

A Bayesian Journey through Dirichlet Processes

BOCCONI STUDENTS ASSOCIATION FOR MATHEMATICAL SCIENCES

September 4, 2026

Contents

1	Introduction	4
1.1	Bayesian vs Frequentist Approach	4
1.2	Bayes' Theorem	4
1.3	From Bayes' Rule to Tractable Posteriors	5
1.4	Some Canonical Conjugate Models	6
1.4.1	Binomial Model	6
1.4.2	Poisson Model	8
1.4.3	Normal Model (known variance)	9
1.4.4	Normal Model (unknown variance)	10
1.5	Posterior Inference Beyond Conjugate Models	11
1.6	Gibbs Sampling and MCMC Methods	12
2	Fundamentals	15
2.1	Exchangeability	15
2.2	Representation theorem	16
2.3	Hierarchical Models	17
2.3.1	The Hierarchical Structure	17
2.3.2	Shrinkage	17
2.3.3	Parametric Foundations: Theoretical Justification of Bayesian Methods . .	18
2.4	Posterior Predictive Distribution	19
3	Conceptual bridge: from parametric to non-parametric	22
3.1	Finite Mixture Models	22
3.2	The Bottlenecks of Finite Mixtures	24
3.3	Posterior Inference via Gibbs Sampling	25
4	What if we let $K \rightarrow \infty$?	27
4.1	From Finite Mixtures to the Infinite Limit	27

4.2	The Chinese Restaurant Process	28
4.3	Motivating the Dirichlet Process through Bayesian Histograms	29
4.4	Measure-Theoretic Foundations	30
4.5	The Dirichlet Process	31
4.6	Stick-Breaking Construction	34
4.7	Dirichlet Process Mixtures	35
4.7.1	The Kernel Mixture Framework	35
4.7.2	Hierarchical Formulation and Predictive Inference	36
4.7.3	Posterior Computation and Marginal Gibbs Sampling	37
4.7.4	Blocked Gibbs Sampler via Stick-Breaking	38
4.7.5	Hyperprior for the Concentration Parameter	39
4.7.6	Non-Conjugate Models and Auxiliary Variable Methods	40
5	A practical application	42
5.1	Modeling Financial Returns: The Heavy-Tail Problem	42
5.2	Implications for Risk Management: VaR and Expected Shortfall	43
5.3	Data and Computational Setup	44
5.4	Empirical Results and Density Estimation	45
5.5	Risk Quantification Comparison	45
6	Conclusion	46
A	Appendix	48
A.1	Derivation of Marginal Dirichlet Process Moments	48
A.2	Algorithmic Steps of the Blocked Gibbs Sampler	48
A.3	Comprehensive Notation Taxonomy	49
A.4	Notation Convention: $P(\cdot)$ vs $p(\cdot)$	50

1 Introduction

1.1 Bayesian vs Frequentist Approach

Statistical inference can be broadly framed through two dominant paradigms: the frequentist and the Bayesian approach. While the frequentist perspective interprets probability as a long-run frequency and focuses on fixed but unknown parameters, it often relies on assumptions of repeated sampling and stable data-generating processes.

However, in many real-world applications, these assumptions are difficult to justify. Data are rarely generated under identical conditions, and decision-making typically occurs in environments characterized by limited information and continuous structural change.

In contrast, the Bayesian framework provides a fundamentally different perspective, interpreting probability as a degree of belief. This allows prior information to be formally incorporated and updated as new data become available, leading to a coherent and sequential learning process.

In this sense, Bayesian inference is not only a statistical tool but a natural framework for modeling how information is processed and beliefs are revised over time. This feature becomes particularly relevant in settings where uncertainty is pervasive and data alone are insufficient to fully characterize the underlying phenomena.

A natural illustration of this difference arises in financial markets, where agents continuously update their expectations in response to new information.

Consider, for instance, the problem of assessing the probability that a central bank will increase interest rates at an upcoming policy meeting. Before the release of key macroeconomic indicators, market participants have implicit beliefs about this event, often reflected in the prices of interest rate derivatives. When new information becomes available (e.g. an inflation report exceeding expectations) these beliefs are immediately revised, leading to observable changes in market-implied probabilities.

This process is inherently Bayesian: initial beliefs (priors) are updated through incoming data (likelihood), resulting in revised expectations (posteriors). In contrast, a purely frequentist approach would struggle to incorporate such sequential information updates in a coherent and interpretable way.

1.2 Bayes' Theorem

The Bayesian approach to statistical inference can be understood as a formal mechanism for learning about unknown quantities from observed data. Let $\theta \in \Theta$ denote an unknown parameter representing the characteristics of a population, and let $y \in \mathcal{Y}$ denote the observed dataset, viewed as a realization from a larger sample space.

Before observing the data, uncertainty about θ is described through a prior distribution $p(\theta)$, which encodes initial beliefs about the parameter. The relationship between the parameter and the data is modeled through the sampling distribution (or likelihood) $p(y | \theta)$, which specifies how the data are generated conditional on θ .

Once data are observed, Bayesian inference proceeds by updating prior beliefs to obtain the posterior distribution $p(\theta | y)$, representing the revised state of knowledge about the parameter. This update is governed by Bayes' Theorem:

$$p(\theta | y) = \frac{p(y | \theta) p(\theta)}{\int_{\Theta} p(y | \tilde{\theta}) p(\tilde{\theta}) d\tilde{\theta}} \quad (1)$$

Bayes' Theorem does not prescribe what prior beliefs should be, but rather provides a coherent rule for how such beliefs must be updated in light of new information. This sequential updating mechanism lies at the core of Bayesian learning and forms the foundation for the models developed in the following sections.

1.3 From Bayes' Rule to Tractable Posteriors

Bayes' Theorem provides a formal rule for updating beliefs, but its direct application quickly raises practical questions. In particular, the posterior distribution can be rewritten as

$$p(\theta | y) \propto p(y | \theta) p(\theta).$$

The denominator in the original expression,

$$p(y) = \int_{\Theta} p(y | \theta) p(\theta) d\theta,$$

plays the role of a normalizing constant, ensuring that the posterior distribution integrates to one.

The first expression highlights an important aspect: the posterior is obtained by combining the prior and the likelihood. However, even when both components are simple, their product does not necessarily lead to a distribution with a known or tractable form.

As a result, deriving the posterior distribution explicitly can be difficult, and in many cases impossible, without resorting to numerical methods.

A particularly important class of models avoids this difficulty. A prior distribution is said to be conjugate to a given likelihood if the posterior distribution belongs to the same family as the prior. We formalize the notion of conjugacy as follows:

Definition (Conjugate) A class $\mathcal{P}$ of prior distributions for θ is called *conjugate* for a sampling model $p(y \mid \theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta \mid y) \in \mathcal{P}.$$

In this case, the updating process becomes analytically tractable, and the posterior can be obtained in closed form.

The relevance of conjugacy is therefore twofold. On the one hand, it provides computational simplicity, allowing for explicit posterior expressions. On the other hand, it offers a clear interpretation of Bayesian updating, where posterior parameters can often be seen as a combination of prior information and data-driven evidence.

In the following sections, we will explore three conjugate families in detail, using them as building blocks for more complex Bayesian models.

1.4 Some Canonical Conjugate Models

To make the concept of conjugacy concrete, we now examine a set of classical models where posterior distributions can be derived analytically. These examples play a central role in Bayesian statistics, as they provide both intuition and practical tools for inference.

In particular, we consider the Binomial, Poisson, and Normal models, each paired with its corresponding conjugate prior.

Through these examples, we will see how conjugacy allows for simple and interpretable posterior updates, where prior information and observed data combine in a transparent way.

1.4.1 Binomial Model

Let $Y_1, \dots, Y_n \mid \theta$ be i.i.d. Bernoulli random variables with parameter $\theta \in [0, 1]$. The likelihood function is given by

$$p(y_1, \dots, y_n \mid \theta) = \theta^{\sum_{i=1}^n y_i} (1 - \theta)^{n - \sum_{i=1}^n y_i}.$$

Equivalently, defining the sufficient statistic

$$S = \sum_{i=1}^n Y_i,$$

we have

$$S \mid \theta \sim \text{Binomial}(n, \theta),$$

with likelihood

$$p(S = s \mid \theta) \propto \theta^s (1 - \theta)^{n-s}.$$

Conjugate prior A natural conjugate prior for θ is the Beta distribution:

$$\theta \sim \text{Beta}(a, b),$$

with density

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in [0, 1].$$

In the special case $a = 1$ and $b = 1$, the Beta distribution reduces to the uniform distribution on $[0, 1]$.

Posterior distribution Combining prior and likelihood, we obtain

$$p(\theta \mid y_1, \dots, y_n) \propto \theta^{s+a-1}(1-\theta)^{n-s+b-1}.$$

Thus, the posterior distribution is

$$\theta \mid y_1, \dots, y_n \sim \text{Beta}(a + s, b + n - s).$$

Normalizing constant The normalizing constant can be obtained explicitly in this case. Indeed, using the identity

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

we have

$$p(y_1, \dots, y_n) = \int_0^1 p(y_1, \dots, y_n \mid \theta) p(\theta) d\theta = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+s)\Gamma(b+n-s)}{\Gamma(a+b+n)}.$$

Thus, the posterior density can be written in normalized form as

$$p(\theta \mid y_1, \dots, y_n) = \frac{\Gamma(a+b+n)}{\Gamma(a+s)\Gamma(b+n-s)} \theta^{a+s-1}(1-\theta)^{b+n-s-1}.$$

Interpretation The posterior parameters can be interpreted as an update of the prior parameters through the data:

$$a \rightarrow a + s, \quad b \rightarrow b + (n - s).$$

This illustrates how conjugacy leads to simple and interpretable Bayesian updating.

1.4.2 Poisson Model

Let $Y_1, \dots, Y_n \mid \theta$ be i.i.d. Poisson random variables with parameter $\theta > 0$. The likelihood function is

$$p(y_1, \dots, y_n \mid \theta) = \prod_{i=1}^n \frac{\theta^{y_i} e^{-\theta}}{y_i!} \propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta}.$$

Defining the sufficient statistic

$$S = \sum_{i=1}^n Y_i,$$

we have

$$S \mid \theta \sim \text{Poisson}(n\theta),$$

and the likelihood can be written as

$$p(S = s \mid \theta) \propto \theta^s e^{-n\theta}.$$

Conjugate prior A conjugate prior for θ is given by the Gamma distribution:

$$\theta \sim \text{Gamma}(a, b),$$

with density

$$p(\theta) \propto \theta^{a-1} e^{-b\theta}, \quad \theta > 0, \quad a, b > 0.$$

Posterior distribution Combining prior and likelihood, we obtain

$$p(\theta \mid y_1, \dots, y_n) \propto \theta^{a+s-1} e^{-(b+n)\theta}.$$

Thus, the posterior distribution is

$$\theta \mid y_1, \dots, y_n \sim \text{Gamma}(a + s, b + n).$$

Normalizing constant The normalizing constant can also be computed explicitly. Using the identity

$$\int_0^\infty \theta^{a-1} e^{-b\theta} d\theta = \frac{\Gamma(a)}{b^a},$$

we obtain

$$p(y_1, \dots, y_n) = \frac{b^a}{\Gamma(a)} \cdot \frac{\Gamma(a + s)}{(b + n)^{a+s}}.$$

Thus, the posterior density can be written in normalized form as

$$p(\theta \mid y_1, \dots, y_n) = \frac{(b + n)^{a+s}}{\Gamma(a + s)} \theta^{a+s-1} e^{-(b+n)\theta}.$$

Interpretation The posterior parameters are updated as

$$a \rightarrow a + s, \quad b \rightarrow b + n,$$

showing again how conjugacy leads to simple and interpretable Bayesian updating.

These results highlight that conjugate priors arise naturally within the exponential family framework. Posterior distributions retain the same functional form, with parameters updated through sufficient statistics of the data. This unifying structure underlies many classical Bayesian models and will extend naturally to more complex settings considered in the following sections.

