- Posterior distribution of x:

$$p(x \mid \boldsymbol{z}, \alpha) \propto x^{\alpha}(1-x)^{n-1}, \qquad 0 < x < 1,$$

  so that

$$x \mid \alpha, \boldsymbol{z} \sim \mathsf{Beta}(\alpha+1, n).$$

4.5. MCMC Algorithm

Algorithm 1 implements an MCMC scheme based on successive updates from the full conditional distributions of the nonparametric Bayesian multilevel model. Starting from initial values for the allocation variables $\boldsymbol{z}$, the component-specific parameters $\{(\boldsymbol{\beta}_k, \sigma_k^2)\}$, and the concentration parameter α, each iteration $b = 1, \ldots, B$ proceeds as follows. First, the allocation indicators z_i^b are updated one at a time from their full conditional distribution $p(z_i = k \mid \mathbf{y}, \boldsymbol{z}_{-i}, \alpha^{b-1})$, which combines the CRP prior and the component-wise likelihood. This step determines the current number of occupied components K^b and the corresponding groups of observations. Then, for each occupied component $k = 1, \ldots, K^b$, the regression coefficients $\boldsymbol{\beta}_k^b$ are drawn from their Normal full conditional $p(\boldsymbol{\beta}_k \mid \mathbf{y}_k, \boldsymbol{z}^{b-1}, \sigma_k^{2^{b-1}})$, and the variances $\sigma_k^{2^b}$ are updated from the Inverse-Gamma full conditional $p(\sigma_k^2 \mid \mathbf{y}_k, \boldsymbol{z}^{b-1}, \boldsymbol{\beta}_k^{b-1})$. Finally, the concentration parameter α^b is sampled from its full conditional $p(\alpha \mid \boldsymbol{z}^{b-1}, x)$, which is obtained via the auxiliary-variable representation described earlier. Repeating this cycle for B iterations yields a Markov chain whose stationary distribution is the joint posterior of all unknown quantities.

Algorithm 1 MCMC Algorithm for the Nonparametric Bayesian Multilevel Model.

Require: Number of samples B; initialize $z_i = 1 \ \ \forall i$, $\{\boldsymbol{\beta}_k^0\}$, $\{\sigma_k^{2^0}\}$, and α^0

for $b = 1$ to B **do**
  for each $i = 1$ to n **do**
    Sample z_i^b from $p(z_i = k \mid \mathbf{y}, \boldsymbol{z}_{-i}, \alpha^{b-1})$
  end for
  for each $k = 1$ to K^b **do**
    Sample $\boldsymbol{\beta}_k^b$ from $p(\boldsymbol{\beta}_k \mid \mathbf{y}_k, \boldsymbol{z}^{b-1}, \sigma_k^{2^{b-1}})$
    Sample $\sigma_k^{2^b}$ from $p(\sigma_k^2 \mid \mathbf{y}_k, \boldsymbol{z}^{b-1}, \boldsymbol{\beta}_k^{b-1})$
  end for
  Sample α^b from $p(\alpha \mid \boldsymbol{z}^{b-1}, x)$
end for

5. Application

This dataset contains information on fuel consumption and estimated carbon dioxide emissions for new vehicles offered for retail sale in Canada between January 2023 and February 2025. It includes both hybrid vehicles and those that operate exclusively on fuel. The data were obtained from the official Government of Canada portal, where they are freely available to the public `https://open.canada.ca/data/en/dataset/`. The database comprises a total of 2,457 vehicles and five key variables, described in detail in Table 1. These variables provide a comprehensive overview of fuel consumption patterns and associated emissions, enabling an in-depth analysis that can inform future sustainability and energy-efficiency policies in the automotive sector.

For this application, the response variable y corresponds to CO_2 emissions (g/km). The model can be specified as

$$y_i \mid z_i = k, \mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k^2 \overset{\text{ind}}{\sim} \mathsf{N}\big(\mathbf{x}_i^\top \boldsymbol{\beta}_k, \sigma_k^2\big), \qquad (\boldsymbol{\beta}_k, \sigma_k^2) \mid F \sim F, \qquad \boldsymbol{z} \mid \alpha \sim \mathsf{CRP}(\alpha),$$

for $k = 1, 2, \ldots$, where $\mathbf{x}_i^\top = [1, x_{i,1}, x_{i,2}, x_{i,3}]$. Here, $x_{i,1}$ denotes the combined fuel consumption (L/km), $x_{i,2}$ is a dummy variable equal to 1 if the fuel type is premium, and $x_{i,3}$ is a dummy variable equal to 1 if the fuel type is regular.

Table 1: Description of variables in the vehicle CO2 emissions database.

Variable	Description	Additional Notes
Vehicle Make	Vehicle brand	N/A
Vehicle Segment	Classification of the vehicle according to its size and design	See available categories in Table 2 and 3
Fuel Type	Type of fuel used to power the vehicle. In the case of hybrids, this is the type of fuel used when in gasoline mode	X = Regular gasoline; Z = Premium gasoline; D = Diesel; E = E85; N = Natural gas
Combined Consumption (L/100 km)	Fuel consumption in liters per 100 kilometers	A weighted average of fuel consumption was calculated, assigning a weight of 55% to urban area driving and 45% to highway driving.
CO2 Emissions (g/km)	Vehicle emissions of carbon dioxide shown in grams per kilometer	N/A

The choice of explanatory variables in this application follows standard practice in modeling vehicle emissions and is guided by a combination of prior domain knowledge and simple exploratory analysis of the database (including the boxplots in Figures 5 and 6 and the bivariate association in Figure 7). Combined fuel consumption $x_{i,1}$ is included as the main physical driver of CO_2 emissions, for which a strong approximately linear relationship with y is observed. The dummy variables $x_{i,2}$ and $x_{i,3}$ capture systematic differences in emission levels across fuel types, with regular gasoline as one of the categories and the remaining fuels (diesel, E85, natural gas) forming the reference group. We deliberately adopt this parsimonious specification, relying on the nonparametric random-effects structure to account for residual heterogeneity across brand–segment combinations, rather than introducing a larger set of covariates or using automatic variable-selection procedures.

