POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Bayesian Models for XRF Data in Microchip Recycling Applications

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Claudio Strino, 10799640

Abstract: The recovery of valuable and critical elements from microchip residual waste requires reliable quantitative characterization of chemical compositions under severe data-scarcity constraints. In this context, chemical analyses generate multivariate count data, which can equivalently be represented in compositional form as proportions constrained to the simplex. These alternative representations call for statistical models that explicitly account for the structural features of multivariate counts and simplex-valued data.
This thesis develops Bayesian models for chemical composition data arising from printed circuit board recycling applications. Two complementary modeling perspectives are considered. When observations are available as counts, multinomial sampling schemes are adopted and compositional distributions are employed as prior models for the latent composition of the material. When observations are expressed directly as normalized compositions, the same families of distributions are used instead as likelihood functions defined on the simplex.
Classical Dirichlet formulations are taken as baseline specifications and are compared with more flexible alternatives, including the Flexible Dirichlet and Extended Flexible Dirichlet distributions, as well as the Logistic–Normal model under ILR and ALR transformations. These models allow for increasing levels of flexibility in capturing dispersion, dependence structure, and potential heterogeneity across components. Model assessment is primarily conducted through full posterior predictive distributions rather than point estimates, focusing on uncertainty quantification and predictive performance for the elemental concentrations of interest.
Finally, a Bayesian sequential sampling framework is proposed, togheter with a stopping rule that explicitly balances posterior uncertainty, expected gain, and sampling costs, providing a decision-support tool for motherboard recycling applications.

Advisor:
Alessandra Guglielmi

Academic year:
2024-2025

Key-words: Dirichlet–multinomial hierarchical model, Generaization of Dirichlet distributions, Compositional data analysis, logistic–normal model, log-ratio transformation

1. Introduction

The rapid growth of electronic devices has led to an unprecedented increase in electronic waste (e-waste), with discarded microchips representing a particularly complex and valuable component of this stream. Residual waste from microchips contains a wide range of chemical elements, including precious metals and rare earth elements, whose recovery is of both economic and environmental interest. Designing efficient recycling strategies for such materials therefore requires a reliable quantitative characterization of their chemical composition.

Within this context, the research group Materials for Energi and Environment (MAT4En2) of the Department of Chemistry, Materials and Chemical Engineering at Politecnico di Milano is involved in a research project whose ultimate goal is the recovery of pure chemical elements, such as precious metals (PM), strategic raw materials (SRM), and critical raw materials (CRM) (see Commission, Directorate-General for Internal Market, and SMEs (2023)) from end-of-life motherboards. The term *end-of-life* refers to several possible conditions under which a printed circuit board (PCBs) can no longer be used, such as mechanical failure of the device, malfunction of the motherboard itself, thermal degradation, or long-term demagnetization.

End-of-life PCBs are first subjected to a coarse cut-milling process, which reduces the material to small fragments or grains. The resulting material is then analyzed using X-ray Fluorescence (XRF), a non-destructive analytical technique used to determine elemental concentration. XRF relies on the emission of characteristic X-rays from a sample following excitation by an incident X-ray beam, allowing for the identification and quantification of the elements present (Acquafredda (2019)). While XRF requires minimal sample preparation, it is a costly technique and returns, for each measurement, a concentration vector representing the proportions of the detected elements relative to the total mass. Due to these costs, the MAT4En2 research group aims at reducing the number of XRF measurements required to accurately infer the overall chemical composition of a batch of residual waste. Being these end-of-life materials, thus wastes, they are supplied in large disomogeneous batches which ask for a large number of analysis to determine the right chemical composition. This leads to a sequential decision problem, determining when the information collected from sampled measurements is sufficient to obtain a reliable estimate of the element composition, without analyzing the entire batch. This motivates the construction of an optimal stopping strategy to balance inferential accuracy and measurement costs.

Addressing this problem requires a coherent probabilistic model for the element concentrations and its uncertainty. Depending on the representation adopted, the observations can be treated either as multivariate count vectors or as normalized compositions on the simplex. This dual perspective motivates the use of hierarchical probabilistic models for count data, alongside approaches used in Compositional Data Analysis (CoDA) for simplex-valued observations. Bayesian modeling offers a natural approach in this setting. By treating the latent composition as a random object and combining prior information with observed data, Bayesian methods allow for coherent uncertainty quantification and sequential updating as new samples become available. Within this framework, the Dirichlet distribution represents the classical baseline model for compositional data, owing to its analytical simplicity and conjugacy with multinomial sampling schemes. However, the Dirichlet distribution imposes restrictive dependence structures and may fail to capture the heterogeneity and multimodality often observed in real chemical compositions.

Several extensions of the Dirichlet model have been proposed to overcome these limitations. The Flexible Dirichlet (FD) and the Extended Flexible Dirichlet (EFD) distributions introduce additional parameters that partially decouple the mean structure, dispersion, and component-specific behavior, allowing for richer marginal variability and more complex dependence patterns. An alternative approach is provided by the Logistic–Normal model, which maps compositions to an unconstrained Euclidean space via log-ratio transformations and models them using multivariate Gaussian distributions.

1.1. Dataset description

To analyze the grinded PCBs sample, the material is partitioned into a grid of 93 squares, corresponding to the number of cells required to cover the entire batch. Each cell is treated as an atomic sampling unit and the same reference mass is assigned which is equal to one million "parts", corresponding to 1.76 g. The mass of each cell is measured using a technical balance with a nominal measurement error of ± 0.01 grams. XRF measurements are then performed on selected cells, returning elemental concentrations expressed in parts per million (ppm). Figure 1 illustrates the grid-based representation, while Table 1 lists the $K = 40$ elements that can be detected by the MAT4En2 research group, classified according to the CRM, SRM, and PM categories.

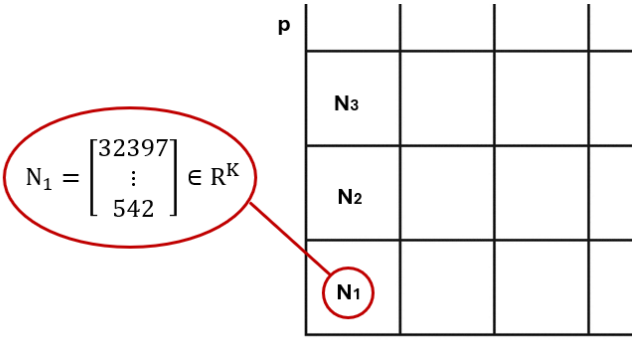


Figure 1: Grid-based representation used to structure the sampling process.

Category	**Elements**
CRM	Co, Ni, Cu, As, Sb, Ba, Ce, Nd, Bi, Nb, Ta, W, La, Y
SRM	Co, Ni, Cu, Bi, Ta, W, Ce, Nd, La
PM	Au, Ag, Pt, Pd, Rh
Other	SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5, F, S, Cl, Br, I, Cr, Zn, Sr, Zr, Sn, Pb

Table 1: Classification of the 40 analyzed elements according to the 2023 EU Critical Raw Materials (CRM), Strategic Raw Materials (SRM), and Precious Materials (PM) frameworks.

Since only a subset of the grid cells is sampled, the dataset consists of $I < 93$ observations. For each sampled cell $i = 1, \ldots, I$, the observed concentration vectors are given by $\boldsymbol{N}_i = (N_{i,1}, \ldots, N_{i,K}) \in \mathbb{R}^K$. The total mass is assumed to be constant and equal to one million parts, meaning $\sum_{k=1}^{K} N_{i,k} = 10^6$. In the hierarchical models developed in this thesis, all samples are assumed to share the same elemental composition, represented by a vector $\boldsymbol{p} \in \mathcal{S}^{K-1}$, where

$$\mathcal{S}^{K-1} = \left\{ (p_1, \ldots, p_K) : p_k \geq 0, \sum_{k=1}^{K} p_k = 1 \right\} \tag{1}$$

The vector $\boldsymbol{p}$ represents the true elemental composition of the entire batch of residual material. In particular, each p_k can be interpreted as the mean concentration of the k-th element that would be obtained from an XRF measurement of the whole batch. The observed vectors $\boldsymbol{N}_i$ are then interpreted as arising from repeated sampling and measurement variability around this latent composition. Due to the grid-based sampling construction, the observed data do not retain any explicit spatial information. The grid is introduced only as a practical tool for defining sampling

units, and the assignment of material to individual cells is arbitrary. In simple terms, the position of a cell and the specific material it contains are irrelevant for the analysis: all the material could, in principle, be completely mixed before sampling without affecting the inferential target. As a consequence, samples are not associated with spatial coordinates, no spatial dependence can be modeled, and no additional covariates are available. All observations are therefore treated as exchangeable replicates drawn from the same latent compositional structure.

1.2. Objective of the Thesis

The aim of this thesis is to investigate Bayesian models for compositional chemical data arising from motherboard waste, with particular emphasis on inferential robustness and predictive behavior in data-scarce regimes under the two complementary modeling frameworks developed. When ppm data are considered, the observed vectors $\boldsymbol{N}_i$ are modeled through multinomial sampling schemes, while the latent elemental composition of the material is assigned as prior distribution on the simplex. In the compositional-data setting, where normalized compositions are observed directly, probability distributions defined on the simplex are instead employed as likelihood functions. In this framework, the Dirichlet and its extensions are compared with the Logistic–Normal model under ILR and ALR transformations, which is used exclusively in this likelihood-based formulation.

Special attention is devoted to prior specification and model parameterization, with particular emphasis on reparameterized formulations designed to improve numerical stability and sampling efficiency in Markov Chain Monte Carlo estimation, without modifying the underlying probabilistic assumptions. Posterior inference is carried out via MCMC methods, and models are evaluated in terms of posterior predictive behavior, predictive accuracy as measured by the WAIC, and computational performance. The analysis represents a first step toward the ultimate goal of developing an optimal stopping strategy aimed at reducing the number of XRF measurements. A preliminary version of such an algorithm is presented.

1.3. Structure of the Thesis

This thesis is organized as follows. Sections 2 and 3 focus on the modeling and inference of parts-per-million count data. Section 2 introduces two flexible prior distributions specifically designed for count data, while Section 3 builds upon this framework by presenting the corresponding posterior inference given some observed data. In this context, the structure of the models is made explicit, hyperparameter choices are discussed, and details related to Markov Chain Monte Carlo sampling in Stan are provided. The resulting inferential outcomes are then presented and critically discussed. Sections 4 and 5 follow an analogous structure, but shift the attention to models defined directly on compositional observations. Section 4 adapts the same modeling framework to the simplex-valued data setting, introducing the corresponding prior distributions and posterior structures, while Section 5 is devoted to posterior predictive inference for the compositional data.

Section 6 provides an additional discussion chapter in which the results obtained from the count-based and compositional modeling approaches are compared. The comparison is carried out from both an inferential standpoint and with respect to computational performance and simulation aspects. Section 7 addresses the optimal stopping problem, the material developed in the previous chapters is used to outline an initial strategy aimed at determining when additional XRF measurements are no longer informative. Due to time constraints, this section focuses on the current implementation without introducing further methodological extensions. Finally, Section 8 concludes the thesis by summarizing the main contributions and outlining possible directions for future work aimed at improving and extending the proposed modeling framework.

2. Models on ppm data

This section presents the Bayesian modelling framework adopted for the analysis of chemical data arising from microchip recycling processes. The central object of inference throughout the chapter is the latent vector of elemental proportions $\boldsymbol{p} = (p_1, \dots, p_K) \in \mathcal{S}^{K-1}$. When modeling concentration count vectors $\boldsymbol{N}_1, \dots, \boldsymbol{N}_I$ expressed in ppm, we fix $n = \sum_{k=1}^{K} N_{i,k} = 10^6$ and represent the data through the following hierarchical model, conditional on $\boldsymbol{p}$:

$$\begin{aligned} \boldsymbol{N}_i \mid \boldsymbol{p} &\overset{\text{i.i.d.}}{\sim} \text{Multinomial}(n, \boldsymbol{p}), \\ \boldsymbol{p} &\sim \pi_p. \end{aligned} \tag{2}$$

Each p_k represents the probability of observing element k. By marginalizing over $\boldsymbol{p}$, the expectation of each count component is

$$\mathbb{E}[N_{i,k}] = \mathbb{E}[\mathbb{E}(N_{i,k} \mid \boldsymbol{p})] = \mathbb{E}(np_k) = n\,\mathbb{E}(p_k).$$

The marginal variance follows from the law of otal variance,

$$\begin{aligned} \text{Var}(N_{i,k}) &= \mathbb{E}[\text{Var}(N_{i,k} \mid \boldsymbol{p})] + \text{Var}(\mathbb{E}(N_{i,k} \mid \boldsymbol{p})) = \mathbb{E}[np_k(1-p_k)] + \text{Var}(np_k) \\ &= n\big(\mathbb{E}(p_k) - \mathbb{E}(p_k^2)\big) + n^2\,\text{Var}(p_k). \end{aligned} \tag{3}$$

For $k \neq r$, the marginal covariance is obtained analogously as

$$\begin{aligned} \text{Cov}(N_{i,k}, N_{i,r}) &= \mathbb{E}[\text{Cov}(N_{i,k}, N_{i,r} \mid \boldsymbol{p})] + \text{Cov}(\mathbb{E}(N_{i,k} \mid \boldsymbol{p}), \mathbb{E}(N_{i,r} \mid \boldsymbol{p})) \\ &= -\,n\,\mathbb{E}(p_k p_r) + n^2\,\text{Cov}(p_k, p_r) \\ &= n(n-1)\,\text{Cov}(p_k, p_r) - n\,\mathbb{E}(p_k)\mathbb{E}(p_r). \end{aligned} \tag{4}$$

Further details are reported in Appendix A. Due to the positivity constraint on all components of $\boldsymbol{p}$, the sign of the marginal covariances is directly driven by the prior dependence structure of $\boldsymbol{p}$. Inference in this setting targets the posterior distribution of $\boldsymbol{p}$ induced by the combination of the multinomial likelihood and a prior distribution defined in the simplex.

The first prior tested for $\boldsymbol{p}$ is the Dirichlet model, due to its simplicity; its structure is recalled in Appendix B. Two increasingly flexible generalizations are then introduced: the Flexible Dirichlet (FD, 2.1) distribution and the Extended Flexible Dirichlet (EFD, 2.2) distribution, both designed to capture more complex dependence structures among the components of $\boldsymbol{p}$.

