To my beloved parents, siblings, grandparents, and to my whole family.

To all my friends, in particular the dearest.

To my precious supervisor Professor Antonio Lijoi.

A Bayesian Non-parametric Framework for Modelling Implied Volatility in Option Pricing

Theory and Applications

Filippo Grandoni

Supervised by

Professor Antonio Lijoi

Bocconi University

July 2025

Contents

Chapter 1

Introduction

1.1 Overview

The accurate pricing of financial derivatives, particularly options, has been a fundamental challenge in financial economics. Since the early 1970s, the predominant approach has been the Black–Scholes model (Black and Scholes, 1973), which provides a closed-form solution for option pricing under the assumption of log-normal asset price dynamics. In the 1990s, the Heston model (Heston, 1993) introduced stochastic volatility to better capture observed market phenomena. Despite their theoretical elegance, both models rely on rigid parametric assumptions that often fail to reflect real-world financial market complexities.

Bayesian non-parametric (BNP) methods offer an alternative approach, allowing for greater flexibility in modeling the latent processes that underlie observed prices. This study explores Dirichlet Process Mixture Models (DPMMs) and Bayesian non-parametric regression as tools to improve the prediction of the *implied volatility surface*, a latent structure derived from market prices and fundamental to option valuation. Once estimated, this surface can be used within standard option pricing formulas, thereby circumventing one of the key limitations of traditional models: their dependence on fixed, rigid parametric assumptions for computing the IV.

1.2 Motivation

1.2.1 Limitations of Traditional Models

Two fundamental frameworks for option pricing are the Black–Scholes and Heston models. The former assumes constant volatility, while the latter introduces stochastic volatility. Despite enhancements, both rely on strict parametric assumptions such as log-normal returns and mean-reverting volatility.

These assumptions, while analytically convenient, often fail to reflect the empirical characteristics of financial markets. The assumption of log-normality and constant implied volatility in the Black-Scholes model, for instance, is frequently violated in real-world data, where asset returns exhibit fat tails and non-Gaussian properties as per Cont (2001), while IV shows changin behaviours. Empirical examinations have revealed discrepancies between the Black-Scholes model's predictions and observed market prices, as per (MacBeth and Merville, 1979).

Similarly, although the Heston model captures stochastic volatility, its underlying structure may still impose unrealistic constraints on the evolution of market prices. Studies assessing the empirical performance of Heston's stochastic volatility model have identified challenges in accurately capturing the dynamics of financial markets, especially during periods of market stress, as explained in Poon and Granger (2003). As a result, these models can lead to systematic mispricing of derivatives, particularly in periods of heightened market turbulence.

1.2.2 Option Pricing Basics and Market Characteristics

For introducing the main topic of this study, it is appropriate to give some theoretical financial foundations. Options are financial derivatives that provide the holder with the right, but not the obligation, to buy (call option) or sell (put option) an underlying asset at a predetermined strike price on or before a specified expiration date. The valuation of these instruments is influenced by several key factors:

Intrinsic Value: This is the immediate exercise value of the option. It is determined for a call option as the difference between the strike price and the current price of the underlying asset, if the difference is positive; if not, it is zero. It is, on the other hand, the difference, if positive, between the strike price and the current price of the underlying asset for a put option.

Time Value: Reflecting the potential for an option to gain value before expiration, time value diminishes as the expiration date approaches, a phenomenon known as time decay. Factors such as the volatility of the underlying asset, time until expiration, and prevailing interest rates contribute to an option's time value.

Empirical studies of asset returns and option prices reveal several statistical properties that challenge traditional parametric modeling approaches.
Specifically, financial markets exhibit:

Skewness and excess kurtosis: Return distributions frequently display asymmetry and heavy tails, deviating from the Gaussian assumptions embedded in classical models.

Stochastic volatility and volatility clustering: Asset price fluctuations tend to be persistent over time, with periods of high volatility followed by periods of relative calm, a feature that is only partially addressed by stochastic volatility models.