1.4.3 Normal Model (known variance)

Let $Y_1, \dots, Y_n \mid \theta$ be i.i.d. normal random variables:

$$Y_i \mid \theta \sim \mathcal{N}(\theta, \sigma^2), \quad \sigma^2 \text{ known.}$$

The likelihood function is

$$p(y_1, \dots, y_n \mid \theta) \propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 \right\}.$$

Expanding the quadratic term, the likelihood depends on the data through the sufficient statistic

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Conjugate prior A conjugate prior for θ is given by a normal distribution:

$$\theta \sim \mathcal{N}(\mu_0, \tau_0^2).$$

Posterior distribution Combining prior and likelihood, the posterior distribution is

$$\theta \mid y_1, \dots, y_n \sim \mathcal{N}(\mu_n, \tau_n^2),$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{n}{\sigma^2} \bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \right)^{-1}.$$

Interpretation The posterior mean is a weighted average of the prior mean and the sample mean, with weights proportional to their respective precisions. The posterior precision is the

sum of prior and data precisions:

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

1.4.4 Normal Model (unknown variance)

In this setting, the parameter of interest is two-dimensional, (θ, σ^2) . A convenient way to specify a prior distribution exploits the fact that:

$$p(\theta, \sigma^2) = p(\theta \mid \sigma^2) p(\sigma^2).$$

This decomposition allows us to construct conjugate priors by choosing compatible distributions for each component, while preserving analytical tractability of the posterior.

Consider the model

$$Y_i \mid \theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2), \quad i = 1, \dots, n.$$

The likelihood function is

$$p(y_1, \dots, y_n \mid \theta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 \right\}.$$

Conjugate prior A conjugate prior is obtained by specifying a hierarchical structure:

$$\theta \mid \sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2 / \kappa_0), \quad \sigma^2 \sim \text{Inverse-Gamma} \left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2} \right),$$

with $\kappa_0 > 0$ and $\nu_0 > 0$.

The dependence of the prior variance of θ on σ^2 is essential for conjugacy. In particular, assuming that the prior variance is proportional to σ^2 ensures that the joint prior belongs to a conjugate family and leads to tractable posterior updates.

Posterior distribution The posterior distributions are given by

$$\theta \mid \sigma^2, y \sim \mathcal{N}(\mu_n, \sigma^2 / \kappa_n), \quad \sigma^2 \mid y \sim \text{Inverse-Gamma} \left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2} \right),$$

where

$$\begin{aligned} \kappa_n &= \kappa_0 + n, & \mu_n &= \frac{\kappa_0 \mu_0 + n \bar{y}}{\kappa_n}, \\ \nu_n &= \nu_0 + n, \\ \sigma_n^2 &= \frac{1}{\nu_n} \left[\nu_0 \sigma_0^2 + \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{\kappa_0 n}{\kappa_n} (\bar{y} - \mu_0)^2 \right]. \end{aligned}$$

Interpretation The posterior updates combine information from the data and the prior through sufficient statistics. In particular, $\bar{y}$ and the sample variance summarize the data, while κ_0 , ν_0 , and σ_0^2 control the strength and scale of prior information.

1.5 Posterior Inference Beyond Conjugate Models

Consider the normal sampling model

$$Y_1, \dots, Y_n \mid \theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2),$$

and a prior distribution specified independently as

$$\theta \sim \mathcal{N}(\mu_0, \tau_0^2), \quad \sigma^2 \sim \text{Inverse-Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

and the joint distribution can be defined as a "semiconjugate" prior distribution. In contrast to the conjugate case, where the prior variance of θ is proportional to σ^2 , the independence assumption implies that the joint posterior distribution does not belong to a standard family. In particular, the marginal posterior of σ^2 is not of a known form, making direct sampling infeasible.

A possible approach is to approximate the posterior distribution on a discrete grid of parameter values. However, this method quickly becomes computationally impractical as the dimension of the parameter space increases.

Despite the intractability of the joint posterior, the conditional distributions retain simple forms. In particular, the conditional distribution of θ given σ^2 and the data is normal, while the conditional distribution of σ^2 given θ and the data is inverse-gamma.

Consider the joint posterior distribution $p(\theta, \sigma^2 \mid y)$, which is not available in a tractable form. However, the full conditional distributions are available:

$$p(\theta \mid \sigma^2, y), \quad p(\sigma^2 \mid \theta, y).$$

Suppose, for a moment, that we are able to sample a value $\sigma^{2(1)}$ from its marginal posterior distribution $p(\sigma^2 \mid y)$. Then, we could generate

$$\theta^{(1)} \sim p(\theta \mid \sigma^{2(1)}, y),$$

and the pair $(\theta^{(1)}, \sigma^{2(1)})$ would be a sample from the joint posterior distribution.

Although sampling directly from $p(\sigma^2 \mid y)$ is not feasible, we can instead proceed iteratively. Starting from an initial value $\theta^{(1)}$, we generate

$$\sigma^{2(2)} \sim p(\sigma^2 \mid \theta^{(1)}, y),$$

and then

$$\theta^{(2)} \sim p(\theta \mid \sigma^{2(2)}, y).$$

Repeating this procedure, we obtain a sequence

$$(\theta^{(1)}, \sigma^{2(1)}), (\theta^{(2)}, \sigma^{2(2)}), \dots$$

where each update uses only conditional distributions.

This construction shows that it is possible to generate samples from the joint posterior distribution by iteratively sampling from the full conditional distributions, even when the joint distribution itself is not directly accessible, suggesting an alternative strategy, which forms the basis of the Gibbs sampler.

1.6 Gibbs Sampling and MCMC Methods

The iterative procedure described in the previous section can be formalized within the general framework of Markov Chain Monte Carlo (MCMC) methods. The objective of MCMC is to generate a sequence of dependent random variables whose stationary distribution coincides with a target posterior distribution. In the present setting, the target is the joint posterior

$$p(\theta, \sigma^2 \mid y),$$

which is not available in a tractable closed form under the semiconjugate prior specification.

The key insight is that direct simulation from the joint posterior is not always necessary. Even when the joint density is analytically intractable, it may still be possible to sample from the corresponding conditional distributions. In our case, the relevant conditional distributions are

$$p(\theta \mid \sigma^2, y), \quad p(\sigma^2 \mid \theta, y),$$

which are called the *full conditional distributions*, since each one gives the distribution of a parameter conditional on all remaining unknown quantities and on the observed data.

The Gibbs sampler exploits this conditional structure. Rather than attempting to sample directly from the joint posterior, it constructs a Markov chain by successively updating each parameter from its full conditional distribution. Starting from an initial value $(\theta^{(0)}, \sigma^{2(0)})$, the algorithm generates a sequence

$$\{(\theta^{(s)}, \sigma^{2(s)})\}_{s \geq 1}$$

through the recursive scheme

$$\begin{aligned} \theta^{(s+1)} &\sim p(\theta \mid \sigma^{2(s)}, y), \\ \sigma^{2(s+1)} &\sim p(\sigma^2 \mid \theta^{(s+1)}, y). \end{aligned}$$

This construction defines a Markov chain on the parameter space, since the distribution of the next state depends only on the current state. Under standard regularity conditions—such as irreducibility, aperiodicity, and existence of the invariant distribution—the resulting chain admits

$$p(\theta, \sigma^2 \mid y)$$

as its stationary distribution. Therefore, as the number of iterations increases, the distribution of the simulated draws approaches the posterior distribution of interest.

An important feature of the Gibbs sampler is that it replaces a difficult joint simulation problem with a sequence of lower-dimensional conditional simulations. This is often computationally advantageous, especially in models where the full conditionals belong to standard parametric families. In such cases, posterior simulation can be implemented even when the joint posterior is too complicated to normalize, visualize, or integrate analytically.

The logic of the method extends naturally beyond the two-parameter case. Suppose that the parameter vector is

$$\theta = (\theta_1, \dots, \theta_m).$$

The Gibbs sampler then proceeds by iteratively sampling

$$\theta_1^{(s+1)} \sim p(\theta_1 \mid \theta_2^{(s)}, \dots, \theta_m^{(s)}, y),$$

$$\theta_2^{(s+1)} \sim p(\theta_2 \mid \theta_1^{(s+1)}, \theta_3^{(s)}, \dots, \theta_m^{(s)}, y),$$

and so on, until

$$\theta_m^{(s+1)} \sim p(\theta_m \mid \theta_1^{(s+1)}, \dots, \theta_{m-1}^{(s+1)}, y).$$

The resulting chain cycles through the coordinates of the parameter vector, updating each block conditionally on the most recent values of the others. In this sense, Gibbs sampling can be viewed as a systematic coordinate-wise exploration of the posterior distribution.

Once a sufficiently long sequence has been generated, posterior quantities can be approximated by Monte Carlo averages. For any integrable function $g(\theta, \sigma^2)$,

$$\mathbb{E}[g(\theta, \sigma^2) \mid y] \approx \frac{1}{S} \sum_{s=1}^S g(\theta^{(s)}, \sigma^{2(s)}).$$

This approximation is justified by the ergodic theorem for Markov chains: under suitable conditions, empirical averages along the simulated path converge to posterior expectations. As a consequence, Gibbs output can be used to approximate posterior means, variances, quantiles, credible regions, and posterior probabilities of events of interest.

It is worth emphasizing that the draws produced by the Gibbs sampler are generally not independent. Unlike standard Monte Carlo simulation, MCMC output exhibits serial dependence, since each iteration is generated conditionally on the previous one. This dependence does not invalidate the approximation, but it affects the efficiency of the algorithm: highly correlated

draws contain less information than an equally large sample of independent observations. For this reason, practical implementation of MCMC methods often involves issues such as initialization, burn-in, mixing behavior, and convergence assessment.

From a methodological perspective, the Gibbs sampler is particularly important because it shows that posterior simulation may remain feasible even when exact posterior analysis fails. In this way, Bayesian inference is no longer restricted to conjugate models or low-dimensional settings. Instead, the availability of tractable full conditional distributions is sufficient to construct an algorithm whose long-run behavior reproduces the posterior law.

2 Fundamentals

2.1 Exchangeability

The **exchangeability** assumption formalizes the idea that, before inspecting the data in detail, the ordering of observations does not add structural information to the model.

Definition 1 (Exchangeability). Let $p(y_1, \dots, y_n)$ be the joint density of $Y_1, \dots, Y_n$. If

$$p(y_1, \dots, y_n) = p(y_{\pi(1)}, \dots, y_{\pi(n)})$$

for every permutation π of $\{1, \dots, n\}$, then $Y_1, \dots, Y_n$ are exchangeable.

Roughly speaking, $Y_1, \dots, Y_n$ are exchangeable when index labels carry no information about outcomes: reordering the data does not change the probabilistic structure of the sample.

An example aligned with the interest-rate discussion above is the daily variation in market-implied probabilities of a central bank rate hike, extracted from asset prices. Over a short window with a stable macro-financial regime, these observations can be treated as exchangeable: swapping two days does not change the joint distribution used to infer the average uncertainty level in that period. The relevant information is in the sample as a whole, not in the specific time label of each day.

In other words, the joint distribution is invariant under reordering.

This does *not* imply independence: observations may still be dependent, but in a way that is *symmetric* with respect to indices.

This assumption is crucial because it leads directly to de Finetti's theorem: if a sequence is infinitely exchangeable, then it can be represented as conditionally i.i.d. given a latent variable. Therefore, using a prior is not only an operational modeling choice, but also a theoretical consequence of the probabilistic structure.

Proof. Suppose that, conditional on some unknown parameter θ , the variables $Y_1, \dots, Y_n$ are i.i.d. Then, for any realization $(y_1, \dots, y_n)$,

$$p(y_1, \dots, y_n \mid \theta) = \prod_{i=1}^n p(y_i \mid \theta).$$

Marginalizing with respect to the prior distribution $p(\theta)$ gives

$$p(y_1, \dots, y_n) = \int p(y_1, \dots, y_n \mid \theta) p(\theta) d\theta = \int \left(\prod_{i=1}^n p(y_i \mid \theta) \right) p(\theta) d\theta.$$

Now let π be any permutation of $\{1, \dots, n\}$. Since the product does not depend on the order of

its terms,

$$\prod_{i=1}^n p(y_i | \theta) = \prod_{i=1}^n p(y_{\pi(i)} | \theta).$$

Therefore,

$$p(y_1, \dots, y_n) = \int \left(\prod_{i=1}^n p(y_{\pi(i)} | \theta) \right) p(\theta) d\theta = p(y_{\pi(1)}, \dots, y_{\pi(n)}).$$

Hence the joint distribution is invariant under permutation, so $Y_1, \dots, Y_n$ are exchangeable. $\square$

2.2 Representation theorem

We introduced the likelihood–prior pair as the backbone of the Bayesian framework. We now justify this structure from the exchangeability assumption.