Table 2: Vehicle segment description for cars.

Vehicle Segment	Interior Volume
Two-seater (T)	N/A
Minicompact (I)	Less than 2,405 L
Subcompact (S)	2,405–2,830 L
Compact (C)	2,830–3,115 L
Mid-size (M)	3,115–3,400 L
Full-size (L)	3,400 L or more
Station wagon: Small (WS)	Less than 3,680 L
Station wagon: Mid-size (WM)	3,680–4,530 L

TABLE 3: Vehicle segment description for light trucks.

Vehicle Segment	**Weight**
Pickup truck: Small (PS)	Less than 2,722 kg
Pickup truck: Standard (PL)	2,722–3,856 kg
Sport utility vehicle: Small (US)	Less than 2,722 kg
Sport utility vehicle: Standard (UL)	2,722–4,536 kg
Minivan (V)	Less than 3,856 kg
Van: Cargo (VC)	Less than 3,856 kg
Van: Passenger (VP)	Less than 4,536 kg
Special purpose vehicle (SP)	N/A

Figures 5 and 6 display boxplots of CO_2 emissions by vehicle brand and segment. The distributions reveal marked heterogeneity both within and between categories, with clear groupings according to emission levels. In particular, luxury and high-performance brands such as Bugatti, Lamborghini, Rolls-Royce, and Aston Martin exhibit the highest CO_2 emissions, reflected in elevated medians and wide interquartile ranges. By contrast, hybrid vehicles—irrespective of brand or segment—concentrate at the lower end of the emission scale, showing substantially reduced central tendencies and dispersion. These patterns highlight strong systematic differences in emissions across market segments and underscore the need for flexible, cluster-aware regression models capable of capturing such structural heterogeneity.

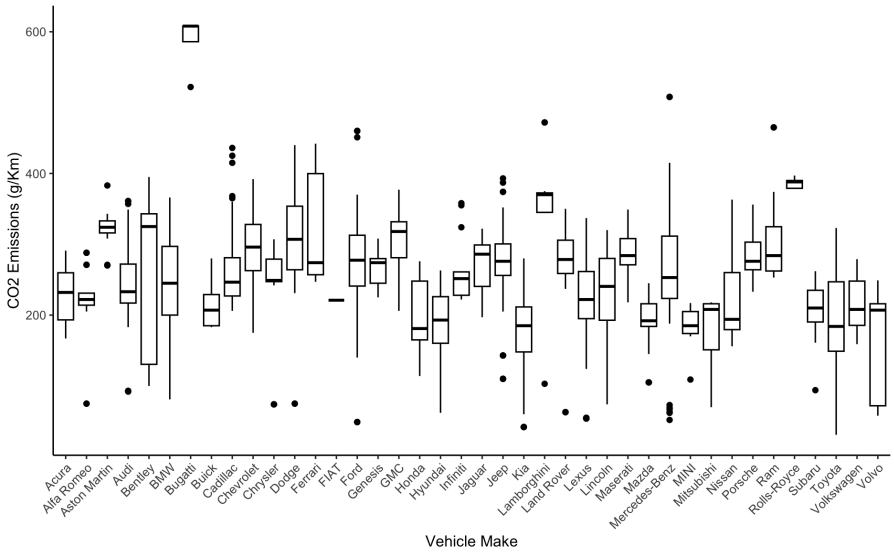


FIGURE 5: Boxplots of CO2 emissions by vehicle brand.

Figure 7 illustrates the relationship between combined fuel consumption and CO_2 emissions, stratified by fuel type. A clear positive linear association is observed: as fuel consumption increases, CO_2 emissions also rise, a pattern supported by a high correlation coefficient of 0.89. In addition, fuel type Z, corresponding to

premium fuel, is systematically associated with higher CO_2 emission levels relative to the other fuel types.

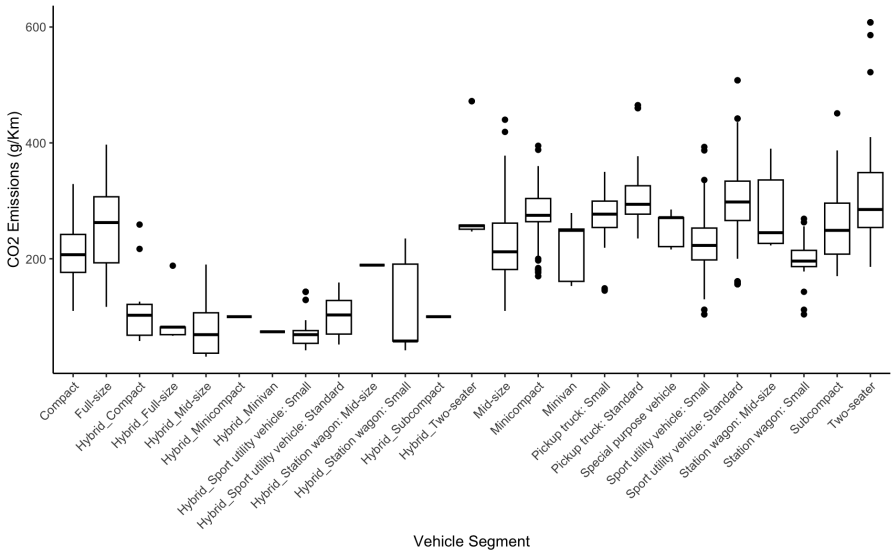


FIGURE 6: Boxplots of CO2 emissions by vehicle segment.