2.1. The Flexible Dirichlet distribution

The Flexible Dirichlet (FD) distribution, introduced in Ongaro and Migliorati (2012), generalizes the Dirichlet distribution by partially decoupling the mean structure from the dependence structure, while preserving analytical tractability. Let

$$Y_k = W_k + Z_k U, \qquad k = 1, \dots, K,$$

where $W_k \overset{\text{i.i.d.}}{\sim} \text{Ga}(\alpha_k, 1)$, $U \sim \text{Ga}(\tau, 1)$, and $\boldsymbol{Z} \sim \text{Multinomial}(1, \boldsymbol{q})$. The random vector

$$\boldsymbol{p} = \frac{\boldsymbol{Y}}{\sum_{k=1}^{K} Y_k}$$

is said to follow an $\text{FD}_K(\boldsymbol{\alpha}, \boldsymbol{q}, \tau)$ distribution.A key property of the FD distribution is its finite mixture representation,

$$f_{\text{FD}}(\boldsymbol{p}) = \sum_{k=1}^{K} q_k\, f_{\text{Dir}}(\boldsymbol{p}; \boldsymbol{\alpha} + \tau \boldsymbol{e}_k), \tag{5}$$

which greatly facilitates both interpretation and posterior computation. Let $\alpha_+ = \sum_{k=1}^K \alpha_k$ and let $\boldsymbol{p} = (p_1, \ldots, p_K)$ follow a Flexible Dirichlet distribution. For arbitrary non–negative integers $n_1, \ldots, n_K$, the joint moments shown in Ongaro and Migliorati (2012) are given by

$$\mathbb{E}\left[\prod_{k=1}^K p_k^{n_k}\right] = \frac{1}{(\alpha_+ + \tau)^{[n_+]}} \prod_{k=1}^K \alpha_k^{[n_k]} \sum_{j=1}^K q_j \frac{(\alpha_j + \tau)^{[n_j]}}{\alpha_j^{[n_j]}}.$$

The first moment and the variance are then given by

$$\mathbb{E}(p_k) = \frac{\alpha_k + q_k\tau}{\alpha_+ + \tau}, \qquad \mathrm{Var}(p_k) = \frac{\mathbb{E}(p_k)(1 - \mathbb{E}(p_k))}{\alpha_+ + \tau + 1} + \frac{\tau^2 q_k(1 - q_k)}{(\alpha_+ + \tau)(\alpha_+ + \tau + 1)}.$$

For $k \neq r$, the covariance is

$$\mathrm{Cov}(p_k, p_r) = -\frac{\mathbb{E}(p_k)\mathbb{E}(p_r)}{\alpha_+ + \tau + 1} - \frac{\tau^2 q_k q_r}{(\alpha_+ + \tau)(\alpha_+ + \tau + 1)}.$$

Although pairwise covariances remain negative, the FD distribution allows for substantially larger magnitudes compared to the Dirichlet model, yielding a more flexible dependence structure and potentially multimodal behavior. The Dirichlet distribution is recovered as a special case of the FD when $\tau = 1$ and $q_k = \frac{\alpha_k}{\alpha_+}$.

2.2. The Extended Flexible Dirichlet distribution

The Extended Flexible Dirichlet (EFD) distribution, introduced by Ongaro and Migliorati (2020), further generalizes the FD model by allowing for component–specific inflation parameters, thereby enabling both negative and positive dependence among components. Let

$$Y_k = W_k + Z_k U_k, \qquad k = 1, \ldots, K,$$

with $W_k \overset{\text{i.i.d.}}{\sim} \mathrm{Ga}(\alpha_k, \beta)$, $U_k \overset{\text{i.i.d.}}{\sim} \mathrm{Ga}(\tau_k, \beta)$, and $\boldsymbol{Z} \sim \mathrm{Multinomial}(1, \boldsymbol{q})$. The random vector

$$\boldsymbol{p} = \frac{\boldsymbol{Y}}{\sum_{k=1}^K Y_k}$$

follows an $\mathrm{EFD}_K(\boldsymbol{\alpha}, \boldsymbol{q}, \boldsymbol{\tau})$ distribution, its density can be represented as a finite mixture

$$f_{\mathrm{EFD}}(\boldsymbol{p}) = \sum_{k=1}^K q_k \, f_{\mathrm{Dir}}(\boldsymbol{p}; \boldsymbol{\alpha} + \tau_k \boldsymbol{e}_k). \tag{6}$$

It is immediate to see that the FD distribution is a special case of the EFD when $\tau_1 = \tau_2 = \cdots = \tau_K$, and that it further collapses to a Dirichlet distribution as discussed in Section 2.1. Let $\boldsymbol{p} = (p_1, \ldots, p_K)$ follow an Extended Flexible Dirichlet distribution with parameters $\boldsymbol{\alpha}$, $\boldsymbol{\tau}$ and $\boldsymbol{q}$. For arbitrary non–negative integers $n_1, \ldots, n_K$, the joint moments (Ongaro and Migliorati, 2020) are given by

$$\mathbb{E}\left[\prod_{k=1}^K p_k^{n_k}\right] = \prod_{k=1}^K \alpha_k^{[n_k]} \sum_{j=1}^K q_j \frac{(\alpha_j + \tau_j)^{[n_j]}}{\alpha_j^{[n_j]}(\alpha_+ + \tau_j)^{[n_+]}}, \qquad \alpha_+ = \sum_{r=1}^K \alpha_r.$$

The expected value admits the closed–form expression

$$\mathbb{E}(p_i) = \frac{\alpha_i}{\alpha_+} k_1 + \frac{\tau_i}{\alpha_+ + \tau_i} q_i, \qquad k_1 = \sum_{r=1}^K q_r \frac{\alpha_+}{\alpha_+ + \tau_r}.$$

The full covariance structure of the EFD distribution, conditional on $\boldsymbol{q}$ and for $j \neq r$, is given by

$$\begin{aligned}\mathrm{Cov}(p_j, p_r \mid q_j, q_r) = \alpha_j \alpha_r \left(k_2 - k_1^2\right) - q_j q_r \frac{\tau_j \tau_r}{(\alpha_+ + \tau_j)(\alpha_+ + \tau_r)} \\ + \frac{\alpha_j p_r \tau_r}{\alpha_+ + \tau_r} \left(\frac{1}{\alpha_+ + \tau_r + 1} - k_1\right) + \frac{\alpha_r p_j \tau_j}{\alpha_+ + \tau_j} \left(\frac{1}{\alpha_+ + \tau_j + 1} - k_1\right),\end{aligned}$$

with

$$k_2 = \sum_{r=1}^{K} \frac{q_r}{(\alpha_+ + \tau_r)(\alpha_+ + \tau_r + 1)}.$$

The EFD distribution allows for positive covariances whenever sufficient heterogeneity is present among the τ_i parameters, providing a substantial increase in modeling flexibility.

2.3. Alternative prior specification and full conditionals

This section specifies the Bayesian models for ppm data introduced in the previous sections. The models differ only for the prior assumed for $\boldsymbol{p}$, the vector of the elements mean concentration. For each model, prior distributions are assigned to all hyperparameters, and the resulting posterior structure is discussed. Since the models involve latent variables and non–conjugate components, posterior distributions are not available in closed form. Inference is therefore carried out using Markov Chain Monte Carlo methods. The full conditional distributions underlying the sampling schemes are reported to clarify the inferential structure and to motivate the computational implementation.

2.3.1 Posterior under the Dirichlet prior

Under a Dirichlet prior on $\boldsymbol{p}$ with parameter $\boldsymbol{\alpha}$, the model is

$$\begin{aligned}\boldsymbol{N}_i \mid \boldsymbol{p} &\overset{\text{i.i.d.}}{\sim} \text{Multinomial}(n, \boldsymbol{p}), \\ \boldsymbol{p} &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K).\end{aligned} \tag{7}$$

For a fixed $\boldsymbol{\alpha}$, conjugacy between the multinomial likelihood and the Dirichlet prior yields a closed–form posterior distribution, that is

$$\boldsymbol{p} \mid \boldsymbol{N}_i, \ldots, \boldsymbol{N}_I, \boldsymbol{\alpha} \sim \text{Dirichlet}\left(\boldsymbol{\alpha} + \sum_{i=1}^{I} \boldsymbol{N}_i\right).$$

During this work, alternative prior specifications are also considered by assigning a prior distribution to $\boldsymbol{\alpha}$. In this case, the hierarchical model is completed by assuming

$$\alpha_k \overset{\text{i.i.d.}}{\sim} \text{Ga}(a, b), \qquad k = 1, \ldots, K,$$

with fixed hyperparameters (a, b). Since conjugacy is no longer available, posterior inference is carried out using MCMC methods.

2.3.2 Posterior under the Flexible Dirichlet prior

When $\boldsymbol{p}$ follows a Flexible Dirichlet prior, posterior inference can be formulated by introducing a latent allocation variable $\boldsymbol{z}$. The conditional posterior distributions associated with this representation are derived in Ascari, Migliorati, and Ongaro (2019). The resulting FD–

Multinomial hierarchical model is defined as

$$\begin{aligned}
\boldsymbol{N}_i \mid \boldsymbol{p} &\overset{\text{i.i.d.}}{\sim} \text{Multinomial}(n, \boldsymbol{p}),\\
\boldsymbol{p} \mid \boldsymbol{z} = \boldsymbol{e}_j &\sim \text{Dirichlet}(\boldsymbol{\alpha} + \tau \boldsymbol{e}_j),\\
\mathbb{P}(\boldsymbol{z} = \boldsymbol{e}_j) &= q_j,\\
\boldsymbol{q} &\sim \text{Dirichlet}(\beta_1, \ldots, \beta_K),\\
\alpha_k &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(a_k, b_k), \qquad \tau \sim \text{Gamma}(c, d).
\end{aligned}$$

For computational reasons, the latent allocation variable $\boldsymbol{z}$ is marginalized out. In particular, the model is implemented in `Stan`, which does not support discrete latent variables. Integrating out $\boldsymbol{z}$ yields a continuous posterior density that can be handled by using MCMC. The resulting marginal likelihood takes the form

$$\mathcal{L}(\{\boldsymbol{N}_i\}, \boldsymbol{p} \mid \boldsymbol{\alpha}, \tau, \boldsymbol{q}) = \left[\prod_{i=1}^{I} \frac{n!}{\prod_{k=1}^{K} N_{i,k}!} \prod_{k=1}^{K} p_k^{N_{i,k}} \right] \sum_{j=1}^{K} q_j \, \text{Dirichlet}(\boldsymbol{p} \mid \boldsymbol{\alpha} + \tau \boldsymbol{e}_j)\,. \tag{8}$$

Posterior inference for the FD–Multinomial model relies on the derivation of the full conditional distributions for each unknown quantity. In a Gibbs sampling framework (Casella and George, 1992; Gilks, 1996), the full conditional distribution of a generic parameter θ is the probability density of θ given all other parameters and the observed data. Formally, let $\boldsymbol{\Theta}$ denote the full set of model parameters and $\mathcal{D}$ the observed data. The full conditional density for a specific parameter $\theta \in \boldsymbol{\Theta}$, denoted as $p(\theta \mid \boldsymbol{\Theta}_{-\theta}, \mathcal{D})$, is proportional to the joint posterior distribution considering only the terms involving θ:

$$p(\theta \mid \text{rest}) \;\propto\; p(\mathcal{D}, \boldsymbol{\Theta}) \;\propto\; \text{likelihood} \times \text{prior structure}.$$

In hierarchical Bayesian models, this simplifies significantly thanks to conditional independence properties. Specifically, for a hyperparameter such as the concentration parameter α_k, the full conditional is derived by combining its prior density with the density of the variables it directly governs. In the FD context, $\boldsymbol{\alpha}$ governs the latent composition vector $\boldsymbol{p}$. Thus, the full conditional for α_k is proportional to the product of the conditional density of $\boldsymbol{p}$ and the prior on α_k:

$$p(\alpha_k \mid \boldsymbol{p}, \boldsymbol{z}, \tau) \;\propto\; p(\boldsymbol{p} \mid \boldsymbol{z}, \boldsymbol{\alpha}, \tau) \cdot \pi(\alpha_k),$$

where $p(\boldsymbol{p} \mid \cdot)$ acts as the likelihood contribution for α_k, and $\pi(\alpha_k)$ represents its prior density. Note that the data $\{\boldsymbol{N}_i\}$ do not appear explicitly in this expression as $\boldsymbol{\alpha}$ is conditionally independent of the data given $\boldsymbol{p}$. When this product matches the kernel of a known distribution, the parameter can be sampled directly; otherwise, as is the case for $\boldsymbol{\alpha}$ and τ, Metropolis-Hastings steps or other numerical sampling methods are required.

For the FD model, full conditional distributions are conveniently derived from the hierarchical representation involving the latent allocation variable $\boldsymbol{z}$. Although $\boldsymbol{z}$ is marginalized out, its explicit introduction is instrumental for analytical purposes, as it reveals the conditional structure of the model. The resulting full conditional distributions are derived in Ascari, Migliorati, and Ongaro (2019) and are reported below for completeness.

$$\begin{aligned}
\boldsymbol{p} \mid (\boldsymbol{z} = \boldsymbol{e}_j), \boldsymbol{\alpha}, \tau, \{\boldsymbol{N}_i\} &\sim \text{Dirichlet}\left(\boldsymbol{\alpha} + \tau \boldsymbol{e}_j + \sum_{i=1}^{I} \boldsymbol{N}_i \right),\\
\mathbb{P}(\boldsymbol{z} = \boldsymbol{e}_j \mid \boldsymbol{p}, \boldsymbol{q}, \boldsymbol{\alpha}, \tau) &= \frac{q_j \, \text{Dirichlet}(\boldsymbol{p} \mid \boldsymbol{\alpha} + \tau \boldsymbol{e}_j)}{\sum_{k=1}^{K} q_k \, \text{Dirichlet}(\boldsymbol{p} \mid \boldsymbol{\alpha} + \tau \boldsymbol{e}_k)},\\
\boldsymbol{q} \mid \boldsymbol{z} &\sim \text{Dirichlet}(\beta_1, \ldots, \beta_j + 1, \ldots, \beta_K).
\end{aligned}$$

The remaining parameters, namely $\boldsymbol{\alpha}$ and τ, are associated with more complex posterior distributions, for which closed–form expressions are not available.

2.3.3 Posterior under Extended Flexible Dirichlet prior

Under the Extended Flexible Dirichlet prior, the hierarchical model, as well as the posterior structure is analogous to the Flexible Dirichlet case, with component-specific inflation parameters τ_j.

$$\begin{aligned}
\boldsymbol{N}_i \mid \boldsymbol{p} &\sim^{i.i.d.} \text{Multinomial}(n, \boldsymbol{p}),\\
\boldsymbol{p} \mid \boldsymbol{z} = \boldsymbol{e}_j &\sim \text{Dirichlet}(\boldsymbol{\alpha} + \tau_j \boldsymbol{e}_j),\\
\mathbb{P}(\boldsymbol{z} = \boldsymbol{e}_j) &= q_j,\\
\boldsymbol{q} &\sim \text{Dirichlet}(\beta_1, \ldots, \beta_K),\\
\alpha_k \overset{\text{i.i.d.}}{\sim} \text{Gamma}(a_k, b_k), &\qquad \tau_k \overset{\text{i.i.d.}}{\sim} \text{Gamma}(c, d).
\end{aligned}$$

The marginal likelihood is

$$\mathcal{L}(\{\boldsymbol{N}_i\}, \boldsymbol{p} \mid \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{q}) = \left[\prod_{i=1}^{I} \frac{n!}{\prod_{k=1}^{K} N_{i,k}!} \prod_{k=1}^{K} p_k^{N_{i,k}}\right] \sum_{j=1}^{K} q_j \, \text{Dirichlet}(\boldsymbol{p} \mid \boldsymbol{\alpha} + \tau_j \boldsymbol{e}_j)\,. \tag{9}$$

The full conditional distributions are analogous to the FDM case, with the scalar inflation parameter replaced by τ_j, and defined accordingly they are updated via MCMC.