Theorem 1 (de Finetti Representation Theorem). *Let $Y_1, Y_2, \dots$ be an infinitely exchangeable sequence of random variables. Then there exists a latent variable θ such that, conditional on θ , the observations are independent and identically distributed.*

That is,

$$Y_1, Y_2, \dots | \theta \stackrel{iid}{\sim} p(y | \theta).$$

Consequently, for every finite sample,

$$p(y_1, \dots, y_n) = \int \left(\prod_{i=1}^n p(y_i | \theta) \right) p(\theta) d\theta.$$

In other words, every infinitely exchangeable sequence can be represented as a mixture of i.i.d. models.

The distribution $p(\theta)$ represents our beliefs about the outcomes of $\{Y_1, Y_2, \dots\}$ induced by the exchangeable belief model.

This result provides the mathematical "license" to use a prior distribution $p(\theta)$. It shows that the parameter θ is not just a convenient tool, but a necessary consequence of the symmetry (exchangeability) we assume in the data.

We can then summarize the results from this and the previous section as follows:

$$\left. \begin{array}{l} Y_1, \dots, Y_n | \theta \text{ are i.i.d.} \\ \theta \sim p(\theta) \end{array} \right\} \implies Y_1, \dots, Y_n \text{ are exchangeable.}$$

When is the condition “ $Y_1, \dots, Y_n$ are exchangeable for all n ” reasonable? For this condition to hold, we must have exchangeability and repeatability. Situations in which repeatability is reasonable include the important case in which $Y_1, \dots, Y_n$ sampled from a finite population of

size $N \gg n$

2.3 Hierarchical Models

In many real-world applications, data possess a "multi-level" or nested structure where observations are naturally grouped. Common examples include house prices within different neighborhoods or stock returns within different industrial sectors.

In such settings, the assumption that all observations across all groups are identically distributed is often too restrictive, if not entirely unrealistic, as it ignores group-specific features (the "pooled" approach). On the other hand, treating each group as completely independent entity is statistically inefficient, as it ignores global trends and shared characteristics (the "unpooled" approach). Hierarchical models allow us to relax these assumptions by assuming that while groups are distinct, their parameters are themselves *exchangeable*.

2.3.1 The Hierarchical Structure

While de Finetti's theorem justifies a single-level prior for a sequence of exchangeable observations, hierarchical modeling extends this logic to the groups themselves. If we have j groups, we assume that the group-specific parameters $\theta_1, \dots, \theta_j$ are exchangeable.

Following de Finetti's Representation Theorem, if exchangeability holds for the group parameters, there must exist a common population distribution (governed by hyperparameters ϕ) from which these parameters are drawn. This lead to a three-level hierarchical framework:

1. **Likelihood (Data Level):** Within each group j , the observations $y_{i,j}$ are conditionally i.i.d. given the group parameter θ_j :

$$y_{i,j} \mid \theta_j \sim p(y \mid \theta_j)$$

2. **Prior (Group Level):** The parameters $\{\theta_1, \dots, \theta_j\}$ are exchangeable and drawn from a shared distribution:

$$\theta_j \mid \phi \sim p(\theta \mid \phi)$$

3. **Hyperprior (Population Level):** The uncertainty regarding the global structure is captured by the distribution of the hyperparameters:

$$\phi \sim p(\phi)$$

2.3.2 Shrinkage

One of the key consequences of hierarchical modeling is what is called **shrinkage**: group-level estimates are pulled toward the common population mean. Intuitively, we can think of each

group estimate as a compromise between its local data signal and the global information shared across groups.

In a simple normal-normal hierarchy, if

$$\bar{y}_j | \theta_j \sim \mathcal{N}\left(\theta_j, \frac{\sigma^2}{n_j}\right), \quad \theta_j \sim \mathcal{N}(\mu, \tau^2),$$

then the posterior mean of θ_j can be written as

$$\mathbb{E}[\theta_j | \bar{y}_j] = w_j \bar{y}_j + (1 - w_j) \mu, \quad \text{where } w_j = \frac{n_j \tau^2}{n_j \tau^2 + \sigma^2}.$$

As a result the expected value is pulled from $\bar{y}_j$ toward μ by an amount depending on group sample size. Therefore, shrinkage is stronger when n_j is small (little information in group j) or when idiosyncratic noise is large (large σ^2). In those cases, the posterior estimate puts more weight on the common mean μ .

Financial Intuition: Shrinkage as Risk Regularization The practical power of hierarchical modeling lies in its ability to perform *partial pooling*. Consider the problem of estimating expected monthly returns across multiple assets within the same industrial sector.

In a non-hierarchical (unpooled) analysis, an asset that is thinly traded (so having only a few observations) might show a sample return of 50%. A naive investor might view this as a gold mine! However, a hierarchical model recognizes that such an extreme estimate, derived from scarce and noisy data, is likely a statistical fluke. Consequently, the estimate is strongly **shrunk** toward the sector-level mean (e.g., from 50% toward a more plausible range of 8%–12%, depending on the posterior uncertainty of the sector).

This behavior is highly desirable for risk management: it regularizes extreme estimates caused by small sample sizes while allowing assets with rich, stable histories (such as the S&P 500) to remain close to their observed means. In essence, the hierarchical model automatically discounts "noise" in favor of the "signal" provided by the broader population.

2.3.3 Parametric Foundations: Theoretical Justification of Bayesian Methods

The previous hierarchical construction provides the theoretical basis for Bayesian regularization: estimates are partially pooled and therefore stabilized when group-level information is weak. In this sense, the parametric framework clarifies why shrinkage improves robustness in finite samples, especially in noisy financial contexts with heterogeneous data availability.

Having established this mechanism at the parameter level, we now move to prediction, where uncertainty in both parameters and future observations is propagated through the posterior predictive distribution.

2.4 Posterior Predictive Distribution

An important feature of Bayesian inference is the ability to generate predictions for future or unobserved data. While the posterior distribution $p(\theta | y)$ summarizes uncertainty about the unknown parameter after observing the data, in many applications the primary goal is to make inference about future observations.

Let $\tilde{y}$ denote a new observation generated from the same data-generating process as the observed sample $y = (y_1, \dots, y_n)$. Conditional on the parameter θ , uncertainty about $\tilde{y}$ is described by the sampling model $p(\tilde{y} | \theta)$. However, since θ is unknown, prediction cannot rely on a fixed parameter value.

The Bayesian solution is to average over the posterior distribution of θ , leading to the posterior predictive distribution:

$$p(\tilde{y} | y) = \int_{\Theta} p(\tilde{y} | \theta) p(\theta | y) d\theta. \quad (2)$$

This expression shows that the predictive distribution is obtained by integrating the sampling model with respect to the posterior distribution of the parameter. Equivalently, the predictive distribution can be interpreted as the posterior expectation of $p(\tilde{y} | \theta)$.

Two important properties follow from this construction. First, the predictive distribution does not depend on unknown parameters, since θ has been integrated out. This makes it directly usable for prediction. Second, the predictive distribution depends explicitly on the observed data through the posterior $p(\theta | y)$, implying that predictions are updated as new information becomes available.

A key implication is that $\tilde{y}$ is not independent of the observed data y . Observing y provides information about θ , which in turn affects the distribution of future observations. In this sense, Bayesian prediction naturally captures the learning process: past data influence beliefs about the parameter, and these updated beliefs determine future predictions.

In simple conjugate models, the posterior predictive distribution often admits a closed form. For example, in the binomial model with a Beta prior, the predictive probability of a future success can be written as

$$\mathbb{P}(\tilde{Y} = 1 | y) = \mathbb{E}[\theta | y], \quad (3)$$

showing that prediction is directly linked to posterior summaries of the parameter.

More generally, the posterior predictive distribution incorporates two sources of uncertainty: the inherent variability of the data-generating process and the uncertainty about the parameter θ . As a result, predictive distributions are typically more dispersed than plug-in predictions based on point estimates, especially when the sample size is small.

This feature is particularly important in applications where uncertainty quantification is essential. In financial contexts, for instance, predicting future returns, volatility, or risk measures requires accounting not only for randomness in the data but also for uncertainty in model pa-

rameters. The posterior predictive distribution provides a coherent probabilistic framework for such tasks, allowing for the construction of predictive intervals and the evaluation of alternative future scenarios.

Finally, posterior predictive distributions also play a central role in model assessment. If the model is well specified, data generated from the predictive distribution should resemble the observed sample in relevant aspects. This idea forms the basis of posterior predictive checks, where discrepancies between observed and predicted data can indicate model misspecification. To make this construction concrete, consider the classical Beta-Binomial model. Let

$$Y_1, \dots, Y_n \mid \theta \sim \text{i.i.d. Bernoulli}(\theta),$$

where $\theta \in [0, 1]$ represents the probability of success. Suppose the prior distribution is

$$\theta \sim \text{Beta}(a, b).$$

If we denote by

$$s = \sum_{i=1}^n y_i$$

the number of observed successes, then by conjugacy the posterior distribution is

$$\theta \mid y \sim \text{Beta}(a + s, b + n - s). \quad (4)$$

Now let $\tilde{Y} \in \{0, 1\}$ be a future Bernoulli observation from the same population. The posterior predictive probability of success is

$$\mathbb{P}(\tilde{Y} = 1 \mid y) = \int_0^1 \mathbb{P}(\tilde{Y} = 1 \mid \theta) p(\theta \mid y) d\theta = \int_0^1 \theta p(\theta \mid y) d\theta = \mathbb{E}[\theta \mid y]. \quad (5)$$

Since the posterior distribution is $\text{Beta}(a + s, b + n - s)$, we obtain

$$\mathbb{P}(\tilde{Y} = 1 \mid y) = \frac{a + s}{a + b + n}. \quad (6)$$

Similarly,

$$\mathbb{P}(\tilde{Y} = 0 \mid y) = \frac{b + n - s}{a + b + n}. \quad (7)$$

This example makes the logic of posterior prediction particularly transparent. The predictive probability of a future success is not obtained by substituting a single fixed estimate of θ , but by averaging over the entire posterior distribution of the parameter. In this case, the result coincides with the posterior mean of θ .

A simple special case is obtained under the uniform prior, corresponding to $a = b = 1$. Then

$$\mathbb{P}(\tilde{Y} = 1 \mid y) = \frac{1 + s}{n + 2}. \quad (8)$$

For example, if $n = 10$ and $s = 2$, then

$$\mathbb{P}(\tilde{Y} = 1 \mid y) = \frac{3}{12} = 0.25.$$

Notice that this predictive probability differs from the sample proportion $s/n = 0.2$. The difference reflects the contribution of prior information and illustrates a fundamental Bayesian feature: prediction is regularized by posterior uncertainty, especially in small samples.

More generally, this example shows that the posterior predictive distribution naturally combines the information in the observed data with prior beliefs, while producing a prediction that no longer depends on unknown parameters. For this reason, posterior predictive inference provides a coherent bridge between parameter estimation and forecastin.

3 Conceptual bridge: from parametric to non-parametric

3.1 Finite Mixture Models

What happens when the observed data contain multiple subgroups that are not labeled in the dataset?

A classical example is height data from a mixed population: we may want to model male and female height distributions separately, but the sex label is not observed for each individual $i = 1, \dots, N$. Instead of forcing a single complex multimodal model, a cleaner strategy is to represent the data as a mixture of simple parametric components.

What are FMM? Finite Mixture Models (FMM) represent the data distribution as a weighted sum (a convex combination) of component distributions:

$$p(y_i | \pi, \theta) = \sum_{k=1}^K \pi_k f(y_i | \theta_k),$$

where $\pi_k \geq 0$, $\sum_{k=1}^K \pi_k = 1$, and each component $f(\cdot | \theta_k)$ corresponds to a latent subgroup.

A key advantage of finite mixtures is that they can generate highly flexible distributions (including skewness, heavy tails, and multimodality) while each individual component remains simple and interpretable.

Latent Indicator Variables (Z) Direct estimation of the observed-data likelihood is difficult because the sum is inside a product: where $L(\pi, \theta | y)$ denotes the likelihood function.

$$L(\pi, \theta | y) = \prod_{i=1}^N \left(\sum_{k=1}^K \pi_k f(y_i | \theta_k) \right).$$

Expanding this product requires summing over all possible allocations of N observations to K components, i.e., K^N terms, which quickly becomes computationally prohibitive.