The dataset used in this study contains multiple observations for each combination of vehicle segment and brand. This structure allows the analysis to focus on identifying clusters at the segment–brand level rather than modeling each observation y_i in isolation. Specifically, we aim to find groupings for y_{ij}, where y_{ij} denotes observation i belonging to segment–brand combination j. The model is thus specified as

$$y_{ij} \mid z_j = k, \mathbf{x}_{ij}, \boldsymbol{\beta}_k, \sigma_k^2 \overset{\text{ind}}{\sim} \mathsf{N}\big(\mathbf{x}_{ij}^\top \boldsymbol{\beta}_k, \sigma_k^2\big), \qquad (\boldsymbol{\beta}_k, \sigma_k^2) \mid F \sim F, \qquad \boldsymbol{z} \mid \alpha \sim \mathsf{CRP}(\alpha),$$

for $k = 1, 2, \ldots$, where $\mathbf{x}_{ij}$ denotes the covariate vector for observation i in segment–brand combination j.

Under this model, the posterior distribution of $\boldsymbol{z}$ in (14) must be modified so that the new full conditional for the allocation variable is

$$p\big(z_j = k \mid \mathbf{y}, \boldsymbol{z}_{-j}, \boldsymbol{\beta}_k, \sigma_k^2, \alpha\big),$$

for $j = 1, \ldots, m$, where m is the total number of segment–brand combinations (here $m = 224$). The CRP prior then becomes

$$p(z_j = k \mid \boldsymbol{z}_{-j}, \alpha) = \begin{cases} \dfrac{m_{-j;k}}{m + \alpha - 1}, & \text{if } k \text{ is an existing component,} \\ \dfrac{\alpha}{m + \alpha - 1}, & \text{if } k \text{ is a new component,} \end{cases}$$

where $m_{-j;k}$ is the number of segment–brand combinations currently assigned to group k.

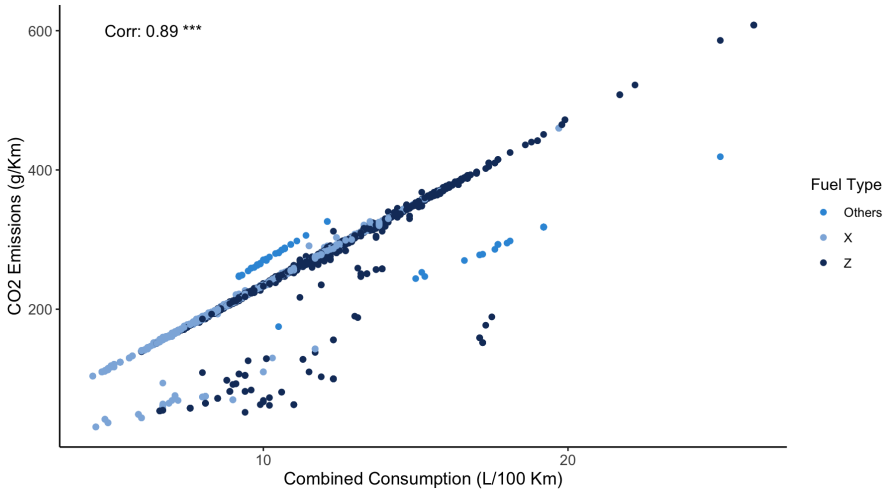


FIGURE 7: Combined fuel consumption versus CO_2 emissions by fuel type. The sample correlation coefficient is 0.89 and is statistically significant at the 5% level.

To fit the Bayesian nonparametric multilevel model, we adopt the unit information prior methodology of Kass & Wasserman (1995). In this framework, the prior mean vector for the regression coefficients is set to the ordinary least squares estimate, $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}_{\text{OLS}}$ (OLS stands for ordinary least squares), and the prior covariance matrix is specified as

$$\boldsymbol{\Sigma}_0 = 224\,\hat{\sigma}^2_{\text{ols}}\,(\mathbf{X}^\top\mathbf{X})^{-1}.$$

For the variance prior, we take $\nu_0 = 1$ and $\sigma_0^2 = \hat{\sigma}^2_{\text{OLS}}$. For the concentration parameter, we adopt $\alpha \sim \mathsf{Gamma}(1, 10)$ with $\mathsf{E}(\alpha) = 0.10$ to moderately constrain the expected number of clusters while allowing flexibility to discover data heterogeneity. The robustness of our findings to this choice is assessed through a comprehensive sensitivity analysis presented in Appendix A.

For model fitting, a total of 150,000 MCMC iterations were generated. The first 50,000 iterations were discarded as burn-in, and to reduce autocorrelation, every tenth remaining draw was retained, yielding a final posterior sample of 10,000 iterations. Figure 8 displays the trace plot of the log-likelihood, which indicates satisfactory convergence of the chain. The fitted model identifies 9 distinct clusters. Table 4 reports the distribution of both individual vehicles and segment–brand combinations across these 9 groups.