3. Posterior Inference on ppm data

The purpose of this section is to report posterior inference obtained under the different priors introduced in Section 2 for different sample sizes. The analysis focuses on posterior predictive behavior and parameter estimates, while also investigating how posterior uncertainty evolves as more samples become available.

All Bayesian models were fitted in Stan (CmdStan 2.38.0 interfaced via cmdstanpy in Python). Two parallel Markov chains were run for each dataset, with 1,000 warmup iterations and 2,000 post-warmup iterations per chain, namely sampling iterations, yielding 4,000 posterior draws in total. We used the default No-U-Turn Sampler (NUTS). In some scenarios the target acceptance statistic `adapt_delta` was increased up to 0.95 or 0.98 to improve exploration of more challenging posterior geometries. Convergence of the MCMC chains was assessed using the $\hat{R}$ statistic (requiring $\hat{R} < 1.01$) and effective sample size; representative traceplots and additional visual diagnostics are reported in Appendix C. All simulations were run on an OMEN by HP Laptop 15-dh1xxx equipped with an Intel Core i7-10750H CPU (6 cores, 12 threads) running at 2.60 GHz, 16 GB of RAM, and Microsoft Windows 11 Home as operating system.

3.1. Hyperparameters prior specification

To ensure comparability across models, several fixed prior configurations are considered, including homogeneous, heterogeneous, and mixed specifications. In the mixed setting, some parameters follow homogeneous priors, while others are assigned heterogeneous ones. Table 2 summarizes the main homogeneous fixed prior configurations used throughout the analysis. Under this specification, all parameters are assumed to be independently and identically distributed according to the distributions reported in the table.

Table 2: Homogeneous fixed prior configurations used.

Configuration	α_k	τ / τ_k	β_k
Flat 1	Gamma$(1, 1)$	Gamma$(1, 10)$	1
Flat 2	Gamma$(2, 0.5)$	Gamma$(2, 0.5)$	1
Moderate	Gamma$(10, 1)$	Gamma$(2, 0.2)$	3
Informative 1	Gamma$(20, 2)$	Gamma$(5, 0.5)$	5
Informative 2	Gamma$(50, 2.5)$	Gamma$(5, 0.5)$	5

The assumption of identically distributed parameters is subsequently relaxed by introducing a heterogeneous fixed prior specification. The selected configurations are reported in Table 3. Based on a preliminary inspection of the data, each component is assigned to one of three groups, labelled *Weak*, *Medium*, or *Strong*. The grouping is determined by the cumulative concentration observed in the first three samples. Specifically, two thresholds δ_1 and δ_2 are selected such that

$$\frac{\sum_{i=1}^{3} N_{i,k}^{\text{Weak}}}{3n} < \delta_1 < \frac{\sum_{i=1}^{3} N_{i,k}^{\text{Medium}}}{3n} < \delta_2 < \frac{\sum_{i=1}^{3} N_{i,k}^{\text{Strong}}}{3n}, \quad n = \sum_{k=1}^{K} N_{i,k}$$

For each group, the shape and rate parameters of the Gamma prior for α_k are chosen so as to approximate the empirical mean and dispersion of the corresponding components.

Table 3: Heterogeneous fixed prior configurations used.

Group	α_k	τ/τ_k	β_k
Weak	Gamma$(1, 1)$	Gamma$(1, 1)$	1
Medium	Gamma$(4, 2)$	Gamma$(5, 0.5)$	3
Strong	Gamma$(50, 2.5)$	Gamma$(5, 0.2)$	5

Finally, mixed prior configurations are explored by combining homogeneous and heterogeneous specifications across different parameters. All combinations considered in the analysis are summarised in Table 4.

Table 4: Mixed prior configurations considered in the analysis. The label "mixed" indicates that the corresponding parameter is assigned a heterogeneous prior, as defined in Table 3, while non-mixed parameters are assigned homogeneous priors.

Name	α prior	q prior	τ/τ_k prior
`alpha_heter`	mixed	Dirichlet$(1, \ldots, 1)$	Gamma$(5, 0.5)$
`q_heter`	Gamma$(20, 2)$	mixed	Gamma$(5, 0.5)$
`tau_heter`	Gamma$(20, 2)$	Dirichlet$(1, \ldots, 1)$	mixed
`both_alpha_tau`	mixed	Dirichlet$(1, \ldots, 1)$	mixed
`all`	mixed	mixed	mixed

Convergence diagnostics and effective sample sizes were inspected for all prior configurations. Overall, more stable behavior was observed under homogeneous fixed priors, particularly for configurations Flat 2 and more informative. An additional experiment using exchangeable priors was also performed; however, it resulted in substantially poorer convergence and mixing compared to the fixed prior specifications.

3.2. Posterior Inference for categorical data

Posterior predictive distributions of the latent categorical probabilities $\boldsymbol{p}$ are first examined after observing 3 and 6 samples. Results shown here refer to the *Informative 1* prior configuration with β_k fixed to 1 (Table 2). Figure 2 shows, under DM model, $\boldsymbol{p}$ posterior distributions for three representative components: a highly abundant component (Cu), a weakly observed component (Br), and a never-observed component (La). All never-observed components yield identical posterior distributions under the homogeneous prior specification. This assumption could be relaxed by incorporating prior information on the manufacturing composition of the material, as documented in the relevant literature. Overall, posterior inference under the Dirichlet–Multinomial model is largely driven by the concentration parameters α_k, which are updated coherently as pseudo-counts reflecting the observed frequencies in the data.

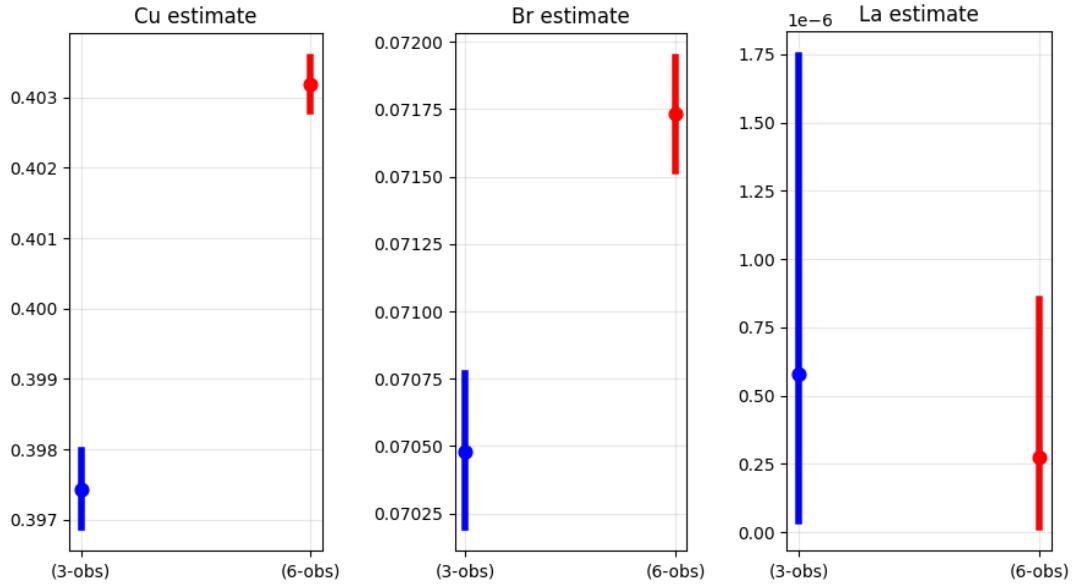


Figure 2: Posterior distributions for three representative components p_k after 3 and 6 observations using DM model: a highly abundant component (left), a weakly observed component (center), and a never-observed component (right).

Figure 3 shifts the focus to the posterior distributions of the same components of $\boldsymbol{p}$ under the Flexible Dirichlet–Multinomial (FDM) model. Results obtained under the Extended Flexible Dirichlet–Multinomial (EFDM) specification are qualitatively analogous and are therefore reported in Appendix D. When moving to these extension formulations, the introduction of the inflation parameters τ and τ_k is intended to capture component-specific deviations from multinomial sampling. Posterior inference shows, however, that these additional parameters tend to compensate for, rather than complement, the concentration parameters. In particular, for the copper component (Cu), α_{Cu} is consistently estimated to be smaller than those of other components, while the corresponding τ (or τ_{Cu} in the EFDM model) increases in order to reproduce the pronounced peak observed in the data. This behavior suggests that the contribution of Cu is primarily encoded through the inflation mechanism rather than through an increase in marginal concentration.

In the EFDM model, this effect is further reinforced by the fact that the τ_k parameters associated with non-dominant components exhibit negligible posterior variability, remaining close to their prior distributions. As a consequence, the extended categorical formulations do not provide a substantial gain in effective flexibility under data scarcity. Instead, they tend to collapse to behavior that is indistinguishable from the standard Dirichlet–Multinomial model, a feature that is also reflected in their very similar posterior distributions.

4. Bayesian models on compositional data

We assume now a different approach: the dara are now proportions, i.e. compositions, obtained by normalizing the observed count vectors $\boldsymbol{N}_i$. Specifically, for each sample i, the compositional vector is defined as

$$\theta_{i,j} = \frac{N_{i,j}}{\sum_{k=1}^{K} N_{i,k}}, \qquad \boldsymbol{\theta}_i = (\theta_{i,1}, \ldots, \theta_{i,K}) \in \mathcal{S}^{K-1},$$

where $\mathcal{S}^{K-1}$ denotes the $(K-1)$-dimensional probability simplex, as defined in (1). This transformation removes the information about the total amount of material and keeps only the relative proportions between components. As a result, each vector describes the elemental composition

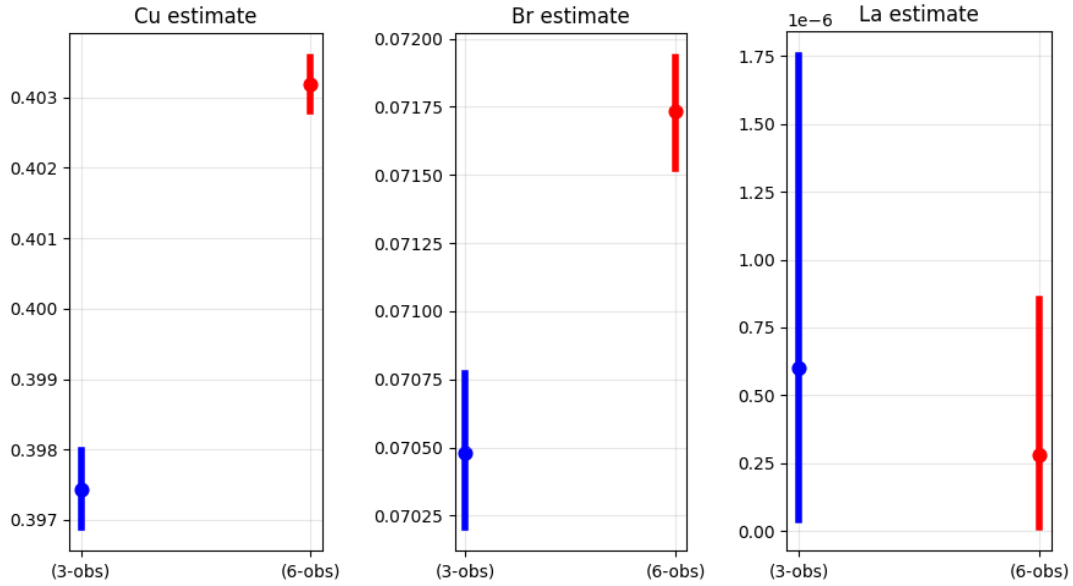


Figure 3: Posterior distributions of three representative components p_k after 3 and 6 observations under the Flexible Dirichlet–Multinomial model: a highly abundant component (left), a weakly observed component (center), and a never-observed component (right).

of a sample independently of its total mass. This is appropriate for our application, since the goal is to estimate the average concentration of each element in the overall material, rather than the absolute quantity measured in each individual subsample.

The modeling framework adopted in this section differs fundamentally from the one introduced for count data. In the count-based formulation, the latent categorical probabilities $\boldsymbol{p}$ represent global parameters shared across samples and inferred through posterior distributions. In contrast, in the compositional setting the vectors $\boldsymbol{\theta}_i$ are directly observed (up to measurement uncertainty) and are therefore treated as random variables rather than fixed parameters. Consequently, posterior inference no longer targets $\boldsymbol{\theta}$ itself, but instead focuses on $\boldsymbol{\theta}$ predictive distribution under the assumed probabilistic model.

The following sections introduce the probabilistic models defined on the simplex. The Dirichlet, Flexible Dirichlet (FD, see Section 2.1), and Extended Flexible Dirichlet (EFD, see Section 2.2) distributions previously discussed are considered again in this setting and are used as likelihood functions for the observed compositional data. In addition, an alternative specification is introduced, namely the Logistic–Normal distribution.

4.1. The Logistic–Normal distribution

The Logistic–Normal distribution, originally introduced by Atchison and Shen (1980), provides a flexible alternative for modeling random compositions by mapping unconstrained Gaussian random vectors to the simplex through log-ratio transformations. Unlike Dirichlet-based models, the Logistic–Normal framework allows for an unconstrained covariance structure among components, thereby enabling richer dependence patterns. In this work, inference under the Logistic–Normal model is conducted by transforming the observed compositions using log-ratio coordinates and specifying a multivariate Gaussian model on the transformed space. Two log-ratio transformations are considered: the additive log-ratio (ALR) and the isometric log-ratio (ILR) transformations.

The additive log-ratio transformation, introduced by Aitchison (1982), maps a composition $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK}) \in \mathcal{S}^{K-1}$ to an unconstrained vector in $\mathbb{R}^{K-1}$ by selecting a reference com-

ponent $r \in \{1, \ldots, K\}$ and forming log-ratios with respect to it:

$$\mathrm{alr}_r(\boldsymbol{x}_i) = \left(\log \frac{x_{i1}}{x_{ir}}, \ldots, \log \frac{x_{i,r-1}}{x_{ir}}, \log \frac{x_{i,r+1}}{x_{ir}}, \ldots, \log \frac{x_{iK}}{x_{ir}} \right).$$

The choice of the reference component for the additive log-ratio (ALR) transformation follows the criteria proposed in the Appendix of Greenacre et al. (2023). Specifically, two complementary aspects are considered. First, this assessment is based on the Procrustes correlation, which compares the sample configuration obtained under a given ALR transformation with the configuration induced by the full log-ratio geometry. The indicator quantifies how well the ALR representation preserves the relative distances between compositions: larger values of the Procrustes correlation correspond to ALR mappings that are closer to an isometric representation of the simplex (Grunsky, Greenacre, and Kjarsgaard (2024)). Second, the variability of the logarithm of the reference component is examined, since a low variance of $\log(x_{\mathrm{ref}})$ implies that individual ALR coordinates behave approximately as additive shifts of $\log(x_j)$, improving numerical stability and interpretability. The final choice reflects a trade-off between near-isometry and interpretability, rather than being driven solely by component abundance or marginal variance.