The standard Bayesian solution introduces latent indicators $z_i = (z_{i1}, \dots, z_{iK})$ such that:

- $z_{ik} = \begin{cases} 1, & \text{if observation } i \text{ belongs to component } k, \\ 0, & \text{otherwise,} \end{cases}$
- $\sum_{k=1}^K z_{ik} = 1$ (exclusive membership).

A priori, this is equivalently a categorical allocation model¹ We denote the full allocation matrix by $Z = (z_1, \dots, z_N)$ and the component parameters by $\theta = (\theta_1, \dots, \theta_K)$.

¹Specifically, the probability mass function of a general multinomial random vector $Z = (Z_1, \dots, Z_K)$ with total

Hierarchical Representation A finite mixture model can be written hierarchically as

$$z_i \mid \pi \sim \text{Multinomial}(1; \pi_1, \dots, \pi_K), \quad y_i \mid z_i, \theta \sim \prod_{k=1}^K f(y_i \mid \theta_k)^{z_{ik}}.$$

Equivalently, conditional on $z_{ik} = 1$,

$$y_i \mid z_{ik} = 1, \theta_k \sim f(y_i \mid \theta_k).$$

This representation makes clear that dependence in the marginal model for y_i arises from integrating over the latent allocation variables.

Complete-Data Likelihood By augmenting the model with the latent indicators Z , the complete-data likelihood factorizes into a much simpler product form, which is the key reason Gibbs sampling becomes feasible. For each observation i , the complete-data contribution is

$$p(y_i, z_i \mid \theta, \pi) = p(z_i \mid \pi) p(y_i \mid z_i, \theta) = \prod_{k=1}^K [\pi_k f(y_i \mid \theta_k)]^{z_{ik}}.$$

For the full sample, the complete-data joint density becomes

$$p(y, z \mid \theta, \pi) = \prod_{i=1}^N \prod_{k=1}^K [\pi_k f(y_i \mid \theta_k)]^{z_{ik}}.$$

Prior on Weights: The Dirichlet Distribution In a Bayesian framework, the mixture weights $\pi = (\pi_1, \dots, \pi_K)$ are random variables on the simplex. They represent prior probabilities of belonging to the k -th latent subgroup, determining how prevalent each component is in the population before observing the specific value of y_i .

The natural conjugate prior for multinomial indicators is

$$\pi \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K).$$

A common weakly informative choice is symmetric, with $\alpha_k = \alpha/K$ for each component.²

For completeness, if $\pi \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$, then its density is

$$f(\pi \mid \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^K \pi_k^{\alpha_k - 1},$$

count n and category probabilities $(\pi_1, \dots, \pi_K)$ is

$$p(Z_1 = z_1, \dots, Z_K = z_K) = \frac{n!}{z_1! \cdots z_K!} \prod_{k=1}^K \pi_k^{z_k}, \quad \sum_{k=1}^K z_k = n.$$

A multinomial distribution with $n = 1$ reduces to a categorical distribution, so $z_i \sim \text{Categorical}(\pi_1, \dots, \pi_K)$.

²This specific functional form $\text{Dirichlet}(\alpha/K, \dots, \alpha/K)$ is the key mathematical step toward infinite-dimensional mixture models.

where

$$B(\alpha) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}$$

is the multivariate Beta normalizing constant.

3.2 The Bottlenecks of Finite Mixtures

While FMMs are powerful, relying on a fixed, finite number of components K introduces significant theoretical and computational challenges that motivate the transition to nonparametric methods.

- **Identifiability and Label Switching:**

Finite mixture models are hardly identifiable. The main issue is represented by invariance under label permutations: if we swap the parameters and weights of cluster 1 with those of cluster 2, the overall distribution remains identical. During MCMC simulations, the chains can randomly jump between these symmetric modes, a phenomenon known as *Label Switching*. This makes averaging the posterior draws of specific cluster parameters meaningless without imposing artificial constraints (e.g., ordering the means $\mu_1 < \mu_2 < \dots < \mu_K$).

Formally, relabeling $(1, \dots, K)$ by any permutation $(k_1, \dots, k_K)$ of $\{1, \dots, K\}$ leaves the mixture likelihood unchanged. A Bayesian specification requires a joint prior on

$$(\theta_1, \dots, \theta_K), (\pi_1, \dots, \pi_K), Z.$$

If components are exchangeable *a priori*, then the marginal posterior distributions of the weights are identical across labels, i.e.,

$$p(\pi_k | y) = p(\pi_{\hat{k}} | y), \quad \forall k, \hat{k}.$$

To illustrate this, consider a two-component Gaussian mixture ($K = 2$) with true parameters $\mu_1 = -5$, $\sigma_1 = 0.3$ and $\mu_2 = 5$, $\sigma_2 = 0.7$. Due to permutation invariance, a mathematically identical mode exists at $\mu_1 = 5$, $\sigma_1 = 0.7$ and $\mu_2 = -5$, $\sigma_2 = 0.3$. If an MCMC chain explores both symmetric modes equally during the simulation, computing the naive posterior mean for μ_1 across the entire chain will yield $(-5 + 5)/2 = 0$. This averaged estimate describes neither of the true underlying clusters, effectively destroying the component-specific information.

Hence, one cannot meaningfully estimate a posterior for a specific component (subpopulation) k unless component k is made distinguishable from the others through additional structure or identifiability constraints.

- **The Problem of Choosing K :**

In finite mixture models, K must be fixed in advance. But how do we know the true

number of latent subgroups? In practice, one typically estimates multiple models for different values of K and compares them using information criteria (e.g., BIC or DIC) or Bayes factors. This workflow is computationally expensive and may still lead to underfitting or overfitting.

3.3 Posterior Inference via Gibbs Sampling

Since the joint posterior distribution $p(\theta, \pi, Z \mid y)$ cannot be evaluated analytically, we rely on Markov Chain Monte Carlo (MCMC) methods. The introduction of the conjugate Dirichlet prior and the latent variables Z allows us to implement an elegant Gibbs Sampler. At each iteration, the algorithm proceeds in three steps:

1. **Update Latent Allocations (Z):** Conditioned on the data and the current parameters, each observation is independently assigned to a component. We draw z_i from a Multinomial distribution with probabilities given by Bayes' theorem:

$$P(z_{ik} = 1 \mid y_i, \pi, \theta) = \frac{\pi_k f(y_i \mid \theta_k)}{\sum_{j=1}^K \pi_j f(y_i \mid \theta_j)}$$

2. **Update Mixture Weights (π):** Relying on Dirichlet-Multinomial conjugacy, the weights are drawn from an updated Dirichlet distribution. The prior "counts" α/K are updated by simply adding the number of observations currently assigned to each cluster, $n_k = \sum_{i=1}^N z_{ik}$ as follows:

$$\pi \mid Z \sim \text{Dirichlet}\left(\frac{\alpha}{K} + n_1, \dots, \frac{\alpha}{K} + n_K\right)$$

3. **Update Component Parameters (θ):** The parameters θ_k for each cluster k are drawn from their full conditional posterior, using only the observations currently allocated to that cluster. Let $y_{(k)} = \{y_i : z_{ik} = 1\}$ and $n_k = |y_{(k)}|$ ³. Then:

$$p(\theta_k \mid y, Z, \pi) \propto p(\theta_k) \prod_{i: z_{ik}=1} f(y_i \mid \theta_k) = p(\theta_k) \prod_{i=1}^N f(y_i \mid \theta_k)^{z_{ik}}.$$

Under conjugacy, this yields a closed-form update:

$$\theta_k \mid y, Z \sim p(\theta_k \mid y_{(k)}).$$

Crucial note: If $n_k = 0$, the conditional reduces to the prior⁴ and we sample $\theta_k \sim p(\theta_k)$.

This dynamic of "emptying and repopulating" clusters is what allows nonparametric models to manage an infinite component space.

³This is equivalent to $n_k = \sum_{i=1}^N z_{ik}$ defined above

⁴By mathematical convention, an empty product evaluates to 1. Consequently, the posterior update elegantly collapses directly to the prior: $p(\theta_k \mid y, Z, \pi) \propto p(\theta_k) \cdot 1 = p(\theta_k)$.

The Predictive Distribution (The Bridge to Infinity) After establishing the MCMC dynamics, a natural question arises: where would a *new* observation y_{N+1} be assigned, based on the current clustering? The key step is to integrate out the random weights π :

$$P(z_{N+1} = k \mid Z) = \int P(z_{N+1} = k \mid \pi) p(\pi \mid Z) d\pi = \mathbb{E}[\pi_k \mid Z].$$

From Dirichlet–Multinomial conjugacy,

$$\pi \mid Z \sim \text{Dirichlet} \left(\frac{\alpha}{K} + n_1, \dots, \frac{\alpha}{K} + n_K \right),$$

so⁵

$$\mathbb{E}[\pi_k \mid Z] = \frac{n_k + \alpha/K}{\sum_{j=1}^K (n_j + \alpha/K)} = \frac{n_k + \alpha/K}{N + \alpha}.$$

Hence, for each existing component k ,

$$P(z_{N+1} = k \mid Z) = \frac{n_k + \alpha/K}{N + \alpha} \propto n_k + \frac{\alpha}{K}.$$

A *rich-get-richer* mechanism is then highlighted: larger clusters are more likely to attract new observations.

This naturally reveals the main limitation of finite mixtures: although the allocation mechanism is elegant and computationally tractable, the number of components K must still be fixed in advance. This leads to a fundamental question: instead of painfully trying to guess or select the right K , what if we could let the model itself dictate the number of components based on the data?

⁵If $(X_1, \dots, X_K) \sim \text{Dirichlet}(a_1, \dots, a_K)$, then $\mathbb{E}[X_k] = \frac{a_k}{\sum_{j=1}^K a_j}$. Here, $a_k = \alpha/K + n_k$.

4 What if we let $K \rightarrow \infty$?

4.1 From Finite Mixtures to the Infinite Limit

The predictive allocation rule derived at the end of the previous section already contains the key intuition for the nonparametric limit. Under the symmetric prior

$$\pi \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right),$$

integrating out the random weights gives, for a new observation,

$$P(z_{N+1,k} = 1 \mid Z) = \frac{n_k + \alpha/K}{N + \alpha}, \quad k = 1, \dots, K,$$

where

$$n_k = \sum_{i=1}^N z_{ik}$$

is the number of observations currently allocated to component k , and

$$N = \sum_{k=1}^K n_k$$

is the total sample size.

Suppose now that, among the K available components, only $K^* \leq K$ are currently occupied. Without loss of generality, after relabeling the occupied components if necessary, we may assume that these are components $1, \dots, K^*$, so that

$$n_1, \dots, n_{K^*} > 0, \quad n_{K^*+1} = \dots = n_K = 0.$$

Then the predictive rule can be separated into two cases. For an already occupied component $k \in \{1, \dots, K^*\}$,

$$P(z_{N+1,k} = 1 \mid Z) = \frac{n_k + \alpha/K}{N + \alpha}.$$

For any empty component $k \in \{K^* + 1, \dots, K\}$,

$$P(z_{N+1,k} = 1 \mid Z) = \frac{\alpha/K}{N + \alpha}.$$

We now study the behavior of these probabilities as $K \rightarrow \infty$. For each occupied component $k \leq K^*$,

$$\frac{n_k + \alpha/K}{N + \alpha} \longrightarrow \frac{n_k}{N + \alpha}.$$

Thus, the probability of joining an existing cluster remains positive and is proportional to its current size.

For any single empty component, however,

$$\frac{\alpha/K}{N+\alpha} \rightarrow 0.$$

At first sight, this might suggest that new clusters disappear in the limit. But this conclusion would be misleading: although the probability assigned to each individual empty component vanishes, the number of empty components increases with K . The relevant quantity is therefore the total probability of being allocated to *some previously empty component*:

$$\sum_{k=K^*+1}^K P(z_{N+1,k} = 1 \mid Z) = \sum_{k=K^*+1}^K \frac{\alpha/K}{N+\alpha} = \frac{(K-K^*)\alpha/K}{N+\alpha}.$$

Since K^* is fixed for any finite sample,

$$\frac{(K-K^*)\alpha/K}{N+\alpha} \rightarrow \frac{\alpha}{N+\alpha}.$$

Therefore, in the infinite limit, the predictive allocation mechanism takes the form

$$P(z_{N+1,k} = 1 \mid Z) = \frac{n_k}{N+\alpha}, \quad k = 1, \dots, K^*,$$

for existing clusters, while

$$P(z_{N+1} \text{ is allocated to a new cluster} \mid Z) = \frac{\alpha}{N+\alpha}.$$

This limit reveals the essential structure of Bayesian nonparametric clustering. Existing clusters are reinforced proportionally to their current size, while a positive amount of probability mass is always reserved for the creation of a new cluster. In this way, the model no longer requires the number of mixture components to be fixed in advance: the effective complexity of the clustering structure is allowed to grow with the data.