Each group is associated with a characteristic type of vehicle. Groups 3, 7, 8, and 9 correspond to hybrid vehicles, with Group 7 standing out as it is composed of high-end models from brands such as Ferrari. By contrast, Groups 1 and 6 are associated with vehicles that generate higher emissions. Group 1 includes only pickup trucks, as expected given that they are large vehicles with more powerful engines designed to handle heavy loads. Group 6, in turn, consists mainly of high-end and sports vehicles, such as Porsche, which also exhibit elevated emissions due

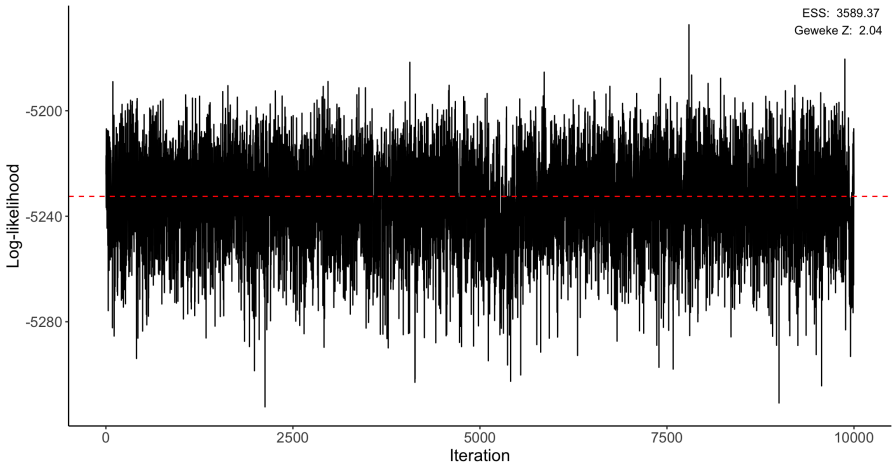


FIGURE 8: Trace plot of the log-likelihood for the vehicle CO_2 emissions application.

to their high-performance engines. Furthermore, Figure 9 shows the 90% credible intervals and posterior means of σ_k^2 for $k = 1, \ldots, 9$. The results indicate that the variance is not homogeneous across groups. In particular, the intervals for Groups 7 through 9, which correspond to hybrid vehicles, reveal greater posterior uncertainty.

TABLE 4: Distribution of vehicles by estimated group.

Group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Vehicles	121	1696	72	303	170	36	25	16	18

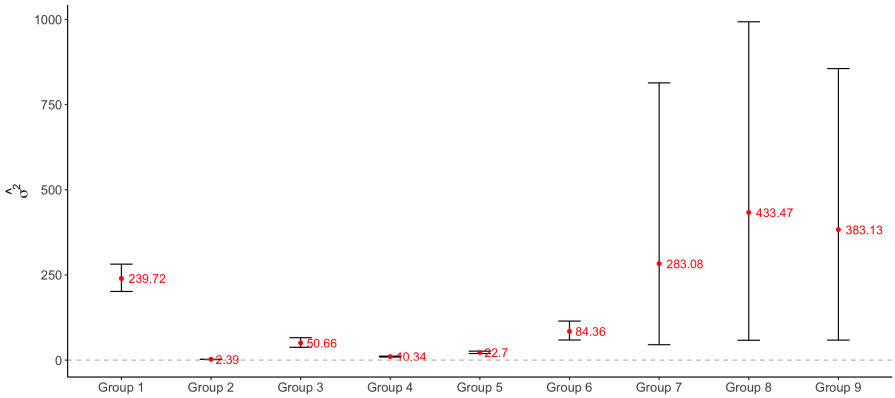
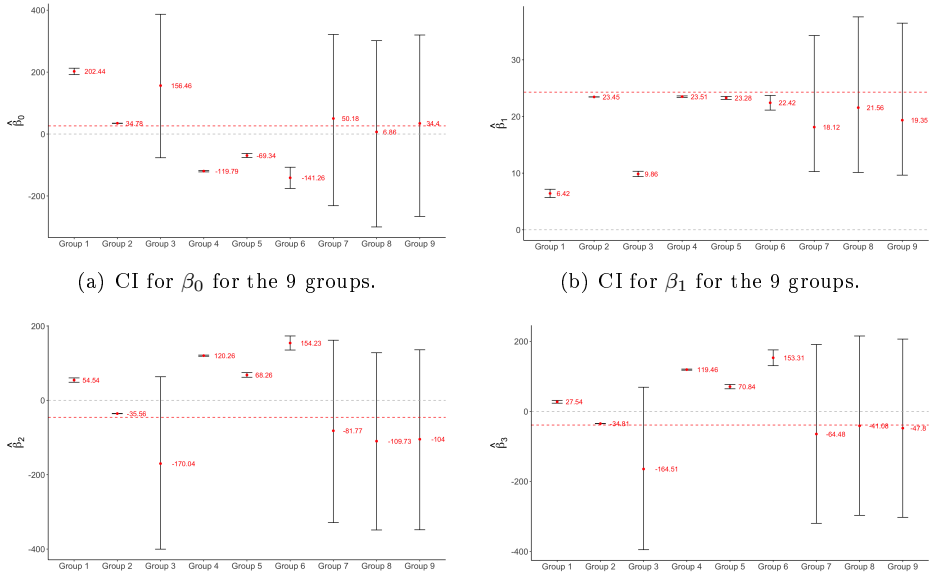


FIGURE 9: 90% credible intervals for σ_k^2, $k = 1, 2, \ldots, 9$.