Components that are never observed across the dataset are excluded from the reference selection procedure. Such components exhibit null variance and would otherwise be systematically selected as denominators, leading to numerical instabilities and undefined log-ratios. Their exclusion ensures well-defined transformations and stable inference.

The isometric log-ratio transformation, introduced by Egozcue et al. (2003), overcomes the main limitations of the ALR by providing orthonormal coordinates in $\mathbb{R}^{K-1}$ that are invariant to the choice of a reference component. Given a composition $\boldsymbol{x}_i \in \mathcal{S}^{K-1}$, the ILR transformation is defined as

$$\mathrm{ilr}(\boldsymbol{x}_i) = \log(\boldsymbol{x}_i)^\top \mathbf{H},$$

where $\log(\boldsymbol{x}_i)$ denotes the component-wise logarithm of the composition and $\mathbf{H}$ is a $K \times (K-1)$ orthonormal basis matrix spanning the simplex subspace of zero-sum vectors.

In this work, the matrix $\mathbf{H}$ is constructed using the Helmert sub-matrix, obtained by removing the first column of the full Helmert matrix. This choice yields an orthonormal basis consistent with the Aitchison geometry and provides a simple and canonical implementation of the ILR transformation (Egozcue et al. (2003); Tsagris, Preston, and Wood (2011); Ankam and Bouguila (2018)). Different orthonormal bases lead to equivalent representations in terms of distances and likelihood, and the Helmert basis is adopted here for computational convenience. Moving to the core of the model, let $\boldsymbol{y}_i \in \mathbb{R}^{K-1}$ denote either the ALR or ILR transformed coordinates of the composition $\boldsymbol{x}_i$. Under the Logistic–Normal model, the transformed observations are assumed to follow a multivariate Gaussian distribution,

$$\boldsymbol{y}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}_{K-1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{10}$$

where $\boldsymbol{\mu}$ is a location vector and $\boldsymbol{\Sigma}$ is a positive definite covariance matrix. Posterior inference is conducted on the unconstrained parameter space, while posterior predictive compositions are obtained by applying the inverse log-ratio transformations and normalizing to the simplex.

The Logistic–Normal model is considered alongside the reparameterized Flexible Dirichlet and Extended Flexible Dirichlet models in order to provide a complementary perspective on compositional variability and dependence structures.

4.2. Predictive probability formulation

In the compositional data setting, inference is formulated in terms of the posterior predictive distribution of a new composition $\boldsymbol{\theta}_{\mathrm{new}}$. Let $\boldsymbol{p} \sim f(\cdot \mid \boldsymbol{\phi})$ denote a generic probabilistic model defined on the simplex and parametrized by $\boldsymbol{\phi}$. The posterior predictive distribution is given by

$$p(\boldsymbol{\theta}_{\mathrm{new}} \mid \boldsymbol{\theta}_{1:I}) = \int f(\boldsymbol{\theta}_{\mathrm{new}} \mid \boldsymbol{\phi})\, p(\boldsymbol{\phi} \mid \boldsymbol{\theta}_{1:I})\, d\boldsymbol{\phi}.$$

This formulation is independent of the specific choice of f and therefore includes the Dirichlet, Flexible Dirichlet, Extended Flexible Dirichlet, and Logistic–Normal models. In practice, the integral is approximated via Markov Chain Monte Carlo (MCMC) methods implemented in Stan, having the same configuration as introduce in Section 3, with posterior predictive draws obtained from the `generated_quantities` block. Under this direct compositional modeling approach, the observed data are treated as proportions, and the original count information is not explicitly modeled. As a consequence, the presence of zero components requires a preprocessing step prior to normalization, as zero values are not compatible with compositional inference in general, and not only with log-ratio transformations. Given the large reference mass of each sample, of the order of 10^6, components that are never observed in a sample are replaced by a unit count. For components that are observed at least once within the sample, zero entries are instead replaced with $\frac{2}{3}$ of the minimum strictly positive observed value. This choice follows a non-parametric multiplicative replacement strategy commonly adopted in compositional data analysis to handle zeros while preserving the relative structure of the composition (Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2003)). The magnitude of the replaced values is explicitly monitored, and in all cases the total mass introduced by zero replacement remains below 0.1% of the overall abundance, ensuring a negligible impact on inference.

During preliminary analyses, posterior inference under the original parameterization of the Flexible Dirichlet (FD) and Extended Flexible Dirichlet (EFD) models proved to be unreliable across a wide range of prior specifications. In particular, standard Markov Chain Monte Carlo implementations exhibited convergence issues and unsatisfactory sampling behavior, preventing a stable and interpretable inferential analysis. For these reasons, the results presented in this work are based on reparameterized formulations of the FD and EFD models. The detailed diagnostic assessment of the original parameterization, together with a systematic investigation of the reparameterized models and their inferential behavior, is deferred to Section 5.1.

4.2.1 Reparameterized Flexible Dirichlet model

The Flexible Dirichlet (FD) model is adopted in its reparameterized formulation (Ascari, Migliorati, and Ongaro (2019);A. Di Brisco, Migliorati, et al. (2017)), in order to improve interpretability and numerical stability. The original parameters are replaced by the triplet $(\boldsymbol{\mu}, \phi, \boldsymbol{q})$ and an auxiliary parameter controlling the separation among mixture components. In particular, the global precision parameter is defined as

$$\phi = \alpha^+ + \tau, \qquad \alpha^+ = \sum_{k=1}^{K} \alpha_k,$$

while the mean composition on the simplex is given by

$$\boldsymbol{\mu} = \frac{\boldsymbol{\alpha}}{\phi} + \tilde{w}\,\boldsymbol{q}, \qquad \tilde{w} = \frac{\tau}{\phi},$$

with $\boldsymbol{\mu} \in \mathcal{S}^{K-1}$. To ensure variation independence and well-defined Dirichlet components, the auxiliary parameter $\tilde{w}$ is rescaled to obtain

$$w = \frac{\tilde{w}}{\min_j \min\left\{\dfrac{\mu_j}{q_j}, 1\right\}}, \qquad w \in (0,1),$$

leading to the final parameterization $(\boldsymbol{\mu}, w, \phi, \boldsymbol{q})$. Under this formulation, prior distributions can be specified directly on quantities with clear statistical meaning:

$$\boldsymbol{\mu} \sim \text{Dirichlet}(e_0, \ldots, e_0), \qquad \boldsymbol{q} \sim \text{Dirichlet}(d_0, \ldots, d_0),$$

$$\phi \sim \text{Gamma}(g_1, g_2), \qquad w \sim \text{Beta}(a_w, b_w),$$

yielding weakly informative priors that decouple prior information on the mean composition, dispersion, and mixture structure.

4.2.2 Reparameterized Extended Flexible Dirichlet model

The Extended Flexible Dirichlet (EFD) model is adopted in its reparameterized formulation, as proposed by Ascari, A. M. Di Brisco, et al. (2024), in order to improve numerical stability and sampling efficiency. The model is expressed in terms of an overall mean composition, mixture weights, a global precision parameter, and component-specific displacement parameters. This reparameterization reduces posterior dependence among the original parameters $(\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{q})$ and enables prior specification on quantities with a direct statistical interpretation.

Let $\boldsymbol{x} \in \mathcal{S}^{K-1}$ denote a random composition. Under the mixture representation, conditional on the latent allocation variable $z \in \{1, \dots, K\}$ with $\mathbb{P}(z = r) = q_r$, the EFD model can be written as

$$\boldsymbol{x} \mid (z = r) \sim \mathrm{Dir}_{\mathrm{mp}}(\boldsymbol{\lambda}_r,\ \kappa_r)\,, \qquad \kappa_r = \alpha^+ + \tau_r, \qquad \alpha^+ = \sum_{k=1}^{K} \alpha_k,$$

where the component-specific mean is defined as

$$\boldsymbol{\lambda}_r = (1 - \tilde{w}_r)\,\bar{\boldsymbol{\alpha}} + \tilde{w}_r\,\boldsymbol{e}_r, \qquad \tilde{w}_r = \frac{\tau_r}{\alpha^+ + \tau_r} \in (0, 1),$$

and $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}/\alpha^+ \in \mathcal{S}^{K-1}$ denotes the common barycenter on the simplex. The overall mean of the EFD distribution satisfies

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{x}) = \sum_{r=1}^{K} q_r \boldsymbol{\lambda}_r,$$

which links the barycenter $\bar{\boldsymbol{\alpha}}$ to the parameters $(\boldsymbol{\mu}, \boldsymbol{q}, \boldsymbol{w})$. To ensure variation independence and automatically satisfy the mixture–mean constraints, the displacement parameters are rescaled component-wise as

$$w_j = \frac{\tilde{w}_j}{\min\{\mu_j/q_j,\ 1\}}, \qquad w_j \in (0, 1), \quad j = 1, \dots, K.$$

Inference can therefore be conducted in terms of the parameter set

$$(\boldsymbol{\mu}, \boldsymbol{q}, \alpha^+, \boldsymbol{w}),$$

from which the original EFD parameters can be recovered when needed. Under this reparameterized formulation, independent priors may be assigned as

$$\boldsymbol{\mu} \sim \mathrm{Dirichlet}(e_0, \dots, e_0), \qquad \boldsymbol{q} \sim \mathrm{Dirichlet}(d_0, \dots, d_0),$$

$$\alpha^+ \sim \mathrm{Gamma}(g_1, g_2), \qquad w_j \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(a_w, b_w),$$

yielding weakly informative priors that separate prior information on the overall mean composition, global dispersion, mixture weights, and component-specific deviations from the barycenter.

4.2.3 Logistic–Normal model prior specification

Posterior inference for the Logistic–Normal model is carried out using Markov Chain Monte Carlo methods implemented in Stan, consistently with the other models considered in this work. Inference is performed on the log-ratio transformed coordinates, while posterior predictive compositions are obtained through inverse transformations.

Although the Logistic–Normal model allows for a fully unconstrained covariance structure on the transformed space, estimating a full covariance matrix of dimension $(K-1) \times (K-1)$ proved numerically unstable in the present application, due to the limited sample size and the relatively high dimensionality of the problem. For this reason, a diagonal covariance structure is adopted, which substantially improves computational stability while still allowing for component-specific variability on the log-ratio scale. Under the diagonal covariance specification, the Logistic–Normal model is parameterized by a location vector $\boldsymbol{\mu} \in \mathbb{R}^{K-1}$ and a vector of scale parameters $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_{K-1})$, where each $\sigma_k > 0$ controls the marginal variability of the corresponding log-ratio coordinate.

Independent priors are assigned to the model parameters. Specifically, the mean vector is assigned a multivariate Normal prior,

$$\boldsymbol{\mu} \sim \mathcal{N}_{K-1}\left(\boldsymbol{\alpha},\, \tau_\mu^2 \mathbf{I}\right),$$

where τ_μ^2 denotes a fixed variance hyperparameter controlling the overall dispersion of the prior mass in the transformed space. For the scale parameters, independent half-Normal priors are adopted,

$$\sigma_k \overset{\text{i.i.d.}}{\sim} \text{Half-Normal}(\tau_\sigma), \qquad k = 1, \ldots, K-1.$$

with scale hyperparameter $\tau_\sigma > 0$. This prior specification regularizes the variability of individual log-ratio coordinates while ensuring positivity of the scale parameters. The assumption of prior independence between $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, together with the diagonal covariance structure, reduces posterior dependence and improves numerical stability in high-dimensional settings with limited sample size.

5. Predictive Inference on compositional data

Posterior inference in the compositional data setting is carried out under several competing models: the standard Dirichlet model with fixed *Informative 1* priors, the reparameterized Flexible Dirichlet (FD) and Extended Flexible Dirichlet (EFD) models, and the Logistic–Normal (LN) model. For the standard Dirichlet specification, posterior predictive summaries are reported for three representative components of $\boldsymbol{\theta}$, corresponding to low-, medium-, and high-abundance regimes, under the *Informative 1* prior setting (Figure 4). As the number of observations increases, the predictive distributions progressively concentrate around the empirical composition.

Under the Dirichlet model, posterior predictive uncertainty is driven almost exclusively by the compositional geometry, without introducing component-specific sources of variability or additional dispersion mechanisms. As a result, predictive credible intervals may appear comparatively wide. This behavior should not be interpreted as a lack of informativeness of the model, but rather as an intrinsic feature of posterior predictive inference in the compositional framework. Indeed, for the Dirichlet model and more generally for all models considered in the compositional setting, the posterior predictive distribution is obtained by averaging over all plausible parameter values supported by the posterior. This integration naturally leads to wider uncertainty bands when compared to posterior summaries of the model parameters themselves, even in the absence of explicit overdispersion or mixture components.

5.1. Original parametrization issues

During preliminary analyses based on the original parameterization of the Flexible Dirichlet (FD) and Extended Flexible Dirichlet (EFD) models, substantial computational and inferential difficulties were encountered, as introduced in Section 4.2. In particular, Markov Chain Monte Carlo sampling frequently exhibited poor mixing behavior for the inflation parameters τ (and τ_k in the EFD case), together with unstable and erratic trajectories for several concentration parameters α_k. These issues persisted across a wide range of prior specifications, including moderately to strongly informative Gamma priors, and were consistently reflected in unsatisfactory convergence diagnostics.

Beyond mixing and convergence issues, the inferential role of the mixture weights $\boldsymbol{q}$ proved difficult to interpret. In many configurations, the posterior distribution of $\boldsymbol{q}$ remained effectively indistinguishable from the prior, indicating limited learning of the mixture structure from the data. In other cases, posterior mass was allocated to mixture components associated with extremely low-abundance elements, yielding interpretations that were difficult to reconcile with the observed compositions. For the Extended Flexible Dirichlet model, some prior configurations produced posterior summaries for (τ_k, q_k) that appeared plausible when considered marginally, but only at the cost of extremely poor chain mixing. Alternative specifications improved sampling behavior but resulted in posterior distributions concentrated near the prior, effectively collapsing the model to a standard Dirichlet formulation.

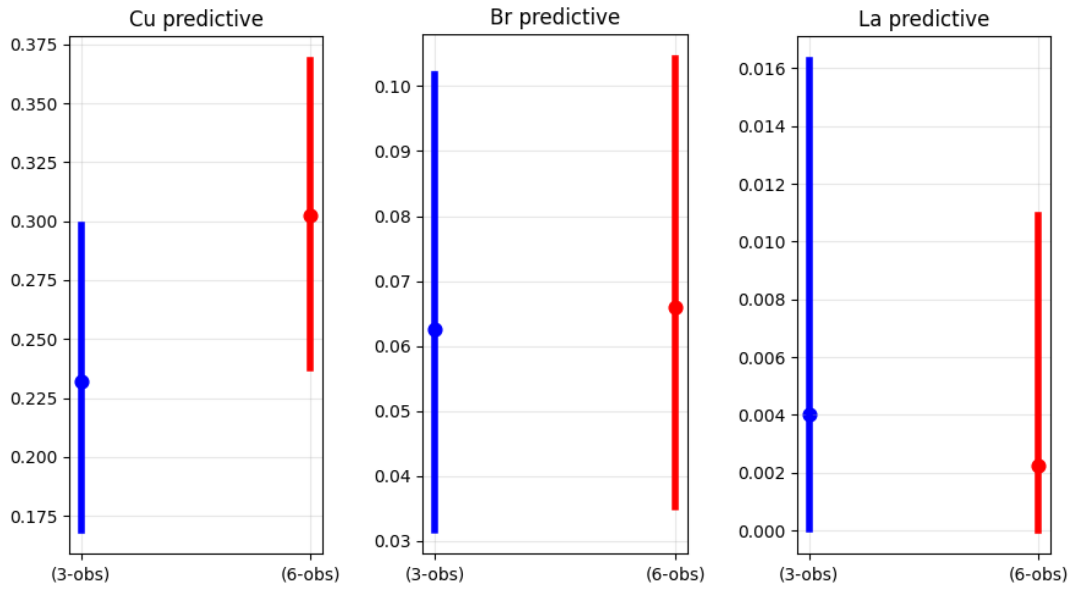


Figure 4: Predictive distributions for three representative components θ_k after 3 and 6 observations using Dirichlet model: a highly abundant component (left), a weakly observed component (center), and a never-observed component (right).