This limiting predictive rule is the fundamental bridge from finite mixture models to random partitions, and it admits a particularly intuitive interpretation through the Chinese Restaurant Process.

4.2 The Chinese Restaurant Process

The limiting predictive rule above admits a simple and intuitive interpretation known as the **Chinese Restaurant Process** (CRP). Imagine a restaurant with an infinite number of tables. Customers enter sequentially, and each customer chooses a table at which to sit. In the clustering interpretation, customers correspond to observations, while occupied tables correspond to clusters.

The first customer necessarily sits at the first table. Suppose now that N customers have already

entered the restaurant, and that these customers induce K^* occupied tables. Let n_k denote the number of customers currently seated at table k , for $k = 1, \dots, K^*$. Then the $(N + 1)$ -th customer is assigned according to the probabilities

$$P(\text{joins occupied table } k \mid Z) = \frac{n_k}{N + \alpha}, \quad k = 1, \dots, K^*,$$

and

$$P(\text{starts a new table} \mid Z) = \frac{\alpha}{N + \alpha}.$$

This is exactly the limiting predictive rule obtained from the finite mixture model as $K \rightarrow \infty$. The interpretation is immediate: larger tables are more likely to attract new customers, because the probability of joining an occupied table is proportional to its current size. This is the same *rich-get-richer* mechanism already observed in the predictive distribution of the finite mixture model.

At the same time, the concentration parameter α controls the tendency to generate new clusters. Larger values of α increase the probability of opening a new table, leading on average to a partition with more clusters. Smaller values of α favor allocations to already occupied tables and therefore induce fewer clusters.

An important conceptual point is that the CRP does not describe a probability distribution over a fixed finite set of labels. Instead, it describes a probability distribution over the partition structure induced by the latent allocation variables. In other words, what matters is not the numerical identity of a component label, but whether two observations are assigned to the same cluster or to different clusters.

Using the notation introduced above, this means that the CRP provides a marginal description of the allocation mechanism for the latent variables Z : conditionally on the current partition of the first N observations, the new vector z_{N+1} is obtained either by assigning the observation to one of the already occupied components with probability proportional to its current size, or by creating a previously unoccupied component with probability proportional to α .

The Chinese Restaurant Process therefore gives the correct intuition for Bayesian nonparametric clustering: observations tend to accumulate in previously occupied groups, but new groups may always emerge as additional data are observed.

However, the CRP only describes the predictive law of the induced partition. To complete the model, we still need a prior on probability measures whose predictive structure reproduces this clustering mechanism. This leads naturally to the Dirichlet process.

4.3 Motivating the Dirichlet Process through Bayesian Histograms

In many statistical problems, the goal is to estimate an unknown probability distribution or density function based on observed data. Let $y_1, \dots, y_n$ be independent and identically distributed observations from an unknown density f . A natural and intuitive approach to

density estimation is the histogram. The sample space is partitioned into a finite number of intervals (bins), and the density is assumed to be constant within each bin. Formally, given a partition $\{b_1, b_2, \dots, b_k\}$, the density can be written as

$$f(y) = \sum_{h=1}^k \mathbf{1}_{b_{h-1} < y \leq b_h} \frac{\pi_h}{b_h - b_{h-1}},$$

where $\pi = (\pi_1, \dots, \pi_k)$ is a vector of probabilities satisfying $\sum_{h=1}^k \pi_h = 1$.

In a Bayesian framework, uncertainty about the density f is expressed through uncertainty about the bin probabilities π . A natural prior for π is the Dirichlet distribution,

$$\pi \sim \text{Dirichlet}(a_1, \dots, a_k),$$

which is conjugate and allows for straightforward posterior updating. After observing the data, the posterior distribution remains Dirichlet with updated parameters $a_h + n_h$, where n_h is the number of observations falling in bin h .

Despite its simplicity and interpretability, this approach has important limitations. The number and location of the bins must be specified in advance, and the resulting estimate can be sensitive to these choices. Moreover, as the dimensionality of the data increases, the number of bins required grows rapidly, making this approach impractical.

These limitations motivate the search for a more flexible framework in which the underlying distribution is modeled directly, without relying on a fixed partition of the sample space. This leads naturally to the idea of placing a prior on the space of probability measures itself.

4.4 Measure-Theoretic Foundations

Let $(\mathcal{X}, \mathcal{B})$ be the measurable space of the observations (the sample space and its Borel σ -algebra), and let $\mathcal{M}_1(\mathcal{X})$ denote the space of all probability measures on $(\mathcal{X}, \mathcal{B})$ ⁶. A **random probability measure** P is defined on a background probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as a measurable mapping:

$$P : (\Omega, \mathcal{F}) \longrightarrow (\mathcal{M}_1(\mathcal{X}), \mathcal{E}),$$

where $\mathcal{E}$ is the smallest σ -algebra on $\mathcal{M}_1(\mathcal{X})$ making the generic evaluation maps $e_B : \mu \mapsto \mu(B)$ measurable for all $B \in \mathcal{B}$.

This formalization establishes a fundamental **duality**:

- **Measure perspective:** For any fixed $\omega \in \Omega$, the realization $P(\omega, \cdot)$ is a valid probability measure over the data space $(\mathcal{X}, \mathcal{B})$.
- **Random variable perspective:** For any fixed measurable set $B \in \mathcal{B}$, the evaluation $P(\cdot, B)$

⁶The notation $\mathcal{M}_1$ indicates more generally positive measures whose total mass over the space evaluates to exactly 1. This highlights that probability measures are a specific realization of it.

is a standard random variable defined on Ω , taking values in $[0, 1]$.

This duality is crucial because it lets us characterize the infinite-dimensional object P through the joint law of the finite-dimensional random variables $P(B_1), \dots, P(B_k)$ associated with any finite partition of $\mathcal{X}$.

Remark 1 (Conceptual interpretation of ω). It is crucial to clarify that, while formally defined as a generic element of the sample space Ω , within the context of the Dirichlet Process, ω represents the **concrete realization** of the random measure. In practical terms, ω should *not* be interpreted as a single numerical value, but rather as the complete set of "instructions" (or seed) that simultaneously determines the entire structure of the model. Fixing a value for ω is equivalent to resolving all uncertainty at once, effectively determining the infinite sequences of weights $\{\pi_h(\omega)\}_{h=1}^{\infty}$ and parameters $\{\theta_h^*(\omega)\}_{h=1}^{\infty}$ that define P . Only once ω is fixed does the object $P(\omega, \cdot)$ cease to be random and becomes a **deterministic probability measure** (i.e., a fixed distribution). This distinction is the key to understanding duality: ω accounts for the variability between different possible distributions, while the resulting measure describes the distribution of data once the underlying "rule" (ω) has been established.

Remark 2 (The Necessity of the Dirichlet Process). In real-world applications, the specific realization ω that characterizes the "true" underlying distribution of the data is unknown. If ω were fixed *a priori*, the model would reduce to a standard parametric framework with a predetermined and rigid structure.

The Dirichlet Process is therefore necessary because it provides a formal way to treat the measure P as a random object. By defining a prior distribution over the infinite-dimensional space of all possible probability measures, the DP allows us to account for our epistemic uncertainty regarding the number, weights, and locations of the clusters. Instead of arbitrarily choosing a single ω , we use the DP to weight **all** possible realizations; the Bayesian machinery then uses the observed data to update this prior, guiding us toward the most probable ω via the posterior distribution. This is the fundamental mechanism that enables the model to be "non-parametric" and adaptive.

4.5 The Dirichlet Process

One may experience as in many statistical problems, the object of interest is not a finite-dimensional parameter, but an unknown probability measure P governing the distribution of the data. This naturally leads to the problem of defining prior distributions on spaces of probability measures. As emphasized by Ferguson [3], a desirable prior in this setting should satisfy two key properties:

1. it should have large support over the space of probability measures,
2. it should admit tractable posterior updating.

Building upon the duality established above, the Dirichlet process provides a flexible and analytically tractable prior for P .

Definition 2 (Dirichlet Process). Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Given two fundamental parameters:

- P_0 (**Base measure**): A fixed probability measure on $(\mathcal{X}, \mathcal{B})$, representing the prior expectation (the mean of the process).
- α (**Concentration/precision parameter**): A positive real number ($\alpha > 0$) that controls the variance of the process (the degree of concentration of the realizations around P_0 ⁷).

A random probability measure P is said to follow a Dirichlet process, denoted as

$$P \sim \text{DP}(\alpha, P_0),$$

if, for any finite measurable partition $B_1, B_2, \dots, B_k$ of the space $\mathcal{X}$, the random vector of probabilities follows a k -dimensional Dirichlet distribution:

$$(P(B_1), P(B_2), \dots, P(B_k)) \sim \text{Dirichlet}(\alpha P_0(B_1), \alpha P_0(B_2), \dots, \alpha P_0(B_k)).$$

For any measurable set $B \in \mathcal{B}$, the marginal distribution satisfies

$$P(B) \sim \text{Beta}(\alpha P_0(B), \alpha(1 - P_0(B))),$$

so that

$$\mathbb{E}[P(B)] = P_0(B), \quad \text{Var}(P(B)) = \frac{P_0(B)(1 - P_0(B))}{1 + \alpha}.$$

Hence, the Dirichlet process is appealing in having a simple specification arising from a model similar to a random histogram but without the dependence on the bins, while also having simple and intuitive forms for the prior mean and variance. The prior can be centered on a parametric model for the distribution of the data through the choice of P_0 , while allowing to control uncertainty in this choice. Moreover, it can be shown that the support of the DP contains all probability measures whose support is contained in the support of the baseline probability measure P_0 .

An important feature of the Dirichlet process is its conjugacy. If $y_1, \dots, y_n \stackrel{iid}{\sim} P$ and $P \sim \text{DP}(\alpha, P_0)$, then the posterior distribution of P is again a Dirichlet process,

$$P \mid y_1, \dots, y_n \sim \text{DP} \left(\alpha + n, \frac{\alpha P_0 + \sum_{i=1}^n \delta_{y_i}}{\alpha + n} \right),$$

where δ_{y_i} is the Dirac delta centered on y_i ⁸ showing that α can be interpreted as a prior sample size. This conjugacy property, together with its large support, makes the Dirichlet process a

⁷Equivalently it expresses the variability of P around P_0

⁸The Dirac measure (or point mass), denoted by δ_x , is a discrete probability measure concentrated entirely at a

fundamental building block in Bayesian nonparametric inference. We now present a proof of this conjugacy property, following the constructive argument of Sethuraman (1994).

Theorem 2 (Conjugacy of the Dirichlet Process). *Let $y \mid P \sim P$ and let P follow a Dirichlet process prior, $P \sim \text{DP}(\alpha, P_0)$. The posterior distribution of P given the observation y is again a Dirichlet process with updated concentration parameter $\alpha + 1$ and updated base measure $(\alpha P_0 + \delta_y) / (\alpha + 1)$:*

$$P \mid y \sim \text{DP} \left(\alpha + 1, \frac{\alpha P_0 + \delta_y}{\alpha + 1} \right),$$

where δ_y denotes the Dirac measure centered at the observed value y . Equivalently, the posterior distribution of P given y can be formally written as the updated Dirichlet measure $D_{\alpha P_0 + \delta_y}$.

This conjugacy property naturally extends to a sample of size n . Suppose that, conditionally on P , the observations $y_1, y_2, \dots, y_n$ are independent and identically distributed (i.i.d.) according to P . We can derive the posterior distribution of P given the entire sample by relying on the principle of sequential Bayesian updating.

By a general property of Bayesian inference, the posterior distribution obtained after observing the first data point y_1 acts as the new prior for the remaining observations. Thus, given y_1 , the random measure P is distributed as $\text{DP}(\alpha + 1, (\alpha P_0 + \delta_{y_1}) / (\alpha + 1))$.