Regarding the regression coefficients, the credible intervals shown in Figures 10(a), 10(b), 10(c), and 10(d) reveal several noteworthy patterns. For Groups 1, 2, 4, 5, and 6, all coefficients are significant, indicating that both combined

fuel consumption and fuel type are key determinants of CO_2 emissions. It is also worth noting that the simple linear regression estimates lie outside these credible intervals, suggesting that the linear model does not adequately capture the heterogeneity present in the data. In particular, Groups 2, 4, 5, and 6 are associated with higher CO_2 emissions per liter consumed per 100 km, as illustrated in Figure 10(b).



(a) CI for β_0 for the 9 groups.

(b) CI for β_1 for the 9 groups.

(c) CI for β_2 for the 9 groups.

(d) CI for β_3 for the 9 groups.

FIGURE 10: 90% credible intervals for $\boldsymbol{\beta}_k$, $k = 1, 2, \ldots, 9$. The gray line marks the zero value, used as a reference to assess the significance of the estimates, and the red line indicates the regression coefficient estimated by simple linear regression.

By contrast, Groups 3, 7, 8, and 9, which correspond to hybrid vehicles, exhibit significance only in the coefficient associated with combined fuel consumption. The variability in fuel type within these groups is minimal, since most hybrids use regular gasoline when operating in gasoline mode, which explains the lack of significance of the fuel-type indicators. The inclusion of these covariates also helps explain the lack of significance of the intercepts for these groups.

Importantly, the Bayesian nonparametric multilevel model, unlike the simple linear regression, allows us to conclude that fuel type is not a significant predictor for hybrid vehicles. This finding highlights the ability of the Bayesian model to uncover relationships that the traditional approach fails to detect. In addition, goodness-of-fit measures improve substantially: the residual standard error decreases from 28.99 to 7.46, and R^2 increases from 0.84 to 0.99 when moving from the simple linear regression to the proposed Bayesian nonparametric multilevel model.

Figure 11 presents the boxplots of CO_2 emissions by group, reinforcing the previous findings. The groups associated with hybrid vehicles exhibit lower emission levels, whereas Groups 1 and 6 are characterized by higher emissions. Figure 12 shows the scatter plot of fuel consumption versus CO_2 emissions, where the differences between clusters are clearly visible, particularly between Group 1 and the groups from 7 onward.

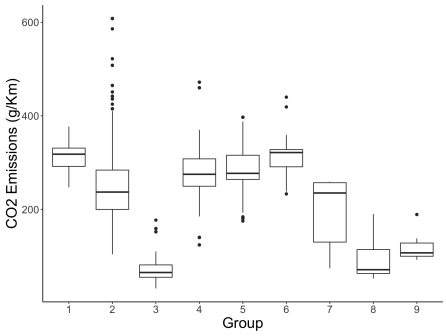


FIGURE 11: Boxplot of CO_2 emissions by estimated group.

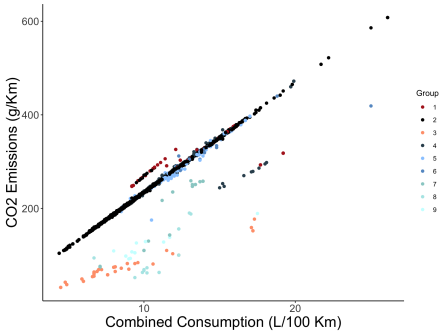


FIGURE 12: Scatter plot of CO_2 emissions versus combined fuel consumption, segmented by estimated cluster.

Finally, we compare the multilevel Bayesian non-parametric model with a Bayesian parametric counterpart. The parametric model uses the same hierarchical mixture regression structure and likelihood as the proposed approach, but replaces the DP prior with a finite mixture of K components, with K fixed in advance. Conditionally on K, all component-specific regression coefficients and variances are estimated under a standard Bayesian formulation, using the same priors as in the non-parametric specification and an MCMC algorithm to sample the latent allocations and model parameters. For the application, we fit this parametric mixture for $K \in \{6, \ldots, 11\}$ and selected $K = 9$ as the best-fitting option according to the goodness-of-fit criteria considered.

Table 5 summarizes the resulting goodness-of-fit and clustering metrics. Both models achieve very similar overall fit. The parametric version attains slightly smaller DIC and WAIC values, whereas the non-parametric model attains a lower mean squared error (MSE) and essentially the same R^2. More importantly for our purposes, the non-parametric model yields better group formation, with lower intra-cluster variance and higher inter-cluster variance, indicating more cohesive and better separated clusters. In terms of computational cost, each fixed-K parametric model is comparatively simple to fit, but it must be re-estimated for every value of K in the chosen range, while the DP mixture is fitted only once at the cost of sampling the additional concentration parameter α and allowing a random number of occupied components. In this application, the total runtime of the non-parametric model was on the order of two to three times that of a single parametric fit, but comparable to the combined cost of fitting multiple parametric

models for different values of K. A more exhaustive predictive comparison based on cross-validation would certainly be of interest, but lies outside the scope of this applied illustration.

TABLE 5: Comparison of goodness-of-fit metrics for the nonparametric and parametric models.