For these reasons, inference in this work is conducted using reparameterized formulations of the FD and EFD models. The alternative parameterizations are designed to decouple the global mean structure, overall dispersion, and mixture components, thereby improving chain mixing, and computational stability. Importantly, this reparameterization preserves the probabilistic structure of the original models.

5.2. Posterior predictive analysis under the FD distribution

Based on an extensive exploratory analysis, a reference prior configuration for the reparameterized Flexible Dirichlet (FD) model is identified as yielding the most stable convergence diagnostics and satisfactory MCMC mixing. The selected specification is

$$\begin{aligned} \boldsymbol{\mu} &\sim \text{Dirichlet}(5,\ldots,5), & \boldsymbol{q} &\sim \text{Dirichlet}(1,\ldots,1), \\ \phi &\sim \text{Gamma}(25,\ 0.5), & w &\sim \text{Beta}(1,20). \end{aligned} \tag{11}$$

where the Dirichlet prior on $\boldsymbol{\mu}$ is centered on a uniform composition, the mixture weights $\boldsymbol{q}$ are a priori uniform, the precision parameter ϕ is relatively informative, and the normalized perturbation weight w is concentrated near zero. This choice reflects a conservative prior belief favoring limited inflation effects while preserving flexibility in the baseline composition.

Posterior predictive distributions for selected components under this baseline specification are reported in Figure 5. The figure illustrates the predictive behavior of the FD model for elements belonging to different abundance regimes: a highly abundant component, a weakly observed component, and a never-observed component. Even as additional observations are incorporated, predictive uncertainty remains substantial, especially for low-abundance and unobserved elements, indicating limited contraction of predictive intervals.

Fixing this reference prior configuration, we further investigate the impact of the Beta hyperparameters (a_w, b_w) governing the perturbation weight. Distinct inferential regimes emerge depending on their relative magnitude. When $b_w > a_w$ and the two parameters are relatively close (e.g., ratios $b_w : a_w = 15:1$ or closer), chain mixing deteriorates substantially and posterior inference becomes dominated by a single spiked realization of the mixture weights. In this regime, most of the perturbation mass is allocated to the most frequently observed component,

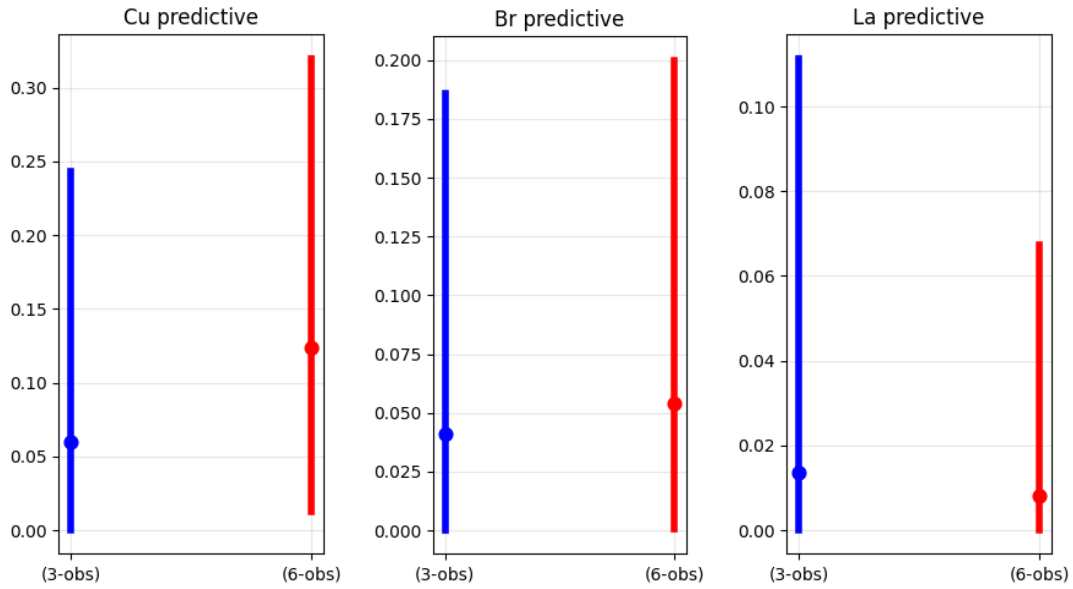


Figure 5: Predictive distributions for three representative components θ_k after 3 and 6 observations using FD model: a highly abundant component (left), a weakly observed component (center), and a never-observed component (right).

such as Cu, leading to poor posterior exploration and unreliable inference. Conversely, when $a_w > b_w$ (e.g., ratios $a_w : b_w = 8:1$ or $3:1$), chains exhibit improved mixing and run in parallel, but often converge to qualitatively different posterior modes. In particular, one chain may allocate most of the perturbation mass to q_{Cu}, while another favors $q_{\mathrm{SiO_2}}$, inducing apparent multimodality in the posterior distribution of $\boldsymbol{q}$.

Overall, the most reliable inferential behavior is obtained under the Beta$(1, 20)$ prior on w. Under this specification, the posterior distribution of $\boldsymbol{q}$ remains substantially flatter than its prior counterpart and the perturbation mass τ stays close to zero. Posterior distributions of selected mixture weights, namely q_{Cu} and q_{La}, corresponding to components at opposite abundance regimes, are reported in Figure 6, showing negligible posterior differentiation across elements.

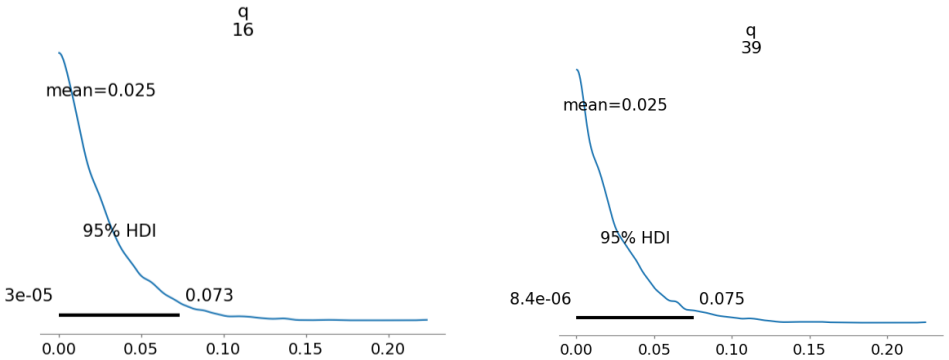


Figure 6: Posterior distribution of q_{Cu}, highest abundace element (left) vs q_{La}, never seen element (right)

These results indicate that, under the reparameterized FD model, posterior behavior is strongly influenced by the prior specification of the normalized perturbation parameter w, which directly controls the magnitude of the inflation parameter τ. While reparameterization improves numerical stability and sampling efficiency, it is not sufficient on its own to prevent the model

from collapsing toward a Dirichlet-like regime. When priors are chosen to induce substantial posterior mass on τ, the resulting mixture structure becomes unstable and posterior predictive distributions are unreliable. Conversely, when priors favor small values of τ, posterior inference converges toward an almost Dirichlet regime, with mixture weights $\boldsymbol{q}$ remaining largely uninfluenced by the data and predictive intervals failing to contract meaningfully as additional observations are incorporated.

5.3. Posterior predictive analysis under the EFD distribution

A similar sensitivity to prior specification is observed when adopting the Extended Flexible Dirichlet (EFD) model; however, the induced behavior appears overall more stable than in the FD case. Owing to its additional layer of flexibility, the EFD specification is able to accommodate heterogeneous perturbation patterns while preserving satisfactory convergence properties over a wider range of prior configurations.

Under the EFD model, the prior specification mirrors that of the FD model, with the exception of the mean parameter and the total concentration. In particular, the adopted priors are defined as

$$\begin{aligned} \boldsymbol{\mu} &\sim \text{Dirichlet}(5,\ldots,5), & \boldsymbol{q} &\sim \text{Dirichlet}(1,\ldots,1), \\ \alpha_+ &\sim \text{Gamma}(25,\, 0.5), & w_j &\overset{\text{i.i.d.}}{\sim} \text{Beta}(1,20). \end{aligned} \tag{12}$$

where the Dirichlet prior on $\boldsymbol{\mu}$ is centered on a uniform composition, the mixture weights $\boldsymbol{q}$ are a priori uniform, the total concentration parameter α_+ carries strong prior information, and the perturbation weight $\boldsymbol{w}$ remains centered near zero. Figure 7 reports posterior predictive distributions for selected components of $\boldsymbol{\theta}$ under alternative prior configurations for the perturbation weight.

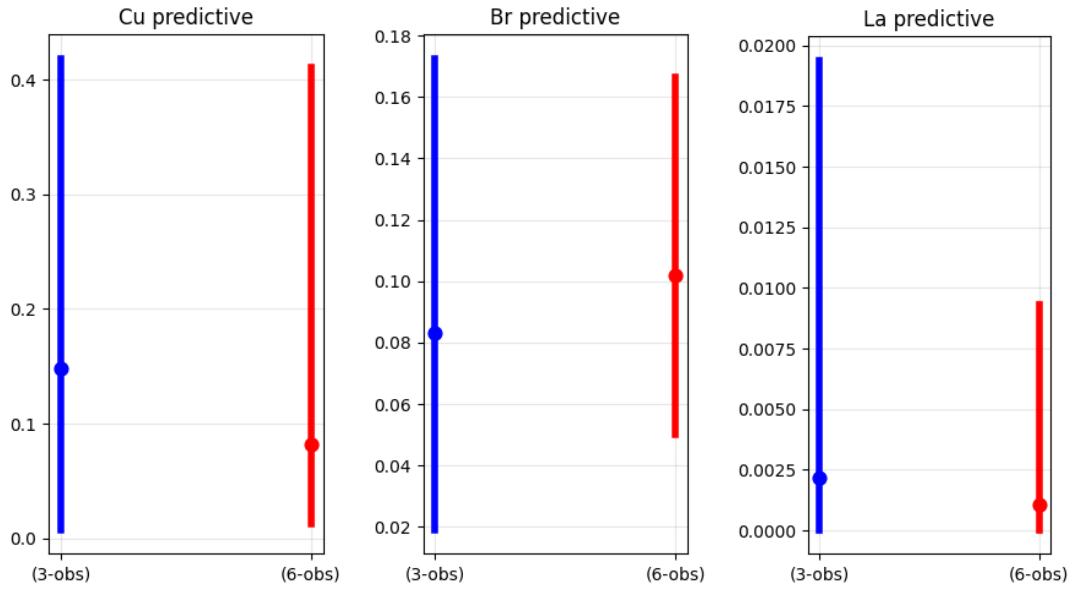


Figure 7: Posterior predictive distributions for representative components θ_k under the EFD model and alternative prior configurations for the perturbation weight.

As in the FD case, the relative magnitude of the Beta hyperparameters (a_w, b_w) plays a key role in shaping posterior inference, although the resulting effects are more regularized under the EFD specification. When $b_w > a_w$ and the ratio assumes moderate-to-large values, the posterior distribution of the perturbation mass exhibits a stable spiked behavior, with most of the mass being allocated to the most frequently observed component, namely Cu. In contrast to the FD case, this regime is characterized by well-mixed and stable chains, indicating that the EFD model

can sustain a dominant perturbation component without compromising posterior exploration. To illustrate this effect, Figure 8 reports posterior distributions of selected components of $\boldsymbol{q}$, focusing on q_{Cu} and q_{La}, which represent elements at opposite abundance regimes. Unlike the FD posterior analysis (Figure 6), this behavior appears more consistent with the underlying data structure.

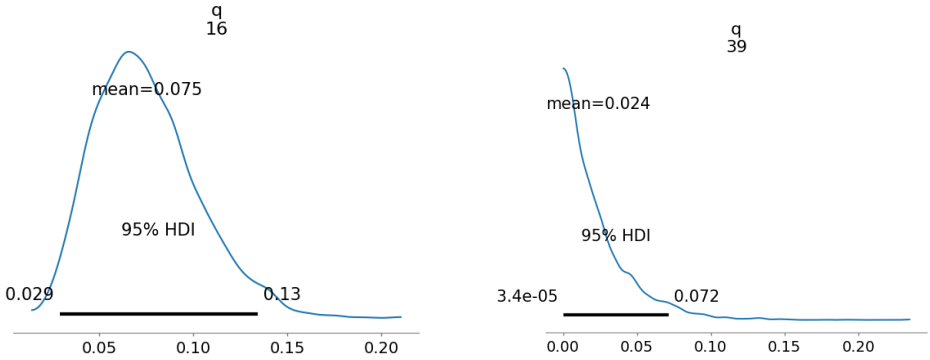


Figure 8: Posterior distribution for some component of $\boldsymbol{q}$, under EFD distribution, in case of spiker component (left) and with a unchanged one (right)

Conversely, when $a_w > b_w$, instability arises already for moderate ratios (e.g., $a_w : b_w = 3:1$). In this setting, chains exhibit good internal mixing but converge to distinct posterior modes. In particular, some chains allocate most of the perturbation mass to q_{Cu}, while others favor $q_{\mathrm{Al_2O_3}}$, leading to parallel chains exploring different regions of the posterior support. This result, shown in Appendix E, in an unreliable posterior analysis and, consequently, in distorted predictive distributions. For these reasons, prior configurations with $a_w > b_w$ are excluded from further consideration.

The reparameterized Extended Flexible Dirichlet (EFD) model exhibits a more structured and stable allocation of flexibility. In particular, configurations in which posterior mass on the τ_k parameters remains smaller than that of the corresponding α_k lead to more coherent and interpretable inference. A notable feature of the EFD model is the emergence of a stable bimodality in the posterior predictive distribution of the Cu component, shown in Figure 9. While this behavior may give the visual impression of excessively wide predictive intervals in aggregate summaries (e.g., Figure 7), it is more appropriately interpreted as the coexistence of two distinct predictive regimes. This behavior is consistent with the properties of the EFD model discussed in Ongaro and Migliorati (2020). Specifically, one regime corresponds to an inflated Cu component, with predictive mass concentrated around values close to 0.4, while the alternative regime reflects a non-inflated scenario with substantially lower Cu abundance. In each case, the remaining components adapt accordingly to satisfy the compositional constraint.