When the second observation y_2 is drawn, we can apply the theorem again treating this posterior Dirichlet process as our prior, which yields:

$$P \mid y_1, y_2 \sim \text{DP} \left(\alpha + 2, \frac{\alpha P_0 + \delta_{y_1} + \delta_{y_2}}{\alpha + 2} \right).$$

Applying this recursive logic across all n observations, it directly follows that the posterior distribution of P given the full dataset $(y_1, \dots, y_n)$ is:

$$P \mid y_1, \dots, y_n \sim \text{DP} \left(\alpha + n, \frac{\alpha P_0 + \sum_{i=1}^n \delta_{y_i}}{\alpha + n} \right).$$

This final expression is extremely intuitive: the updated parameter measure is simply the sum of the prior mass (αP_0) and the empirical mass of the observed data ($\sum_{i=1}^n \delta_{y_i}$). It clearly highlights how the concentration parameter α acts as a *prior sample size*, balancing the weight of the initial guess P_0 against the evidence provided by the n observations.

single point $x \in \mathcal{X}$. For any measurable subset $B \in \mathcal{B}$, it is formally defined as:

$$\delta_x(B) = \begin{cases} 1, & \text{if } x \in B \\ 0, & \text{if } x \notin B \end{cases}$$

The integral of any measurable function f with respect to the measure δ_x simply evaluates the function at the point x :

$$\int_{\mathcal{X}} f(y) d\delta_x(y) = f(x)$$

4.6 Stick-Breaking Construction

The definition of the Dirichlet process through finite-dimensional distributions does not provide direct intuition on the form of realizations $P \sim \text{DP}(\alpha, P_0)$. A more constructive representation, known as the stick-breaking construction, offers a useful characterization of such random probability measures.

Under this representation, a draw P from a Dirichlet process can be expressed as

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(\cdot), \quad \text{with} \quad \pi_h = V_h \prod_{l < h} (1 - V_l)^9,$$

where $V_h \sim \text{Beta}(1, \alpha)$ independently for $h = 1, 2, \dots$, and $\theta_h \sim P_0$ independently. Here, the symbol $\cdot$ denotes a generic measurable set or argument on which the random measure P acts, so $P(\cdot)$ should be read as the whole measure rather than its evaluation at a specific point. Moreover, δ_{θ} denotes a point mass at θ , the θ_h are referred to as atoms, and π_h are the corresponding weights.

This construction ensures that the weights $(\pi_h)_{h \geq 1}$ are nonnegative and sum to one almost surely. The name “stick-breaking” arises from the following interpretation: starting with a stick of unit length, we break off a portion V_1 to obtain $\pi_1 = V_1$. The remaining length $1 - V_1$ is then broken again by sampling V_2 , yielding $\pi_2 = V_2(1 - V_1)$. This process continues indefinitely, with each step allocating a fraction of the remaining stick to the next component.

Remark 3. Sampling a realization P from a Dirichlet Process $\text{DP}(\alpha, P_0)$ is equivalent to determining the infinite sequences of mixing proportions $\{\pi_h\}_{h=1}^{\infty}$ and atoms $\{\theta_h^*\}_{h=1}^{\infty}$ that constitute the random measure:

$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*} \tag{9}$$

where the weights are generated via the **stick-breaking construction** and the atoms are independent and identically distributed samples from the **base measure** P_0 .

This formulation emphasizes that the randomness of P enters through two channels: the *locations* of the atoms, which are scattered according to P_0 , and the *masses* assigned to them, which are partitioned according to the concentration parameter α . Once these sequences are drawn, the measure P is fully specified as a discrete probability distribution.

To further clarify this point, we now examine more closely the role of the parameter α in controlling the distribution of the weights.

Since $\mathbb{E}[V_h] = \frac{1}{1+\alpha}$, it follows that

$$\mathbb{E}[V_h] \rightarrow 1 \quad \text{as } \alpha \rightarrow 0, \quad \mathbb{E}[V_h] \rightarrow 0 \quad \text{as } \alpha \rightarrow \infty.$$

Hence, for small values of α , the first stick-breaking weights remain of order one, whereas for

⁹Equivalently, the weight sequence can be denoted by $\pi \sim \text{stick}(\alpha)$, meaning that the random weights are distributed according to the stick-breaking law induced by the variables V_h .

large values of α they become asymptotically negligible, leading to a more diffuse allocation of mass across many components.

An important consequence of this representation, as anticipated above, is that realizations from a Dirichlet process are almost surely discrete, consisting of a countable collection of atoms drawn from P_0 . This property explains why the Dirichlet process is not suitable as a direct prior for continuous densities, but instead serves as a building block for more flexible models such as Dirichlet process mixtures as explained in the next section.

4.7 Dirichlet Process Mixtures

Despite its appealing mathematical properties, such as conjugacy and large support, the Dirichlet process (DP) has fundamental limitations when applied directly to continuous density estimation.

1. **Lack of spatial smoothness:** A density estimator should allow observations to inform the density of nearby regions. However, the DP induces negative correlation between the measures of disjoint sets, ignoring the topological distance between them.
2. **Discreteness:** Realizations from a DP are *almost surely discrete*. As $P \sim \text{DP}(\alpha, P_0)$ is an atomic measure, it cannot possess a continuous density function with respect to the Lebesgue measure.

To overcome these issues, the DP is typically employed as a prior on an unknown mixing distribution within a hierarchical model, leading to the **Dirichlet Process Mixture (DPM)** model.

4.7.1 The Kernel Mixture Framework

A Kernel Mixture Model defines a complex probability density by blending together infinite copies of a simpler base distribution, called a kernel. Formally, it is expressed as the integral:

$$f(y | P) = \int K(y | \theta) dP(\theta), \quad (10)$$

where $K(\cdot | \theta)$ is a continuous kernel density (such as a Gaussian) and P is the *mixing measure* that governs the distribution of parameters θ .

If we place a Dirichlet process prior on the mixing measure, $P \sim \text{DP}(\alpha, P_0)$, and apply the stick-breaking representation

$$P(d\theta) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*}(d\theta),$$

then the integral simplifies into an infinite sum through the following steps:

$$f(y) = \int K(y | \theta) P(d\theta) = \int K(y | \theta) \left(\sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*}(d\theta) \right).$$

By linearity of the integral,

$$f(y) = \sum_{h=1}^{\infty} \pi_h \int K(y | \theta) \delta_{\theta_h^*}(d\theta).$$

Using the sifting property of the Dirac measure, we obtain

$$f(y) = \sum_{h=1}^{\infty} \pi_h K(y | \theta_h^*), \quad (11)$$

where $\pi \sim \text{stick}(\alpha)$ and the atoms $\theta_h^* \sim P_0$ represent the unique parameters of the infinite possible clusters.

4.7.2 Hierarchical Formulation and Predictive Inference

The DPM can be expressed through a hierarchical structure that links individual observations y_i to their specific parameters θ_i :

$$y_i | \theta_i \sim K(\cdot | \theta_i), \quad (12)$$

$$\theta_i | P \sim P, \quad (13)$$

$$P \sim \text{DP}(\alpha, P_0). \quad (14)$$

In this framework, P acts as the "engine" that weights the parameter space. Because P is discrete, multiple observations will share the same parameter value, effectively forming clusters.

The Polya Urn Scheme The predictive distribution for a new parameter θ_{n+1} , given a set of previously observed parameters $\theta_{1:n} = (\theta_1, \dots, \theta_n)$, is obtained by marginalizing out the random measure P . This yields the **Blackwell-MacQueen prediction rule**:

$$\theta_{n+1} | \theta_1, \dots, \theta_n \sim \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\theta_i}. \quad (15)$$

This formulation elegantly decomposes the logic of the model into two components:

- **Innovation** ($\frac{\alpha}{\alpha+n} P_0$): With probability proportional to α , θ_{n+1} is a new value sampled from P_0 . This corresponds to the creation of a new cluster.
- **Reinforcement** ($\frac{1}{\alpha+n} \sum \delta_{\theta_i}$): With probability proportional to the number of existing

observations, θ_{n+1} matches a previously seen value. By grouping the n observations into k distinct clusters of size n_h , this term becomes $\sum_{h=1}^k \frac{n_h}{\alpha+n} \delta_{\theta_h^*}$.

This mechanism is intuitively described by the **Pólya Urn scheme**. Imagine an urn containing an initial "latent mass" α from P_0 . Each time we draw from α , we "discover" a new parameter (a new color) and add a ball of that color to the urn. If we draw an existing color, we return it and add another identical ball.

This *rich-get-richer* property ensures that the model adapts its complexity to the data: it favors established clusters for common behaviors while allowing the birth of new clusters to capture structural breaks or anomalies in the asset's price-path.

4.7.3 Posterior Computation and Marginal Gibbs Sampling

The conditional representation derived above provides a useful foundation for posterior computation. In principle, one could update the parameters θ_i sequentially using their full conditional distributions, which retain the same structure as the prior when the base measure P_0 is conjugate to the kernel K . However, this approach may lead to poor mixing, as it does not explicitly separate cluster allocations from component-specific parameters.

A more convenient strategy is to work directly with the clustering structure and introduce explicit allocation variables. Since many observations will share the same parameter value, we can separate the unique parameters from the cluster assignments. Let $\theta^* = (\theta_1^*, \dots, \theta_{k^*}^*)$ be the distinct cluster-specific parameter values currently active, and define $S_i = c$ whenever $\theta_i = \theta_c^*$. Thus, S_i simply records which cluster observation i belongs to.

Based on this representation, an alternative marginal Gibbs sampler can be constructed, which alternates between updating the cluster allocations and the corresponding cluster-specific parameters. In the allocation step, each observation is assigned either to an existing cluster or to a new one.

The allocation probability is proportional to:

Existing cluster c	New cluster
$\propto n_c^{(-i)} K(y_i \mid \theta_c^*)$	$\propto \alpha \int K(y_i \mid \theta) dP_0(\theta)$
Weight from current partition: $n_c^{(-i)}$	Innovation weight: α
Data fit: $K(y_i \mid \theta_c^*)$	Prior predictive fit: $\int K(y_i \mid \theta) dP_0(\theta)$

Here, $n_c^{(-i)}$ denotes the size of cluster c after removing observation i from the current partition. This term comes from the conditional prior induced by the Dirichlet process and captures the

usual rich-get-richer effect: larger occupied clusters are more likely to attract observation i . By contrast, the term $\alpha \int K(y_i | \theta) dP_0(\theta)$ corresponds to the creation of a new cluster: α controls the prior propensity to generate a new component, while the integral is the prior predictive contribution obtained by averaging the kernel under the base measure P_0 .

In the parameter update step, the distinct values θ_c^* are sampled from their posterior distribution,

$$p(\theta_c^* | -)^{10} \propto P_0(\theta_c^*) \prod_{i:S_i=c} K(y_i | \theta_c^*),$$

which is simply the posterior distribution under the parametric model that assigns the prior distribution P_0 to the parameters θ^* and then updates this prior with the likelihood for those subjects assigned to cluster c .

This marginal approach *avoids explicit representation* of the infinite-dimensional mixing measure P and allows the number of clusters to vary across iterations, providing a flexible and computationally efficient alternative for posterior inference in Dirichlet process mixture models.

4.7.4 Blocked Gibbs Sampler via Stick-Breaking

An alternative approach to posterior computation avoids marginalizing out the random probability measure P and instead relies on its stick-breaking representation. While marginalization simplifies the structure of the model, it prevents direct inference on P and may *not be feasible* in more general settings. By working with the stick-breaking construction, one can retain an explicit representation of the mixing measure.

Since later stick-breaking weights are multiplied by the remaining stick length, they tend to become smaller as the index h increases, although they are not necessarily monotonically decreasing in every realization. In particular, one considers

$$P \approx \sum_{h=1}^N \pi_h \delta_{\theta_h^*},$$

where the truncation level N is chosen such that the remaining mass $\sum_{h=N+1}^{\infty} \pi_h$ is **negligible**. In practice, moderate values such as $N = 30$ or 50 are often sufficient, providing an upper bound on the number of occupied clusters. Under this approximation, posterior inference can be carried out using a blocked Gibbs sampler.

Let $S_i \in \{1, \dots, N\}$ denote the allocation of observation i to component c , that is, $S_i = c$ if $\theta_i = \theta_c^*$. The algorithm alternates between the following steps:

¹⁰Here the symbol $-$ means “all the remaining quantities are held fixed.” In this context, it is shorthand for conditioning on all variables other than θ_c^* , such as the current allocations, the data, the other cluster-specific parameters, and any fixed hyperparameters.