Metric	Non-Parametric	Parametric
DIC	11850.98	**11831.17**
WAIC	10573.13	**10479.21**
MSE	**55.65**	59.65
R^2	0.99	0.99
Intra-cluster	**8870.99**	10102.94
Inter-cluster	**113.49**	104.97

6. Conclusions

This study shows that the multilevel Bayesian nonparametric model can overcome key limitations of parametric approaches in settings where approximately linear relationships coexist with latent clustering and an unknown number of groups. The work specifies the full probabilistic structure of the model, derives the corresponding full conditional distributions, and develops an MCMC algorithm that enables posterior inference and practical implementation.

Through applications to real data, the multilevel Bayesian nonparametric model is found to be effective both for clustering and for estimating linear relationships within clusters. Its main advantage over parametric formulations is that it does not require the number of groups to be fixed in advance; instead, the clustering structure adapts automatically to the complexity and variability of the data.

The results highlight two particularly relevant scenarios in which the proposed model is especially useful. First, in situations where traditional methods incorrectly discard existing linear relationships because they fail to account for the underlying grouped structure. Second, in situations where a traditional linear model provides a global fit but does not adequately characterize specific subpopulations, potentially leading to systematic underestimation or overestimation for certain individuals. In both cases, the multilevel Bayesian nonparametric model proves to be a valuable and effective alternative.

In the comparison with a parametric Bayesian model, it is observed that in controlled settings such as simulations (not shown), both approaches perform similarly. In contrast, in real applications, global goodness-of-fit measures based on information criteria tend to favor the parametric model slightly, likely reflecting the greater flexibility and complexity of the nonparametric specification. However, these criteria do not fully capture the practical complexity of fitting the parametric model, which requires specifying a prior range for the number of groups, fitting the model for each candidate value, and then selecting the best option. By comparison, the nonparametric model requires only a single fit.

On the other hand, clustering performance metrics consistently indicate better segmentation under the Bayesian nonparametric model than under the parametric alternative. Taking into account the improved clustering, competitive goodness-of-fit, and the practical advantage of avoiding repeated fits for different numbers of groups, the nonparametric specification emerges as the more attractive option overall.

Despite these advantages, the proposed model has some limitations. First, although not observed in the applications considered here, it is possible in principle that the number of groups may not stabilize adequately in some settings. Second, the model is substantially more demanding in terms of computation time, which can be up to three times longer than that required to fit a parametric Bayesian model. Third, variable selection remains challenging: some coefficients may be non-significant in specific groups, and their inclusion can influence the estimation of other, significant effects.

Several directions for future research follow naturally. One promising line is the use of variational inference methods (Blei et al., 2017), which would facilitate the analysis of larger datasets and improve scalability. It would also be interesting to modify the base distribution, for example by using a Student-t distribution instead of the Normal to better accommodate outliers, or a Laplace distribution to induce regularization and enable covariate selection in a manner analogous to Lasso-type models (Tibshirani, 1996). For the variance component, adopting a Half-Cauchy prior on the standard deviation (Polson & Scott, 2011), rather than an inverse-Gamma prior on the variance, could increase robustness.

Another relevant extension is to adapt the model to handle non-Gaussian responses, such as count or binary outcomes, which would broaden its applicability to more complex data structures, including binary or weighted networks. It would also be natural to incorporate covariates into the mixture weights through logistic or softmax functions (Murphy, 2012), yielding covariate-dependent mixture models that support more accurate supervised segmentation. In addition, more flexible specifications could be explored in which the mixture weights or component parameters vary with covariates, for example using dependent Dirichlet processes (DDP) or semiparametric/additive regressions within each component to capture nonlinear relationships.

For multivariate response settings, a further generalization would allow each component to have a full covariance matrix, which is useful for capturing correlations among multiple outcomes. More complex dependence structures, such as autoregressive terms or random effects, could also be incorporated, extending the applicability of the proposed model to data with temporal or spatial dependence.

Finally, the code for the Bayesian nonparametric model is publicly available in a GitHub repository `https://github.com/Camilacruzdepaula/Bayesian_Non_Parametric_Model`. This open access facilitates replication of the results presented here and provides a practical tool for researchers and practitioners interested in Bayesian nonparametric modeling, who may use this implementation as a starting point. By making the code available, we aim to encourage the exchange of ideas and collaboration within the scientific community, thereby contributing to the development of Bayesian nonparametric statistics and its applications.

Statements and declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

During the preparation of this work the authors used ChatGPT-5 in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

[Received: September 2025 — Accepted: November 2025]

References

Barnes III, T. G., Jefferys, W., Berger, J., Mueller, P. J., Orr, K. & Rodriguez, R. (2003), 'A Bayesian analysis of the cepheid distance scale', *The Astrophysical Journal* **592**(1), 539.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer, New York.

Blackwell, D. & MacQueen, J. B. (1973), 'Ferguson distributions via Pólia urn schemes', *Annals of Statistics* **1**(2), 353–355.

Blei, D. M. (2007), 'Lecture 1: Bayesian nonparametrics', Lecture notes for COS 597C: Bayesian Nonparametrics. Scribes: Peter Frazier and Indraneel Mukherjee.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017), 'Variational inference: A review for statisticians', *Journal of the American statistical Association* **112**(518), 859–877.

Bouchard-Côté, A. (2011), Statistical Modeling with Stochastic Processes, PhD thesis, University of British Columbia, Vancouver, Canada.

Clyde, M. & George, E. I. (2000), 'Flexible empirical Bayes estimation for wavelets', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **62**(4), 681–698.