An additional advantage of the reparameterized EFD model is the increased variability across the τ_k parameters. Unlike the EFDM model, where minor components often retain posterior distributions close to their priors, the EFD specification allows for more pronounced component-specific differentiation, as illustrated in Figure 10. In principle, such heterogeneity in the τ_k parameters could induce non-trivial dependence structures among components Ongaro and Migliorati (2020). However, across all EFD configurations considered in this work, posterior predictive covariance matrices do not exhibit statistically significant positive off-diagonal entries. This suggests that dependence among components is still primarily induced by the simplex constraint, rather than by explicit correlation mechanisms encoded in the model.

Overall, these results indicate that, unlike the FD model, the EFD specification does not collapse to the standard Dirichlet model even under prior configurations favoring small perturbations. Instead, the EFD model preserves a distinct posterior structure, yielding posterior predictive distributions that remain clearly separated from the baseline Dirichlet case across different prior settings.

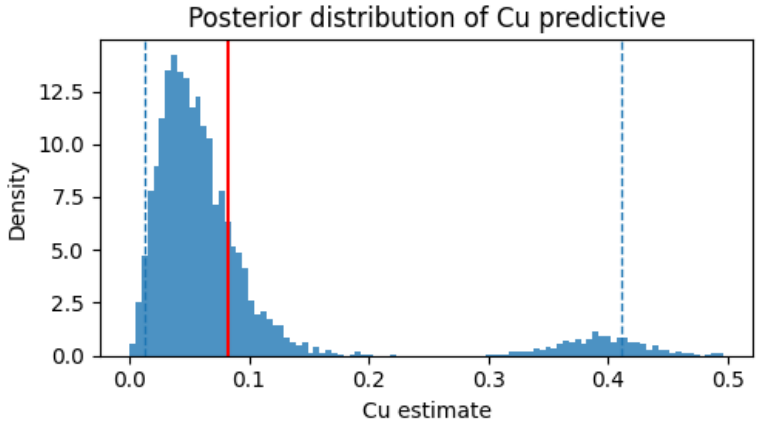


Figure 9: Posterior predictive distribution for the Cu component under the EFD model, highlighting the presence of bimodality and the coexistence of inflated and non-inflated regimes.

5.4. Posterior predictive analysis under LN distribution

Posterior predictive summaries for the Logistic–Normal (LN) model are reported in Figure 12. Both additive log-ratio (ALR) and isometric log-ratio (ILR) representations are considered to assess the robustness of posterior inference with respect to the chosen coordinate system. For both transformations, the prior specification is defined as

$$\boldsymbol{\mu} \sim \mathcal{N}(0,\, 5), \qquad \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{K-1}^2), \qquad \sigma_k \overset{\text{i.i.d.}}{\sim} \text{Half-Normal}(0, 2), \qquad k = 1, \ldots, K-1.$$

where a diagonal covariance structure is adopted to reduce numerical complexity while preserving stable convergence and effective chain mixing.

For the ALR representation, posterior predictive inference depends critically on the choice of the reference component. Following a data-driven procedure proposed in the literature, elements exhibiting structural zeros are removed prior to reference selection, as they induce excessively large denominators in the log-ratio transformation and result in unstable inference. Applying this procedure, Figure 11 component 7 (Na_2O) is identified as a suitable reference, yielding stable posterior trajectories, satisfactory mixing, and predictive behavior closely aligned with that obtained under the ILR transformation.

Posterior inference under the ILR transformation exhibits consistently good mixing and stable convergence across all chains, independently of the specific components involved. To facilitate a direct comparison between the two representations, Figure 12 reports paired posterior predictive summaries under ALR and ILR transformations for selected components spanning different abundance regimes, including highly abundant (Cu), intermediate high (Br), intermediate low (Nd) and low-abundance (La) elements.

The comparison with Logistic–Normal models offers an alternative perspective on model flexibility. By operating in an unconstrained Euclidean space through log-ratio transformations, the LN framework allows for richer covariance structures than those achievable under Dirichlet-based compositional models. Although prior specification is less intuitive due to reduced parameter interpretability, posterior inference consistently recovers non-trivial covariance patterns. In particular, under the isometric log-ratio (ilr) transformation, weak but positive posterior covariance emerges between selected component pairs, such as SiO_2 and Al_2O_3. According to domain knowledge provided by chemical engineers, this behavior is consistent with material usage in PCBs production, where these compounds may either partially substitute one another or be jointly employed due to shared properties. From this perspective, the observed positive posterior covariance is physically plausible and illustrates the ability of the log-normal (LN) model to capture co-variation structures that remain inaccessible to Dirichlet-based approaches.

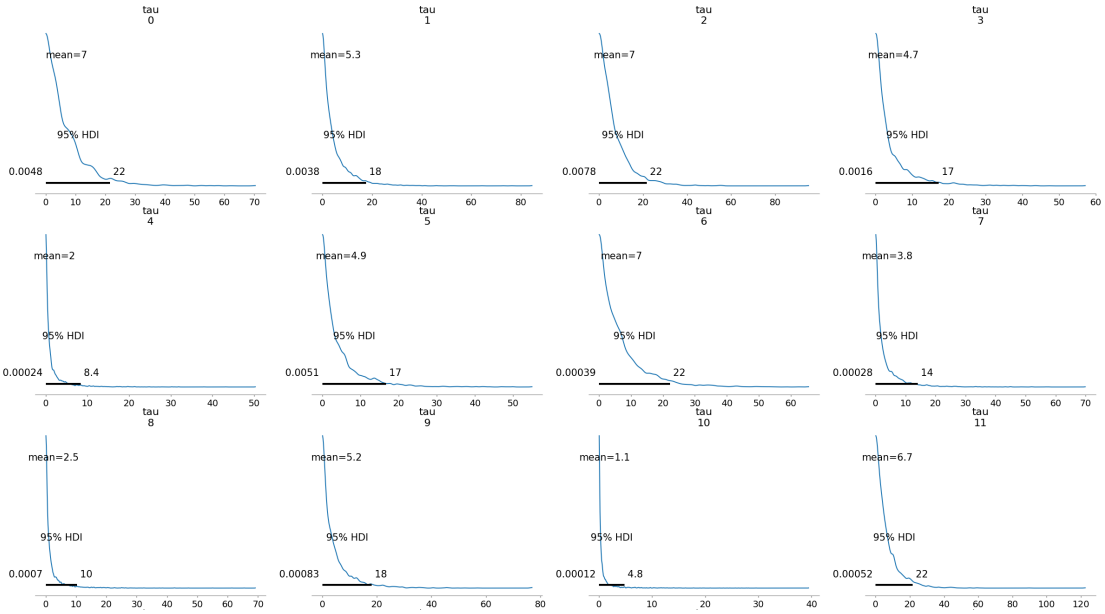


Figure 10: Posterior distributions of the component-specific inflation parameters τ_k under the EFD model, with $w_j \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, 20)$.

Moreover, the posterior predictive behaviour under the LN model appears more stable as the number of observations increases, making it a promising candidate for sequential data acquisition strategies and stopping-rule algorithms based on predictive convergence. At the same time, credible intervals remain wider than in the categorical setting, yet substantially narrower than those obtained under Dirichlet-based compositional extensions, providing a balanced representation of the high heterogeneity characterizing the data.

6. Discussion

This section provides a comparative discussion of the proposed modeling approaches, focusing on computational efficiency, predictive performance, and qualitative posterior behavior. The comparison is carried out across both categorical (ppm-based) and compositional modeling frameworks under identical simulation settings. A first comparison concerns computational cost. All runtimes reported below correspond to the total time required to complete a full predictive analysis under identical MCMC settings (see Section 3), including two consecutive runs based on 3 and 6 observed samples, respectively. This choice reflects the practical cost associated with producing predictive summaries such as Figures 2, 3, 4, 5, 7, and 12. Within the categorical framework, the Dirichlet–Multinomial model exhibits relatively low computational cost, with a total runtime of approximately 43 seconds. In contrast, its flexible extensions are substantially more demanding: both the Flexible Dirichlet–Multinomial (8) and the Extended Flexible Dirichlet–Multinomial (9) models require on the order of 6 minutes and 20–30 seconds, reflecting the increased complexity induced by mixture components and inflation parameters.

Within the compositional framework, computational requirements are generally lower for simpler models. The standard Dirichlet model completes the full predictive analysis in approximately 2 seconds, while the reparameterized Flexible Dirichlet (11) requires around 54–57 seconds. The reparameterized Extended Flexible Dirichlet (12) again incurs a substantially higher cost, with runtimes of about 6 minutes and 30 seconds. Logistic–Normal models (10) remain computationally efficient even when accounting for both runs. The additive log-ratio (alr) and the isometric log-ratio (ilr) formulation runs both in about 4–4.3 seconds. These runtimes remain negligible compared to those of the flexible Dirichlet-based models, despite the additional linear transformations involved.

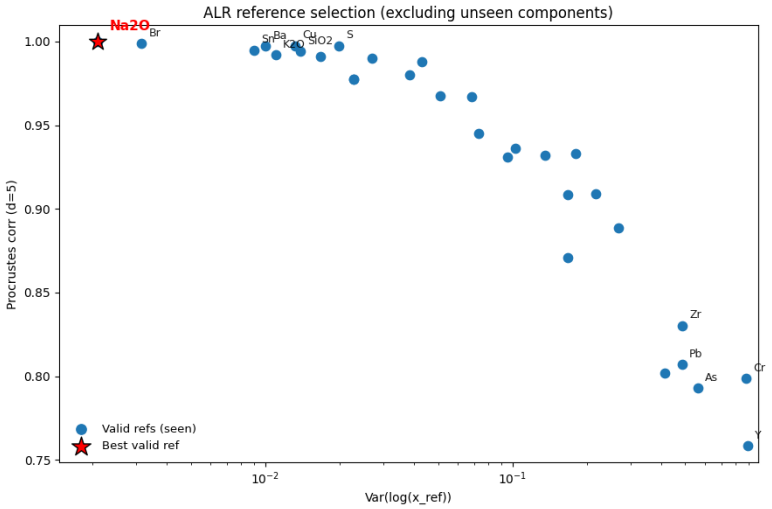


Figure 11: ALR reference selection based on the Procrustes correlation between PCA scores under ALR and CLR geometries. The red star denotes the selected reference component (Na_2O).

Moving to predictive comparison, model comparison based on predictive criteria further supports the conclusions drawn from posterior analysis. The Watanabe–Aikake Information Criterion (WAIC, Watanabe and Opper (2010)) was computed for each model configuration. Within the compositional framework, the Dirichlet and Flexible Dirichlet models yield comparable WAIC values (on the order of ≈ 1600), indicating similar predictive performance. The Extended Flexible Dirichlet model consistently improves predictive fit, achieving lower WAIC values (approximately `waic_efd` ≈ 1550), reflecting the benefit of component-specific flexibility.

A substantial improvement is observed under the Logistic–Normal model, for which WAIC values are dramatically lower, ranging approximately between -50 and -200. This reduction highlights the superior predictive capability of the Logistic–Normal formulation within the compositional setting. It is important to emphasize, however, that WAIC values obtained under categorical and compositional likelihoods are not directly comparable, due to the fundamentally different data representations and likelihood structures.

Within the categorical framework, WAIC values are several orders of magnitude larger, with values around -9×10^4 for the Dirichlet–Multinomial model and around -10^5 for its flexible extensions. While these values confirm improvements within the categorical class, they should not be contrasted numerically with those obtained in the compositional setting.

Beyond predictive scores, the models differ markedly in their treatment of uncertainty. Models fitted to ppm count data tend to produce very narrow posterior credible intervals, which may in some cases appear overly optimistic. In contrast, models defined directly on compositional data yield wider posterior and predictive intervals. While this behavior more faithfully reflects uncertainty at the composition level, it also implies that predictive intervals may expand rapidly as the prediction horizon increases. In particular, when considering predictions far ahead in the sequence (e.g., for $\boldsymbol{\theta}_{I+m}$ with large m), predictive uncertainty may grow substantially. This aspect is especially relevant in view of the optimal stopping problem addressed later in this work, as the choice of stopping rule may depend critically on the rate at which predictive uncertainty increases.

Among the compositional models considered, the Logistic–Normal formulation with ilr transformation exhibits the most stable predictive behavior. Compared to Dirichlet-based alternatives, predictive trajectories under the ilr parameterization are more centralized and display reduced oscillations when comparing results with differnt number of sample observed. This results in more stable posterior summaries and smoother predictive distributions. These properties, together with its favorable computational and predictive performance, make the Logistic–Normal

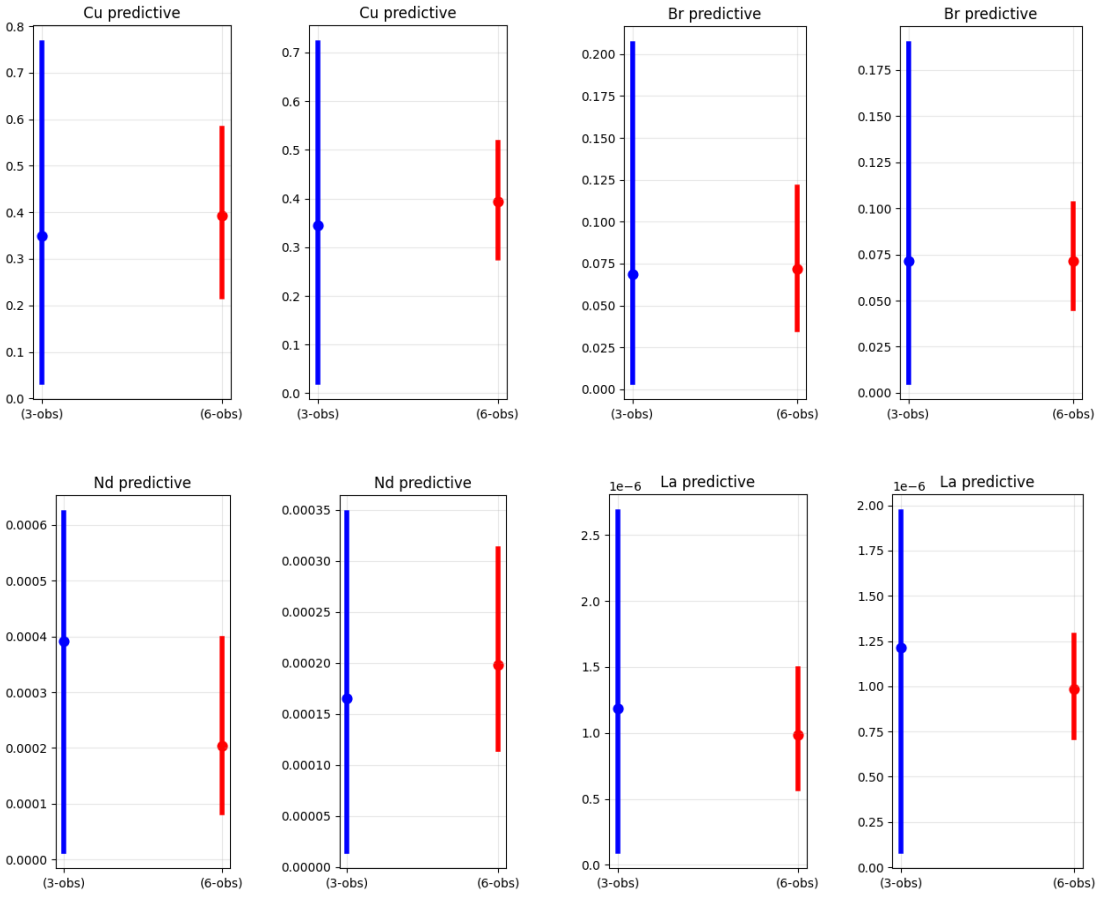


Figure 12: Posterior predictive summaries under the Logistic–Normal model comparing isometric log-ratio (ILR, left panels) and additive log-ratio (ALR, right panels) representations for selected components across different abundance regimes.

model with ilr transformation a robust choice for an optimal stopping algorithm which could use several predictive sampling.