1. Update the allocation variables by sampling

$$\Pr(S_i = c \mid -) = \frac{\pi_c K(y_i \mid \theta_c^*)}{\sum_{h=1}^N \pi_h K(y_i \mid \theta_h^*)}, \quad c = 1, \dots, N.$$

2. Update the stick-breaking weights $V_c, c = 1, \dots, N - 1$, from

$$V_c \sim \text{Beta} \left(1 + n_c, \alpha + \sum_{c'=c+1}^N n_{c'} \right),$$

where n_c is the number of observations assigned to cluster c .

3. Update the component-specific parameters θ_c^* from their posterior distribution,

$$p(\theta_c^* \mid -) \propto P_0(\theta_c^*) \prod_{i:S_i=c} K(y_i \mid \theta_c^*),$$

which coincides with the standard Bayesian update under the parametric model defined by P_0 and K .

This algorithm is straightforward to implement and provides a practical alternative to marginal approaches. In addition, the induced density

$$f(y) = \sum_{h=1}^N \pi_h K(y \mid \theta_h^*)$$

can be monitored across iterations to assess convergence and perform posterior inference.

4.7.5 Hyperprior for the Concentration Parameter

The concentration parameter α plays a crucial role in controlling the number of clusters in the model. Fixing α may lead to limited flexibility, and it is often preferable to assign a prior distribution to α , allowing the data to inform its value.

A common choice is a Gamma prior,

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha).$$

For the blocked Gibbs sampler, this prior is conditionally conjugate, leading to the full conditional distribution

$$\alpha \mid - \sim \text{Gamma} \left(a_\alpha + N - 1, b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h) \right).$$

This additional update step allows adaptive control of the clustering behavior. In particular, smaller values of α tend to favor fewer clusters, while larger values encourage more dispersed

allocations. Importantly, the Bayesian framework naturally penalizes overfitting through the marginal likelihood, avoiding the proliferation of unnecessary mixture components.

An important and often subtle aspect of Dirichlet process mixture models is the choice of the base measure P_0 . While P_0 is frequently selected for computational convenience, for example within conjugate parametric families, its specification has a significant impact on the induced clustering behavior of the model.

In particular, P_0 determines the prior distribution of the cluster-specific parameters θ_h^* , and hence influences where clusters are likely to be located in the parameter space. If P_0 is overly diffuse, assigning mass over a very wide region, then the marginal likelihood tends to penalize the introduction of multiple clusters, as each new cluster is associated with parameters in low-probability regions under P_0 . As a result, the model may exhibit a tendency to favor fewer clusters, potentially leading to underfitting.

Conversely, a more informative choice of P_0 that concentrates mass near plausible values of the data can facilitate the formation of multiple clusters and improve model flexibility. For this reason, it is often recommended to standardize the observed data prior to analysis and to choose P_0 accordingly, for example by centering it at zero with unit variance. This ensures that the prior is appropriately scaled relative to the data, avoiding unintended sensitivity to the units of measurement.

Overall, both P_0 and the concentration parameter α play complementary roles in controlling the complexity of the model, and careful specification of these quantities is crucial for obtaining reliable posterior inference.

Note. In practice, it is important to verify that the truncation level N is sufficiently large, for example by monitoring the maximum occupied component index. While Gibbs samplers based on stick-breaking representations are effective, they may suffer from slow mixing due to multimodality of the posterior. Techniques such as label-switching moves or split-merge algorithms can be employed to improve mixing in such cases.

4.7.6 Non-Conjugate Models and Auxiliary Variable Methods

When the base measure P_0 is not conjugate to the kernel K , the marginal Gibbs sampler becomes difficult to implement due to the intractability of the integral

$$\int K(y_i | \theta) dP_0(\theta).$$

To address this issue, an alternative approach introduces auxiliary variables, allowing for an exact Gibbs update without requiring analytical integration.

The auxiliary variable method proposed by Neal [6] proceeds by augmenting the state space with a finite number m of auxiliary parameters, which represent candidate values for new

clusters. Let $\theta_1^*, \dots, \theta_{k^-}^*$ denote the distinct cluster-specific parameters associated with the observations excluding i , where k^- is the number of clusters in θ_{-i} . To update the allocation variable S_i , we introduce m auxiliary parameters $\tilde{\theta}_{k^-+1}, \dots, \tilde{\theta}_{k^-+m}$ sampled independently from P_0 .

Since the observations y_i are exchangeable, we can consider the conditional distribution of S_i given the remaining allocations S_{-i} . The probability of assigning observation i to an existing cluster $c \in \{1, \dots, k^-\}$ is given by

$$P(S_i = c \mid S_{-i}) = \frac{n_{-i,c}}{n-1+\alpha},$$

where $n_{-i,c}$ denotes the number of observations assigned to cluster c , excluding observation i .

The probability of assigning observation i to a new cluster is

$$P(S_i \text{ new} \mid S_{-i}) = \frac{\alpha}{n-1+\alpha},$$

which is divided equally among the m auxiliary components. Hence, for each auxiliary component $c = k^- + 1, \dots, k^- + m$, we have

$$P(S_i = c \mid S_{-i}) = \frac{\alpha}{m(n-1+\alpha)}.$$

Combining these prior weights with the likelihood, the full conditional distribution used in the Gibbs update is proportional to

$$P(S_i = c \mid S_{-i}, \theta^*, \tilde{\theta}, y_i) \propto \begin{cases} \frac{n_{-i,c}}{n-1+\alpha} K(y_i \mid \theta_c^*) & \text{for } c = 1, \dots, k^-, \\ \frac{\alpha}{m(n-1+\alpha)} K(y_i \mid \tilde{\theta}_c) & \text{for } c = k^- + 1, \dots, k^- + m. \end{cases}$$

If S_i is assigned to one of the auxiliary components, a new cluster is created with parameter value given by the corresponding $\tilde{\theta}_c$. After the update, all auxiliary parameters not associated with any observation are discarded.

The cluster-specific parameters are then updated from their posterior distribution

$$p(\theta_c^* \mid -) \propto P_0(\theta_c^*) \prod_{i:S_i=c} K(y_i \mid \theta_c^*),$$

as in a standard finite mixture model.

This algorithm provides an exact Gibbs sampling scheme for Dirichlet process mixture models in the non-conjugate setting. The auxiliary variables are introduced only temporarily during each update and allow the sampler to avoid intractable integrals while preserving the correct posterior distribution as the stationary distribution of the Markov chain.

5 A practical application

5.1 Modeling Financial Returns: The Heavy-Tail Problem

The theoretical framework developed in the previous sections provides a powerful tool for density estimation and clustering. To demonstrate its practical relevance, we apply the Dirichlet Process Mixture (DPM) model to a fundamental problem in quantitative finance: the modeling of asset returns and the quantification of tail risk, following the methodology proposed by Das et al. [2].

Remark 4. While the original paper extends this framework to multivariate portfolio analysis using elliptical copulas and introduces additional advanced risk metrics (such as Wang’s Transform), we restrict our empirical application to the univariate case. A full theoretical treatment of these financial tools is beyond the scope of our manuscript, which remains focused on showing the foundational mechanics and posterior computation of Bayesian statistical models.

Let S_t denote the price of a financial asset at time t . The standard practice in financial modeling, dating back to the Black-Scholes framework, assumes that the asset price follows a Geometric Brownian Motion (GBM). Under this assumption, the continuous compounded returns (log-returns), defined as $r_t = \log(S_t/S_{t-1})$, are independent and normally distributed:

$$r_t \sim \mathcal{N}(\mu, \sigma^2),$$

where μ represents the drift (expected return) and σ^2 is the constant volatility.

However, decades of empirical evidence have consistently shown that real-world financial returns severely violate the normality assumption. Empirical return distributions exhibit significant skewness and *excess kurtosis* (heavy tails). In a standard normal distribution, extreme market events (crashes or sudden rallies) are considered statistically virtually impossible. In reality, these extreme events occur with alarming frequency.

Parametric attempts to fix this issue often involve fitting specific heavy-tailed distributions (like the Student’s t -distribution) or finite mixture models. As discussed in Section 3, while a finite mixture of normals can capture different "volatility regimes" (e.g., a high-probability low-volatility component for normal trading days, and a low-probability high-volatility component for market crashes), determining the exact number of regimes K in advance is practically impossible and computationally inefficient.

The Bayesian Nonparametric approach offers an elegant and data-adaptive solution. By treating the log-returns as observations in our DPM model, we allow the data to naturally cluster into

an unspecified number of volatility regimes. Formally, we model the log-returns r_t as:

$$\begin{aligned} r_t \mid \mu_t, \sigma_t^2 &\sim \mathcal{N}(\mu_t, \sigma_t^2), \\ (\mu_t, \sigma_t^2) \mid P &\sim P, \\ P &\sim \text{DP}(\alpha, P_0). \end{aligned}$$

In this setup, the Dirichlet Process prior automatically identifies the underlying structure of the market. Most trading days will be assigned to a dominant cluster characterized by average returns and low variance. However, during periods of market stress, the stick-breaking mechanism will allocate observations to new, high-variance clusters, effectively generating the heavy tails observed in empirical data without the need to force a rigid parametric shape.

5.2 Implications for Risk Management: VaR and Expected Shortfall

The choice of the underlying probability distribution for asset returns is not merely an academic exercise; it has profound consequences for risk management. Financial institutions rely on risk metrics to determine the amount of capital they must hold in reserve to survive adverse market movements.

To formalize this, let X represent the loss of a financial position over a given time horizon (where positive values of X represent losses). A risk measure is a function that assigns a real number to the random variable X , representing the capital required to cover the risk.

Value at Risk (VaR) The most widely used risk metric in the financial industry is Value at Risk. For a given confidence level $q \in (0, 1)$ (typically 0.95 or 0.99), VaR_q is defined as the smallest loss such that the probability of the actual loss exceeding it is at most $1 - q$:

$$\text{VaR}_q(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq q\}.$$

In statistical terms, VaR is simply the q -th quantile of the loss distribution. While simple to interpret, VaR suffers from severe theoretical and practical deficiencies. Most notably, it is not a *coherent* risk measure because it lacks **sub-additivity**: the VaR of a diversified portfolio can paradoxically be strictly greater than the sum of the VaRs of its individual components. Furthermore, VaR is completely "blind" to tail severity; it tells us the minimum loss in the worst 5% of cases, but provides absolutely no information on *how bad* those losses can get.

Conditional Value at Risk (CVaR) To overcome the limitations of VaR, Artzner et al. (1999) formalized the concept of coherent risk measures. The most prominent alternative is the Conditional Value at Risk (CVaR), also known as Expected Shortfall (ES). CVaR evaluates the expected loss strictly in the tail of the distribution, conditional on the loss exceeding the VaR

threshold:

$$CVaR_q(X) = \mathbb{E}[X \mid X > VaR_q(X)].$$

Unlike VaR, CVaR is a coherent risk measure, meaning it correctly rewards diversification and explicitly accounts for the magnitude of extreme losses.

The DPM Advantage in Risk Measurement The true power of the DPM model is revealed when estimating CVaR. If a risk manager assumes a standard Normal distribution (GBM), the exponential decay of the Gaussian tails will lead to an artificially low CVaR, providing a false sense of security.

By substituting the standard parametric loss distribution with the **posterior predictive distribution** obtained from the Blocked Gibbs Sampler, we incorporate both the structural heavy-tails of the market and the epistemic uncertainty regarding the parameters. The resulting CVaR calculated under the DPM framework will be significantly more conservative and realistic, ensuring that the tail risk driven by rare but devastating market shocks is adequately quantified.

5.3 Data and Computational Setup

To empirically demonstrate the advantages of the Dirichlet Process Mixture model, we analyze the daily log-returns of the S&P 500 index. The dataset comprises 10,079 daily observations spanning from January 1, 1985 to January 1, 2025.

Prior to estimation, the log-returns are standardized to have zero mean and unit variance. This standardization is crucial, as discussed in Section 4.7.3, to ensure that the hyperparameters of the base measure P_0 are appropriately scaled relative to the data, avoiding unintended sensitivity to the units of measurement.

For the posterior computation, we employ the **Blocked Gibbs Sampler** described in Section 4.7.4. We approximate the infinite-dimensional Dirichlet Process by truncating the stick-breaking representation at a maximum of $N = 30$ components.