Dunson, D. B. (2010), 'Nonparametric Bayes applications to biostatistics', *Bayesian nonparametrics* **28**, 223–273.

Ewens, W. J. (1990), Population genetics theory-the past and the future, *in* 'Mathematical and statistical developments of evolutionary theory', Springer, pp. 177–227.

Ferguson, T. S. (1973), 'A Bayesian analysis of some nonparametric problems', *The annals of statistics* pp. 209–230.

Foti, N. J. & Williamson, S. A. (2013), 'A survey of non-exchangeable priors for Bayesian nonparametric models', *IEEE transactions on pattern analysis and machine intelligence* **37**(2), 359–371.

Frigyik, B. A., Kapila, A. & Gupta, M. R. (2010), Introduction to the Dirichlet distribution and related processes, UWEE Technical Report UWEETR-2010-0006, University of Washington, Department of Electrical Engineering.

Gamerman, D. & Lopes, H. F. (2006), *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, Chapman and Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis*, 3 edn, Chapman and Hall/CRC, Boca Raton, FL.

Hanson, T. & Johnson, W. O. (2002), 'Modeling regression error with a mixture of Pólia trees', *Journal of the American Statistical Association* **97**(460), 1020–1033.

Hanson, T. & Johnson, W. O. (2004), 'A Bayesian semiparametric AFT model for interval-censored data', *Journal of Computational and Graphical Statistics* **13**(2), 341–361.

Hoff, P. D. (2009), *A first course in Bayesian statistical methods*, Vol. 580, Springer.

Jara, A. (2017), 'Theory and computations for the dirichlet process and related models: An overview', *International Journal of Approximate Reasoning* **81**, 128–146.

Kamper, H. (2013), 'Gibbs sampling for fitting finite and infinite Gaussian mixture models', Technical report.

Kass, R. E. & Wasserman, L. (1995), 'A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion', *Journal of the american statistical association* **90**(431), 928–934.

MacEachern, S. N. (1999), Dependent nonparametric processes, *in* 'ASA proceedings of the section on Bayesian statistical science', Vol. 1, Alexandria, VA, pp. 50–55.

Müeller, P., Quintana, F. A. & Page, G. (2018), 'Nonparametric Bayesian inference in applications', *Statistical Methods & Applications* **27**(2), 175–206.

Müller, P., Erkanli, A. & West, M. (1996), 'Bayesian curve fitting using multivariate normal mixtures', *Biometrika* **83**(1), 67–79.

Müller, P. & Mitra, R. (2013), 'Bayesian nonparametric inference–why and how', *Bayesian analysis (Online)* **8**(2), 10–1214.

Müller, P., Quintana, F. A., Jara, A. & Hanson, T. (2015), *Bayesian nonparametric data analysis*, Vol. 1, Springer.

Murphy, K. P. (2012), *Machine learning: a probabilistic perspective*, MIT press.

Navarro, D. J. & Perfors, A. (2023), The chinese restaurant process. Lecture notes, University of Adelaide.

Orhan, E. (2012), 'Bayesian statistics: Dirichlet processes', Lecture notes. Unpublished manuscript.

Polson, N. G. & Scott, J. G. (2011), 'On the half-cauchy prior for a global scale parameter'.

Quintana, F. A., Müller, P., Jara, A. & MacEachern, S. N. (2022), 'The dependent dirichlet process and related models', *Statistical Science* **37**(1), 24–41.

Schörgendorfer, A., Branscum, A. J. & Hanson, T. E. (2013), 'A bayesian goodness of fit test and semiparametric generalization of logistic regression with measurement data', *Biometrics* **69**(2), 508–519.

Sosa, J. & Aristizabal, J.-P. (2022), 'Some developments in bayesian hierarchical linear regression modeling', *Revista Colombiana de Estadística* **45**(2), 231–255.

Teh, Y. W. (2010), Dirichlet processes, *in* C. Sammut & G. I. Webb, eds, 'Encyclopedia of Machine Learning', Springer, pp. 280–287. `https://www.stats.ox.ac.uk/~teh/research/npbayes/Teh2010a.pdf`

Theodoridis, S. (2020), *Machine Learning: A Bayesian and Optimization Perspective*, second edn, Academic Press.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**(1), 267–288.

Walker, S. & Mallick, B. K. (1999), 'A Bayesian semiparametric accelerated failure time model', *Biometrics* **55**(2), 477–483.

West, M. (1992), Hyperparameter estimation in dirichlet process mixture models, Discussion Paper 92-A03, Institute of Statistics and Decision Sciences, Duke University.

Williams, C. K. & Rasmussen, C. E. (2006), *Gaussian processes for machine learning*, Vol. 2, MIT press Cambridge, MA.

Xu, Y., Müller, P., Wahed, A. S. & Thall, P. F. (2016), 'Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times', *Journal of the American Statistical Association* **111**(515), 921–950.

Xuan, J., Lu, J. & Zhang, G. (2019), 'A survey on Bayesian nonparametric learning', *ACM Computing Surveys (CSUR)* **52**(1), 1–36.