7. Optimal Stopping

In microchip recycling applications, the collection and analysis of additional samples typically entails non-negligible laboratory costs and processing time. For this reason, it is natural to adopt a sequential sampling framework in which samples are collected one at a time and a decision rule determines whether further sampling is economically justified.

The stopping rule considered in this work is based on the decision-theoretic framework proposed by Rasmussen and Starr (1979) in the context of species discovery. In the present application, a stopping rule based on undiscovered elements may be overly conservative. From an operational perspective, the objective is not to identify all chemical elements, but rather to quantify the presence of a specific subset of elements that are of economic or environmental relevance.

Let $\mathcal{R} \subset 1, \ldots, K$ denote the set of *elements of interest*, namely those CRM, SRM, or PM elements of particular relevance to the MAT4En2, hereafter referred to as *red elements* (see Table 5). For each element k, let p_k^h denote its true proportion in the batch, and let $\mathbb{E}(p_k \mid \mathcal{F}_n)$

be the Bayesian estimator obtained after n samples. We define

$$u(n) = \sum_{k=1}^{K} \mathbb{E}(p_k \mid \mathcal{F}_n)\ \mathbb{I}[k \in \mathcal{R}],$$

as the posterior estimate of the total probability mass associated with elements in $\mathcal{R}$. More generally, to account for heterogeneous relevance across elements, a weighted version can be introduced:

$$u(n) = \sum_{k=1}^{K} w_k\, \mathbb{E}(p_k \mid \mathcal{F}_n)\ \mathbb{I}[k \in \mathcal{R}],$$

where $w_k \geq 0$ reflects the relative economic or environmental importance of element k.

Category	**Elements**
CRM	Co, Ni, Cu, Y, Ce, Nd, Nb, Ta, W, La
SRM	Co, Ni, Cu, Ce, Nd, Ta, W, La
PM	Au, Ag, Pt, Pd, Rh

Table 5: Classification of the *red elements*

Finally, let $d(n)$ denote the cumulative quantity (or an estimate thereof) of elements in $\mathcal{R}$ recovered after n samples. Let $h(d(n))$ be a non-decreasing gain function associated with this quantity, and let c denote the cost of collecting and analyzing an additional sample.

Since a unit increment in $d(n)$ may be negligible relative to the scale of the problem, we introduce a tolerance parameter $\delta \in [0, 1]$ representing the fraction of the total mass C that is considered practically relevant. The stopping time is then defined as

$$s = \inf \left\{ n \geq 0 : \left[h(d(n) + \delta C) - h(d(n)) \right] \cdot u(n) \leq c \right\}.$$

This formulation explicitly links the decision to continue sampling to three key components: (i) posterior beliefs about the total mass of elements of interest, as summarized by $u(n)$; (ii) the marginal gain induced by additional recovery, encoded by the function h; and (iii) the economic cost of further sampling. As such, it provides a principled Bayesian criterion for terminating the sampling process in microchip recycling applications.

8. Conclusions and Further Works

This thesis started investigating Bayesian models for chemical compositional data arising from PCBs residual waste under severe data-scarcity constraints. Two complementary representations of the same information were considered: ppm count vectors modeled through categorical likelihoods and simplex-valued compositions modeled directly in the compositional framework. Across all analyses, the focus was placed on posterior predictive distributions rather than point estimates, with the practical goal of supporting sequential sampling decisions in XRF-based characterization tasks.

8.1. Final recommendations

Within the compositional framework, the Logistic–Normal (LN) model emerges as the most reliable default choice under strong data scarcity. In particular, the LN model under the ILR transformation provides stable posterior predictive behavior when moving from 3 to 6 observations, with posterior summaries that remain comparatively centralized and less sensitive to oscillations across sequential updates. Although predictive intervals are still relatively wide, they represent a coherent quantification of uncertainty and are expected to shrink in a reasonable manner as additional samples are collected. In this setting, the LN model under the ALR

transformation may also be considered, as after six observations it yields similar point estimates and slightly narrower predictive intervals.

When a larger number of samples becomes available, the reparameterized Extended Flexible Dirichlet (EFD) model represents a promising alternative. Its component-specific inflation parameters allow for richer heterogeneity than the standard Dirichlet model and may enable a more detailed characterization of element-wise deviations from a common barycenter. With more data, this additional flexibility could translate into a more informative posterior structure, potentially inducing non-trivial variability patterns and improving the representation of element-specific behavior. In contrast, the reparameterized Flexible Dirichlet (FD) model provided limited practical gains in the present application. Across both prior sensitivity analyses and predictive summaries, FD either collapses toward Dirichlet-like behavior or becomes unstable when substantial posterior mass is forced onto inflation parameters. As a result, FD appears less suitable as a default model for the data regimes considered in this work.

Within the ppm count formulation, the Dirichlet–Multinomial (DM) model offers the most convenient trade-off between interpretability, computational efficiency, and stability of posterior inference. Flexible categorical extensions are substantially more computationally demanding and, under data scarcity, tend to provide limited effective gains in predictive performance. For this reason, the DM model is recommended in the current data regime. With a larger number of samples, the use of flexible categorical extensions may become worthwhile, with the final choice guided by practical considerations such as simulation time and posterior stability. Predictive criteria such as WAIC should only be compared within the same data representation, since categorical and compositional likelihoods are not directly comparable.

8.2. Future developments

Several directions may further extend and strengthen the present work. A primary development concerns repeating the comparative analysis with a larger number of XRF measurements. This would allow for a more definitive assessment of whether the inflation parameters of the EFD model exhibit stronger component-specific differentiation and whether flexible categorical models are able to move beyond the near-Dirichlet behavior observed under severe data scarcity. Another important direction concerns the refinement of dependence structures. For the Logistic–Normal (LN) model, richer covariance parameterizations may become feasible as additional data are collected, potentially allowing for a more explicit and reliable characterization of the covariance matrix. Similarly, under the EFD model, larger datasets could clarify whether heterogeneity in the τ_k parameters effectively translates into predictive dependence beyond that mechanically induced by the simplex constraint.

A preliminary analysis incorporating 21 additional samples was conducted, providing initial evidence consistent with our expectations regarding model behavior. In the multinomial framework, posterior credible intervals for $\boldsymbol{p}$ remain very narrow, and the flexible extensions continue to behave similarly to the Dirichlet–Multinomial model, showing limited practical departure from the baseline specification. Turning to the two most promising models in the compositional setting, namely the Logistic–Normal and the EFD models, different patterns emerge. The LN model exhibits remarkable stability: posterior predictive intervals for $\boldsymbol{\theta}$ remain of comparable magnitude to those obtained with only 3 or 6 samples, and the estimated covariance structure does not display substantial qualitative changes. This suggests that the LN formulation provides a stable representation of compositional variability, though additional data do not immediately translate into sharp predictive contraction.

In contrast, the EFD model shows a more pronounced response to the increased sample size. Greater heterogeneity in the parameters τ_k is observed, along with the appearance of multiple positive correlations between SiO_2 and other elements of high-abundance. This behavior is consistent with the intended flexibility of the model and indicates that, with more data, the EFD specification may begin to depart meaningfully from Dirichlet-like behavior. Despite these encouraging preliminary findings, a more systematic robustness analysis is required. The prior configurations adopted in the thesis for the 3- and 6-sample settings appear to perform reasonably well also with the enlarged dataset, with only minor exceptions. Nonetheless, a

comprehensive sensitivity analysis under larger sample sizes would be necessary before drawing definitive practical recommendations.

Finally, the optimal stopping framework introduced in this thesis can be substantially extended. Rather than relying on step-by-step predictive assessment, a more principled strategy would consist in minimizing an objective function that aggregates predictive quantities over the entire unsampled portion of the full batch. In this context, models exhibiting stable predictive updates as new observations are incorporated are preferable, since strongly oscillating predictions may lead to rapidly increasing uncertainty when forecasting over longer horizons. Integrating a stable compositional predictive model, such as the LN model with ILR transformation, under the current evidence, within a horizon-based stopping strategy represents a natural and promising direction for future research.

References

Acquafredda, Pasquale (2019). "XRF technique". *Physical Sciences Reviews* 4.8, p. 20180171.

Aitchison, John (1982). "The statistical analysis of compositional data". *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2, pp. 139–160.

Ankam, Divya and Nizar Bouguila (2018). "Compositional data analysis with PLS-DA and security applications". *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, pp. 338–345.

Ascari, Roberto, Agnese Maria Di Brisco, et al. (2024). "A multivariate mixture regression model for constrained responses". *Bayesian Analysis* 19.2, pp. 377–405.

Ascari, Roberto, Sonia Migliorati, and Andrea Ongaro (2019). "Bayesian Inference for a Mixture Model on the Simplex". *Statistical Learning of Complex Data*. Ed. by Francesca Greselin et al. Cham: Springer International Publishing, pp. 103–110. ISBN: 978-3-030-21140-0.

Atchison, Jhon and Sheng M Shen (1980). "Logistic-normal distributions: Some properties and uses". *Biometrika* 67.2, pp. 261–272.

Casella, George and Edward I George (1992). "Explaining the Gibbs sampler". *The American Statistician* 46.3, pp. 167–174.

Champ, Charles W and Andrew V Sills (2022). "The generalized law of total covariance". *arXiv preprint arXiv:2205.14525*.

Commission, European, Entrepreneurship Directorate-General for Internal Market Industry, and SMEs (2023). *Study on the critical raw materials for the EU 2023 – Final report*. Publications Office of the European Union. DOI: doi/10.2873/725585.

Di Brisco, A, Sonia Migliorati, et al. (2017). "A special Dirichlet mixture model for multivariate bounded responses". In: *Cladag 2017 Book of Short Papers*. Universitas Studiorum srl, pp. 1–6.

Egozcue, Juan José et al. (2003). "Isometric logratio transformations for compositional data analysis". *Mathematical geology* 35.3, pp. 279–300.

Gilks, Walter R. (1996). "Markov Chain Monte Carlo in Practice". In: ed. by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Chapman and Hall/CRC. Chap. Full conditional distributions, pp. 75–88.

Greenacre, Michael et al. (2023). "Aitchison's compositional data analysis 40 years on: A reappraisal". *Statistical Science* 38.3, pp. 386–410.

Grunsky, Eric, Michael Greenacre, and Bruce Kjarsgaard (2024). "GeoCoDA: Recognizing and validating structural processes in geochemical data. A workflow on compositional data analysis in lithogeochemistry". *Applied Computing and Geosciences* 22, p. 100149. ISSN: 2590-1974. DOI: https://doi.org/10.1016/j.acags.2023.100149.

Martín-Fernández, Josep A, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn (2003). "Dealing with zeros and missing values in compositional data sets using nonparametric imputation". *Mathematical Geology* 35.3, pp. 253–278.

Ongaro, Andrea and Sonia Migliorati (2012). "A Generalization of the Dirichlet Distribution". *Journal of Multivariate Analysis* 114, pp. 412–426. DOI: 10.1016/j.jmva.2012.08.003.

— (2020). "A New Mixture Model on the Simplex". *Statistical Methods & Applications* 29, pp. 123–148. DOI: 10.1007/s10260-019-00495-6.

Rasmussen, Shelley L and Norman Starr (1979). "Optimal and adaptive stopping in the search for new species". *Journal of the American Statistical Association* 74.367, pp. 661–667.

Tsagris, Michail T, Simon Preston, and Andrew TA Wood (2011). "A data-based power transformation for compositional data". *arXiv preprint arXiv:1106.1451*.

Watanabe, Sumio and Manfred Opper (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of machine learning research* 11.12.

A. Hierarchical Multinomial Covariance

1. The marginal variance of a single component $N_{i,k}$ is obtained by application of the Law of Total Variance,

$$\mathrm{Var}(N_{i,k}) = \mathbb{E}[\mathrm{Var}(N_{i,k} \mid \boldsymbol{p})] + \mathrm{Var}(\mathbb{E}[N_{i,k} \mid \boldsymbol{p}])\,. \tag{13}$$

Conditionally on $\boldsymbol{p}$, each component follows a binomial distribution,

$$N_{i,k} \mid \boldsymbol{p} \sim \mathrm{Binomial}(n, p_k),$$

with

$$\mathbb{E}(N_{i,k} \mid \boldsymbol{p}) = np_k, \qquad \mathrm{Var}(N_{i,k} \mid \boldsymbol{p}) = np_k(1-p_k).$$

Taking expectations with respect to $\boldsymbol{p}$ yields

$$\mathbb{E}[\mathrm{Var}(N_{i,k} \mid \boldsymbol{p})] = n\big(\mathbb{E}(p_k) - \mathbb{E}(p_k^2)\big),$$

and

$$\mathrm{Var}(\mathbb{E}[N_{i,k} \mid \boldsymbol{p}]) = n^2\mathrm{Var}(p_k).$$

Combining the two terms in (13) gives

$$\mathrm{Var}(N_{i,k}) = n\big(\mathbb{E}(p_k) - \mathbb{E}(p_k^2)\big) + n^2\mathrm{Var}(p_k).$$

Using the identity

$$\mathrm{Var}(p_k) = \mathbb{E}(p_k^2) - \mathbb{E}(p_k)^2,$$

the marginal variance can be expressed in the more interpretable form by adding and subtracting $\mathbb{E}(p_k)^2$ obtaining

$$\mathrm{Var}(N_{i,k}) = n\,\mathbb{E}(p_k)\big(1 - \mathbb{E}(p_k)\big) + n(n-1)\,\mathrm{Var}(p_k).$$

2. Let $N_{i,k}$ and $N_{i,r}$, with $k \neq r$, denote two distinct components of the multinomial count vector $\boldsymbol{N}_i$. The marginal covariance is obtained by application of the Law of Total , Champ and Sills (2022),

$$\mathrm{Cov}(N_{i,k}, N_{i,r}) = \mathbb{E}[\mathrm{Cov}(N_{i,k}, N_{i,r} \mid \boldsymbol{p})] + \mathrm{Cov}(\mathbb{E}[N_{i,k} \mid \boldsymbol{p}], \mathbb{E}[N_{i,r} \mid \boldsymbol{p}])\,. \tag{14}$$