The base measure P_0 is chosen to be the conjugate Normal-Inverse-Gamma distribution:

$$P_0(\mu, \sigma^2) = \mathcal{N}(\mu \mid m_0, \sigma^2/c_0) \times \text{Inverse-Gamma}(\sigma^2 \mid a_0, b_0).$$

This conjugate specification allows for exact and highly efficient Gibbs updates for the cluster-specific parameters. We set a weakly informative prior for the concentration parameter $\alpha = 1.0$. The model autonomously identified 8 active volatility regimes to explain the historical log-returns.

The algorithm was run to convergence, and 100,000 posterior predictive samples were generated from the fitted Dirichlet Process Mixture to compute robust risk metrics.

5.4 Empirical Results and Density Estimation

Once the algorithm converges, we can compute the posterior predictive density of the log-returns. Figure 1 compares the empirical distribution of the data (histogram) with two fitted models: the standard parametric Normal distribution and the nonparametric DPM model.

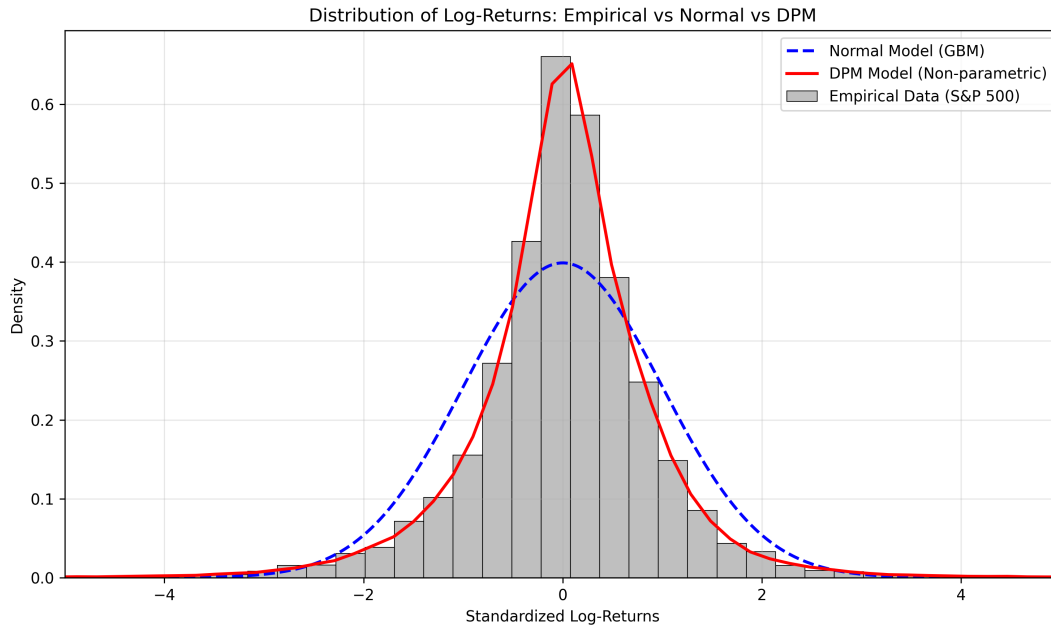


Figure 1: Comparison of the empirical log-return distribution with the standard Normal fit and the DPM posterior predictive density.

Here the difference is dramatic. While the Gaussian model fails to capture the observed tail behavior, the DPM model better reproduces heavy tails and produces substantially larger, and arguably more realistic, tail-risk estimates.

5.5 Risk Quantification Comparison

To quantify the impact of our modeling choice on risk management, we compute the 95% and 99% Value at Risk (VaR) and Conditional Value at Risk (CVaR) under both the standard Normal assumption and our DPM framework. The results are summarized in Table 1.

Metric	Normal Model (GBM)		DPM Model	
	95% Level	99% Level	95% Level	99% Level
VaR	1.6439	2.3265	1.5181	2.8123
CVaR	2.0595	2.6520	2.4359	4.2906

Table 1: Comparison of standardized risk metrics (VaR and CVaR) estimated under the parametric Normal model and the nonparametric DPM model.

Most notably, at the 99% confidence level, the Expected Shortfall (CVaR) estimated by the DPM model increases by over 60% relative to the Normal model. This divergence underscores how

risk management decisions and capital reserve requirements shift when moving from rigid parametric assumptions to a data-adaptive, nonparametric framework.

6 Conclusion

In this manuscript, we have traversed the mathematical landscape of Bayesian inference, tracking its evolution from classical parametric frameworks to infinite-dimensional nonparametric formulations. By addressing the rigid bottlenecks inherent in finite mixture models (specifically the challenge of choosing the configuration space K and the complications of label switching), we illustrated how the Dirichlet Process and its underlying stick-breaking mechanics offer an elegant, data-adaptive solution.

Our practical application to the historical log-returns of the S&P 500 index demonstrates that these theoretical concepts have profound real-world implications. Where standard Geometric Brownian Motion frameworks fail by severely underestimating the frequency and severity of market shocks, the Dirichlet Process Mixture model naturally accommodates historical irregularities, identifying latent volatility regimes dynamically. The resulting 60% increase in the 99% Conditional Value at Risk highlights that Bayesian nonparametrics is not merely a rigorous academic exercise, but an essential framework for robust institutional risk management and realistic tail-risk quantification.

References

- [1] Blackwell, D., & MacQueen, J. B. (1973). *Ferguson distributions via Polya urn schemes*. The Annals of Statistics, 1(2), 353–355.
- [2] Das, S., Halder, A., Lahiri, A., & Dey, D. K. (2018). *Modeling Risk and Return using Dirichlet Process Prior*. arXiv preprint arXiv:1805.00306. Available at: <https://arxiv.org/abs/1805.00306>
- [3] Ferguson, T. S. (1974). *Prior distributions on spaces of probability measures*. The Annals of Statistics, 2(4), 615–629.
- [4] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC.
- [5] Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer Science & Business Media.
- [6] Neal, R. M. (2000). *Markov Chain Monte Carlo Methods for Dirichlet Process Mixture Models*. Journal of Computational and Graphical Statistics, 9(2), 249–265.
- [7] Sethuraman, J. (1994). *A constructive definition of Dirichlet priors*. Statistica Sinica, 4(2), 639–650.

A Appendix

A.1 Derivation of Marginal Dirichlet Process Moments

By definition, for any measurable subset $B \in \mathcal{B}$, a random probability measure $P \sim \text{DP}(\alpha, P_0)$ evaluated at B follows a Beta distribution:

$$P(B) \sim \text{Beta}(\alpha P_0(B), \alpha(1 - P_0(B))).$$

Let $a = \alpha P_0(B)$ and $b = \alpha(1 - P_0(B))$. Using the standard properties of the Beta distribution, we can formally derive the prior expectation and variance of the random measure.

Expectation Derivation The expected value is given by:

$$\mathbb{E}[P(B)] = \frac{a}{a+b} = \frac{\alpha P_0(B)}{\alpha P_0(B) + \alpha(1 - P_0(B))} = \frac{\alpha P_0(B)}{\alpha(P_0(B) + 1 - P_0(B))} = P_0(B).$$

This mathematically justifies interpreting the base measure P_0 as the center of mass or prior mean of the Dirichlet Process.

Variance Derivation The variance of a Beta random variable is defined as $\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$. Substituting our parameters:

$$\begin{aligned} \text{Var}(P(B)) &= \frac{(\alpha P_0(B))(\alpha(1 - P_0(B)))}{(\alpha P_0(B) + \alpha(1 - P_0(B)))^2(\alpha P_0(B) + \alpha(1 - P_0(B)) + 1)} \\ &= \frac{\alpha^2 P_0(B)(1 - P_0(B))}{\alpha^2(\alpha + 1)} \\ &= \frac{P_0(B)(1 - P_0(B))}{\alpha + 1}. \end{aligned}$$

This highlights that the precision parameter α acts inversely to the variance. As $\alpha \rightarrow \infty$, the variance vanishes, forcing the random realizations of P to collapse entirely onto the smooth base measure P_0 .

A.2 Algorithmic Steps of the Blocked Gibbs Sampler

To implement the financial regime-switching application outlined in Section 5, the infinite-dimensional Dirichlet Process Mixture is truncated at a finite bound N . Let $\mathbf{r} = (r_1, \dots, r_T)$ denote the historical series of standardized S&P 500 log-returns.

Consistent with Section 3, we define the latent indicators for each trading day as a row vector $\mathbf{z}_t = (z_{t1}, \dots, z_{tN})$ where $z_{tk} = 1$ if day t belongs to regime k , and 0 otherwise. At each MCMC

iteration s , the algorithm cycles sequentially through the following updates:

Step 1: Cluster Allocation Matrix Update For each individual trading day $t \in \{1, \dots, T\}$, update its latent indicator vector z_t by sampling from a Categorical allocation model (Multinomial with $n = 1$) with probabilities:

$$P(z_{tk} = 1 \mid \mathbf{r}, \pi, \theta) = \frac{\pi_k \mathcal{N}(r_t \mid \mu_k, \sigma_k^2)}{\sum_{j=1}^N \pi_j \mathcal{N}(r_t \mid \mu_j, \sigma_j^2)}, \quad k = 1, \dots, N.$$

Step 2: Weight Refinement via Stick-Breaking Compute the updated component counts representing the number of trading days currently assigned to each volatility regime: $n_k = \sum_{t=1}^T z_{tk}$. Sample the random breaking fractions V_k for $k = 1, \dots, N - 1$ from the updated conditional Beta law:

$$V_k \mid Z \sim \text{Beta} \left(1 + n_k, \alpha + \sum_{k'=k+1}^N n_{k'} \right).$$

Set $V_N = 1$ to force closure on the simplex. Reconstruct the global mixture weights using the multiplication chain rule: $\pi_k = V_k \prod_{l < k} (1 - V_l)$.

Step 3: Regime-Specific Parameter Updates For each active component k , extract the subset of returns matching that regime: $\mathbf{r}_{(k)} = \{r_t : z_{tk} = 1\}$. Since the baseline P_0 is chosen as a conjugate Normal-Inverse-Gamma distribution, sample the updated drift and localized volatility parameters simultaneously from their full conditionals:

$$\sigma_k^2 \mid \mathbf{r}_{(k)}, Z \sim \text{Inverse-Gamma}(a_n, b_n), \quad \mu_k \mid \sigma_k^2, \mathbf{r}_{(k)}, Z \sim \mathcal{N}(m_n, \sigma_k^2 / c_n).$$

Note: If a component empties out ($n_k = 0$), the conditional update elegantly collapses directly to the prior specification, and we sample $(\mu_k, \sigma_k^2) \sim P_0$.

Step 4: Precision Regularization Update Sample the global concentration parameter α to allow the network to autonomously regulate cluster sensitivity:

$$\alpha \mid V \sim \text{Gamma} \left(a_\alpha + N - 1, b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h) \right).$$

A.3 Comprehensive Notation Taxonomy

To optimize clarity throughout the text, Table 2 catalogs the comprehensive operational framework utilized across the theoretical sections and empirical modules of this study.

Symbol	Statistical Definition	Structural Context
r_t	Asset continuous log-return at time t	Empirical data vector
z_{tk}	Latent allocation matrix indicator ($\sum_{k=1}^N z_{tk} = 1$)	Observational cluster map
π_k	Realized mixture probability weight on the simplex	Convex component allocation
V_k	Random breaking fraction ($V_k \sim \text{Beta}(1, \alpha)$)	Stick-breaking backbone
α	Mass concentration parameter	Clustering growth controller
P_0	Prior expectation functional measure	Normal-Inverse-Gamma baseline
δ_x	Dirac delta point-mass operator	Empirical distribution building-block
VaR_q	Quantile loss threshold ($q \in (0, 1)$)	Value at Risk baseline
CVaR_q	Integrated tail expectation profile	Expected Shortfall risk profile

Table 2: Taxonomy of symbols, distributions, and functional operators.

A.4 Notation Convention: $P(\cdot)$ vs $p(\cdot)$

In this manuscript we consistently use:

- $P(\cdot)$ for event probabilities (typically for discrete variables), e.g.

$$P(z_{tk} = 1 \mid r_t, \pi, \theta).$$

- $p(\cdot)$ for densities (continuous variables) or generic probability laws in Bayesian expressions, e.g., $p(\mu_k, \sigma_k^2 \mid \mathbf{r}, \mathbf{Z})$.

For latent indicators z_{tk} , both notations are common in the literature. We retain $P(\cdot)$ when writing point probabilities and $p(\cdot)$ in full-conditionals written up to proportionality.