Appendix. Sensitivity Analysis

The concentration parameter α in a DP mixture model critically determines the expected number of clusters. In the main analysis, we adopted a moderately informative prior $\alpha \sim \mathsf{Gamma}(1, 10)$ with $\mathsf{E}(\alpha) = 0.10$ to reduce the tendency of the model to create an excessive number of components while allowing sufficient flexibility to discover genuine subgroup structure. To assess the robustness of our inference to this prior specification, we conducted a comprehensive sensitivity analysis examining five different prior scenarios for α: $\mathsf{Gamma}(1, 100)$ with $\mathsf{E}(\alpha) = 0.01$ (strongly conservative, favoring very few clusters), $\mathsf{Gamma}(1, 20)$ with $\mathsf{E}(\alpha) = 0.05$ (conservative), $\mathsf{Gamma}(1, 10)$ with $\mathsf{E}(\alpha) = 0.10$ (the prior used in the main analysis), $\mathsf{Gamma}(1, 1)$ with $\mathsf{E}(\alpha) = 1.0$ (weakly informative), and $\mathsf{Gamma}(5, 1)$ with $\mathsf{E}(\alpha) = 5.0$ (diffuse, permitting many clusters). All other hyperparameters, including those for $\boldsymbol{\beta}_k$ and σ_k^2 specified in Section 5, remained fixed across scenarios. Each model was fit using the same MCMC with 60,000 iterations, discarding the first 10,000 as burn-in and retaining every tenth iteration thereafter, yielding 5,000 posterior samples for inference.

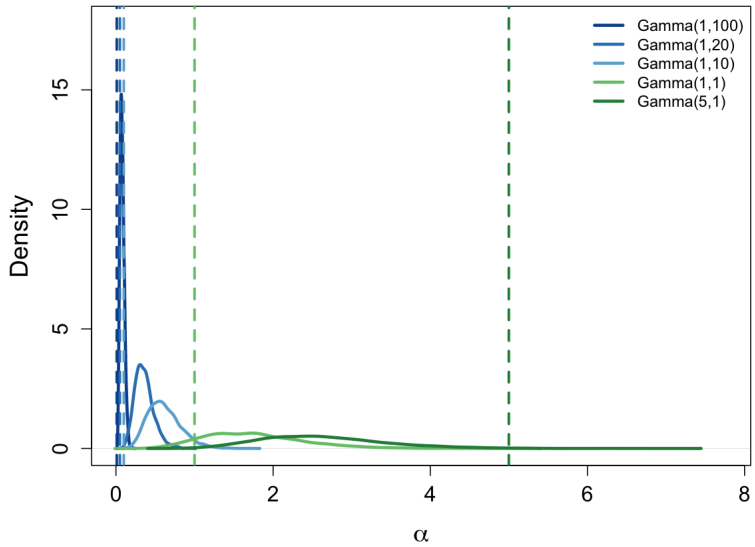


FIGURE A1: Posterior density distributions of α under five prior specifications. Dashed vertical lines indicate prior expected values $\mathsf{E}[\alpha] = a/b$.

Figure A1 displays the posterior distributions of α obtained under each prior scenario, and Table A1 summarizes the key performance metrics. As the prior expectation for α increases from 0.01 to 5.00, the posterior mean of α increases from 0.079 to 2.637. Notably, the modal number of clusters exhibits remarkable stability across scenarios, varying only from 8 to 9 groups, suggesting that the clustering structure is primarily driven by the data rather than the prior specification. Examining the model selection criteria reveals important patterns. The DIC ranges from 5,766 to 10,247 across all specifications, with $\mathsf{Gamma}(5, 1)$ showing a

suspiciously low value that may indicate numerical instability or overfitting; excluding this scenario, $\mathsf{Gamma}(1, 10)$ achieves the lowest DIC of 9,942. WAIC values remain relatively consistent across moderate prior choices, ranging from 10,470 to 11,036. The MSE shows greater variability, with $\mathsf{Gamma}(1, 1)$ yielding the lowest value of 36.36 and $\mathsf{Gamma}(5, 1)$ the highest of 80.76, yet all scenarios maintain high explanatory power with R^2 exceeding 0.98. The intra-cluster distances range from 8,855 to 10,099 and inter-cluster distances from 105 to 118, exhibiting minimal variation and confirming that the fundamental clustering structure is robust to prior specification.

TABLE A1: Performance metrics across five prior specifications for α.

Prior	**E**(α) **(prior)**	**Post.** α **(mean ± SD)**	**Mode** K	**DIC**	**WAIC**	**MSE**	R^2	**Intra-cluster**	**Inter-cluster**
$\Gamma(1, 100)$	0.01	0.079 ± 0.028	8	10240	11036	54.18	0.9898	9966.4	108.4
$\Gamma(1, 20)$	0.05	0.358 ± 0.120	9	10114	10533	51.60	0.9903	8924.7	113.0
$\Gamma(1, 10)$	0.10	0.614 ± 0.211	9	**9942**	10564	69.03	0.9871	10098.8	113.3
$\Gamma(1, 1)$	1.00	1.752 ± 0.641	9	10247	**10470**	**36.36**	**0.9932**	8859.9	105.2
$\Gamma(5, 1)$	5.00	2.637 ± 0.808	9	5766	10516	80.76	0.9849	**8855.3**	**118.4**

We select the hyperparameters $a = 1$ and $b = 10$, corresponding to $\alpha \sim \mathsf{Gamma}(1, 10)$ with $\mathsf{E}(\alpha) = 0.10$, for the main analysis as it achieves the lowest DIC of 9,942 among reliable specifications while maintaining comparable performance across other metrics including MSE, R^2, and WAIC. The posterior distribution shows substantial updating from the prior, with $\alpha = 0.614 \pm 0.211$ considerably exceeding the prior expectation of 0.10, indicating that the data actively inform the clustering structure.