Conditionally on $\boldsymbol{p}$, the multinomial distribution yields

$$\mathbb{E}(N_{i,k} \mid \boldsymbol{p}) = np_k, \qquad \mathbb{E}(N_{i,r} \mid \boldsymbol{p}) = np_r,$$

and

$$\mathrm{Cov}(N_{i,k}, N_{i,r} \mid \boldsymbol{p}) = -np_kp_r.$$

Taking expectations with respect to $\boldsymbol{p}$, the first term in (14) becomes

$$\mathbb{E}[\mathrm{Cov}(N_{i,k}, N_{i,r} \mid \boldsymbol{p})] = -n\,\mathbb{E}(p_kp_r).$$

The second term is given by

$$\mathrm{Cov}(\mathbb{E}[N_{i,k} \mid \boldsymbol{p}], \mathbb{E}[N_{i,r} \mid \boldsymbol{p}]) = \mathrm{Cov}(np_k, np_r) = n^2\mathrm{Cov}(p_k, p_r).$$

Combining the two contributions yields the marginal covariance

$$\mathrm{Cov}(N_{i,k}, N_{i,r}) = n^2 \mathrm{Cov}(p_k, p_r) - n\,\mathbb{E}(p_k p_r), \qquad k \neq r.$$

Using the identity

$$\mathbb{E}(p_k p_r) = \mathrm{Cov}(p_k, p_r) + \mathbb{E}(p_k)\mathbb{E}(p_r),$$

the expression can be equivalently written as

$$\mathrm{Cov}(N_{i,k}, N_{i,r}) = n(n-1)\mathrm{Cov}(p_k, p_r) - n\,\mathbb{E}(p_k)\mathbb{E}(p_r), \qquad k \neq r.$$

B. The Dirichlet distribution

The Dirichlet distribution represents the simplest and most widely used model for random compositions. Let

$$\boldsymbol{p} \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K), \qquad \alpha_k > 0.$$

The probability density function of the Dirichlet distribution is given by

$$f(\boldsymbol{p} \mid \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \qquad \boldsymbol{p} \in \mathcal{S}^{K-1},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_k > 0$, $\mathcal{S}^{K-1} = \left\{ \boldsymbol{p} \in \mathbb{R}_+^K : \sum_{k=1}^K p_k = 1 \right\}$ denotes the $(K-1)$-dimensional simplex, and

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\alpha_+)}$$

is the multivariate Beta function, with $\alpha_+ = \sum_{k=1}^K \alpha_k$.

The first two moments are

$$\mathbb{E}(p_k) = \frac{\alpha_k}{\alpha_+}, \qquad \mathrm{Var}(p_k) = \frac{\alpha_k(\alpha_+ - \alpha_k)}{\alpha_+^2(\alpha_+ + 1)},$$

with covariance, for $k \neq r$,

$$\mathrm{Cov}(p_k, p_r) = -\frac{\alpha_k \alpha_r}{\alpha_+^2(\alpha_+ + 1)}.$$

The Dirichlet distribution imposes strictly negative dependence between components, which may be overly restrictive in applications involving complex generative mechanisms.

C. Convergence Results

This appendix reports convergence diagnostics for all Bayesian models considered in the thesis. Given the high dimensionality of several parameter blocks (e.g., the full set of K concentration parameters α_k or the $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ parameters in the Logistic–Normal model), we do not include traceplots for every single component. Instead, we adopt the following reporting strategy:

- **Representative components:** we show traceplots for a small subset of parameters chosen to cover all revelant fitted models, see Figures 13; 14; 15.
- **Summary across all parameters:** global convergence is summarized via tables reporting the minimum bulk-ESS, minimum tail-ESS, maximum $\widehat{R}$, and the number of divergent transitions for each model, see Table 6.

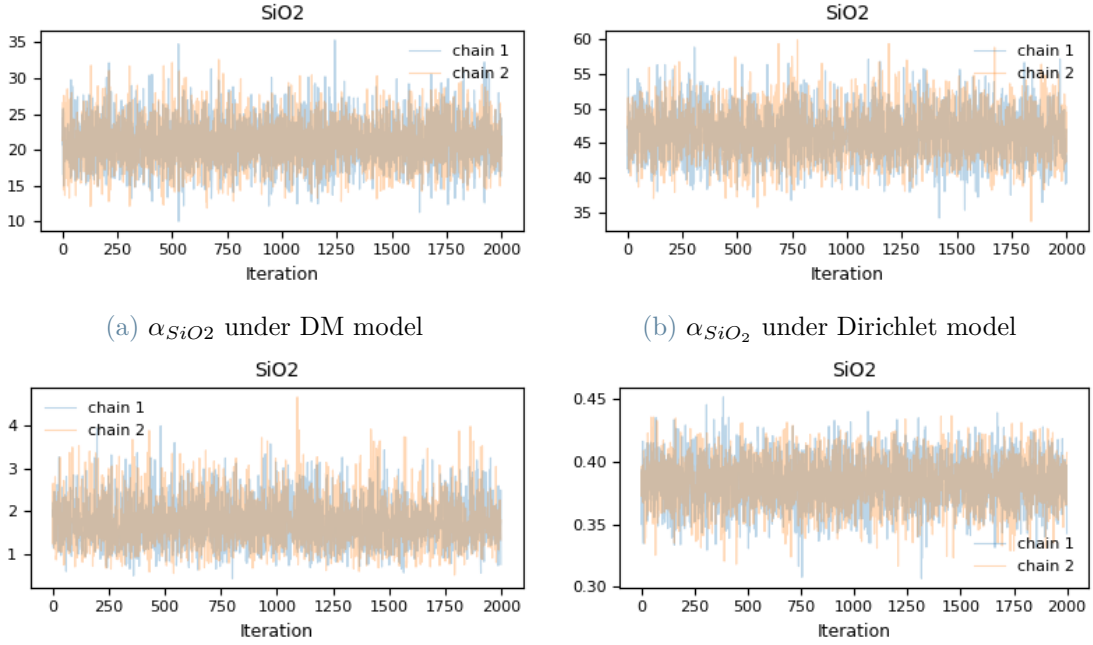
Overall, the traceplots reported below are consistent with satisfactory mixing for the selected parameters and with $\widehat{R} \approx 1$ and adequate ESS values, unless explicitly noted.

D. Additional Results, Categorical Setting

This appendix reports additional posterior results for the categorical setting, focusing on the Extended Flexible Dirichlet–Multinomial (EFDM) model. In the main text, posterior distributions under EFDM and FDM were shown to be essentially indistinguishable; therefore, only EFDM results are presented here to avoid redundancy. Figure 16 displays posterior distributions of selected components p_k after 3 and 6 observations, illustrating model behavior for a highly abundant, a weakly observed, and a never-observed component.

Table 6: Convergence summary for all fitted models. For each model we report the maximum $\widehat{R}$, the minimum bulk-ESS and tail-ESS across monitored parameters, and the number of divergent transitions.

Model	max $\widehat{R}$	min bulk-ESS	min tail-ESS	divergences
Dirichlet–Multinomial	1.004	206.6	239.2	0
FD–Multinomial	1.003	264.2	177.3	0
EFD–Multinomial	1.013	224.3	247.2	0
Dirichlet	1.004	1498.7	2266.5	0
FD (reparam.)	1.005	1395.9	1609.8	0
EFD (reparam.)	1.005	1403.5	1060.0	0
Logistic–Normal (ILR)	1.005	1160.0	1201.5	0
Logistic–Normal (ALR)	1.003	1099.5	1425.4	0



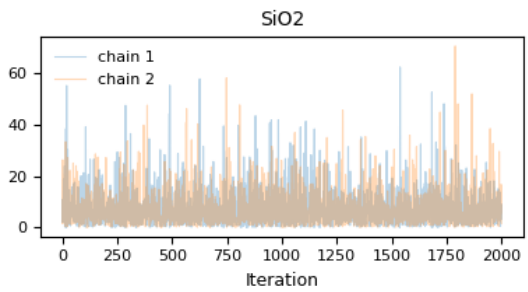
(a) α_{SiO2} under DM model

(b) α_{SiO_2} under Dirichlet model

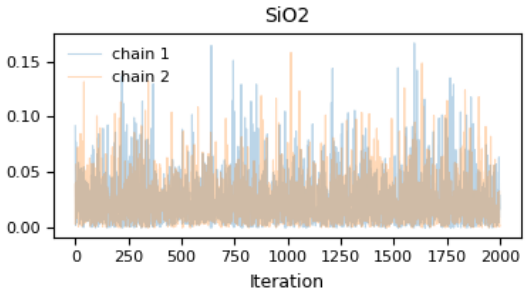
(c) α_{SiO_2} under FD model

(d) $\bar{\alpha}_{SiO_2}$ under EFD model

Figure 13: Traceplots for selected concentration parameters α_{SiO_2} under the Dirichlet-based models. The reported DM model, Dirichlet, FD and EFD model

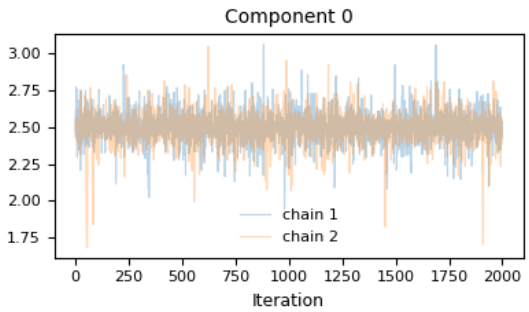


(a) Selected τ_k components

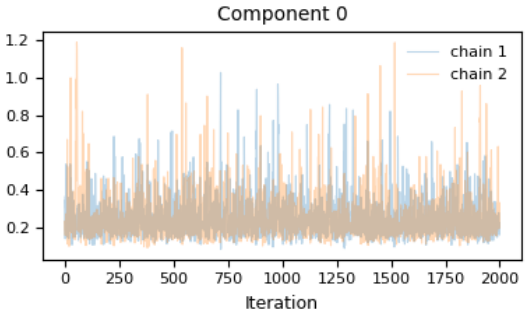


(b) Selected q_k components

Figure 14: Traceplots for inflation parameters τ_k and mixture weights $\boldsymbol{q}$ in the Extended Flexible Dirichlet (EFD) model (reparameterized).



(a) Selected components of $\boldsymbol{\mu}$



(b) Selected dispersion parameters

Figure 15: Traceplots for representative parameters in the Logistic–Normal model. We report selected components of $\boldsymbol{\mu}$ and dispersion parameters under ILR transformation.

E. Additional Results, Compositional Setting

This appendix reports a representative example of the unstable behavior discussed in the main text for models defined on compositional observations. Specifically, we consider the prior configuration with $a_w > b_w$, which leads to multimodal posterior behavior and unreliable inference. The two panels below (Figure 17 illustrate the distinct posterior regimes associated with different allocations of the perturbation mass.

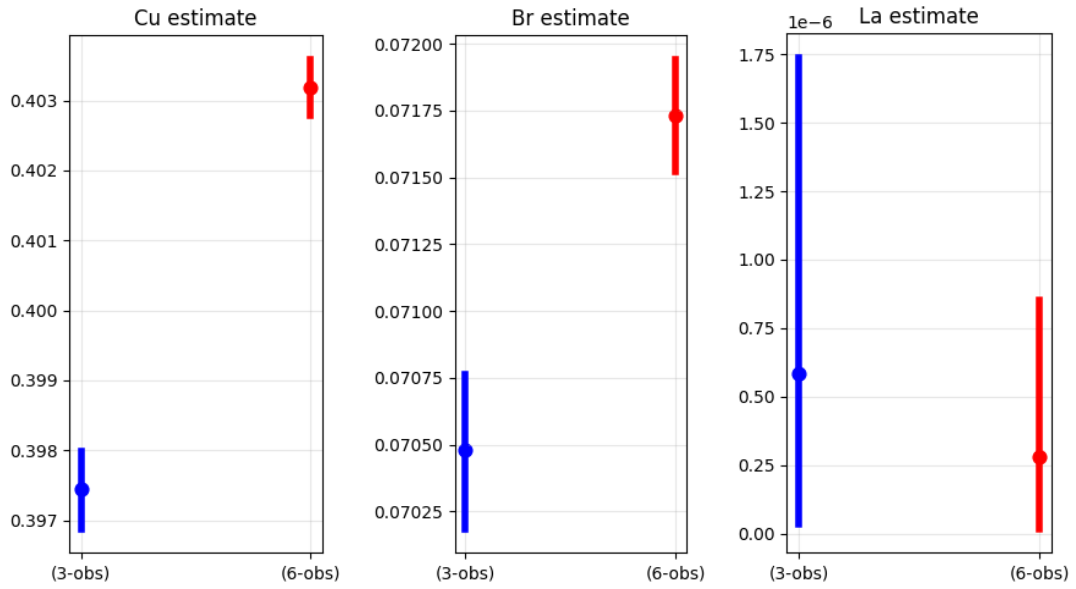


Figure 16: Posterior distributions of three representative components p_k after 3 and 6 observations under the Extended Flexible Dirichlet–Multinomial model: a highly abundant component (left), a weakly observed component (center), and a never-observed component (right).

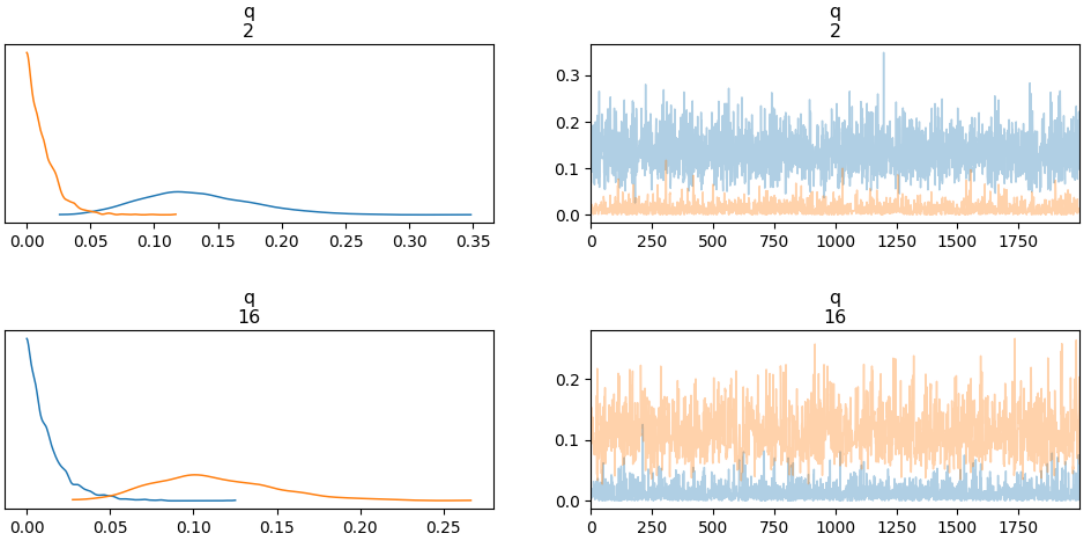


Figure 17: Posterior distributions obtained under a prior configuration with $a_w > b_w$, illustrating the multimodal behavior discussed in the main text. Figure shows results for $q_{Al_2O_3}$ (top) and q_{Cu} (bottom).