Nonlinear dependencies and jumps: Market shocks and discrete price jumps introduce discontinuities that cannot be easily captured by continuous-time parametric processes.

These empirical characteristics highlight how restrictive parametric models are inherently unable to fully capture the complexity of financial markets. The inability of conventional models to flexibly adapt to observed data motivates the investigation of Bayesian non-parametric (BNP) techniques, which provide a more data-driven and adaptive framework. In particular, BNP methods can be used to predict the entire implied volatility surface—an object not directly observable but inferred from market prices—thus improving option pricing without requiring unrealistic structural assumptions.

Formally, *implied volatility* σ_{imp} is defined as the unique value of the volatility parameter which, when inserted into the chosen pricing model (e.g. Black–Scholes), equates the model-derived option price to the observed market price:

$$C_{\mathrm{BS}}\big(S_t,\, K,\, T,\, r,\, \sigma_{\mathrm{imp}}\big) \;=\; C_{\mathrm{mkt}}\big(t;\; S_t,\, K,\, T,\, r\big),$$

where S_t denotes the underlying asset price at the valuation time t, while K represents the option's strike price. The time to maturity is indicated by T, expressed in years, such that the option expires at time $t+T$. The parameter r corresponds to the continuously compounded risk-free interest rate. The implied volatility, denoted by σ_{imp}, is defined as the unique value of volatility that equates the model price to the observed market price. The function $C_{\text{BS}}(\cdot)$ refers to the Black–Scholes pricing formula, or any alternative model employed. Finally, $C_{\text{mkt}}(t; S_t, K, T, r)$ denotes the observed market price of the option at time t.

1.3 Objective

This research investigates the application of Dirichlet Process Mixture Models (DPMMs), introduced by Lo (1984), and Bayesian non-parametric regression techniques to predict the implied volatility surface, which is not directly observable, while it is a crucial factor for option pricing. Rather than modeling option prices directly through fixed structural forms, the focus here is on estimating the latent volatility surface with greater flexibility and statistical rigor, to consequently use it in option pricing.

Building on a flexible Bayesian non-parametric clustering of implied-volatility data, we introduce a two-stage regression framework that first captures regime-specific surface shapes and then refines each cluster with a local linear 'tilt', providing a tractable yet markedly more adaptable alternative to standard parametric approaches, *see Section 3.2.3*.

Indeed, unlike conventional parametric approaches, BNP methods provide a probabilistic framework in which the complexity of the data determines the structure of the model, avoiding the imposition of an a priori functional form. By leveraging the Dirichlet process prior, DPMMs enable automatic model complexity selection, yielding a data-driven approximation of return distributions and volatility patterns.

Bayesian non-parametric regression methods similarly allow for flexible and locally adaptive estimation of the implied volatility surface, capturing nonlinearities and structural shifts across moneyness and time-to-maturity dimensions. Once estimated, this surface can be plugged into standard option pricing formulas such as Black–Scholes or Heston, thereby bypassing their main limitation: the need to assume fixed, often unrealistic parameter values.

This study aims to demonstrate that BNP methods—particularly DPMMs—provide more accurate and robust predictions of implied volatility surfaces, and this can lead to improved option pricing and risk assessment in practice.

1.4 Thesis Structure

This dissertation's remaining sections are organised as follows:

A thorough analysis of Bayesian approaches, with an emphasis on non-parametric techniques, is given in Chapter 2, which also explains their theoretical underpinnings and argues that they are appropriate for use in financial modelling.

The construction of a BNP model for Implied Volatility prediction is the main topic of Chapter 3, which also emphasises the theoretical use of Dirichlet Process Mixture Models in capturing the statistical characteristics of financial returns.

Using S&P 500 option market data, Chapter 4 empirically applies the suggested BNP framework and assesses how accurate its IV prediction is in comparison to more conventional models used as benchmarks.

A review of the main conclusions, their ramifications for financial modelling, and possible avenues for further study in Bayesian non-parametrics and the employment of them in IV prediciton round off Chapter 5.

This thesis aims to provide a data-driven substitute for conventional parametric models by incorporating BNP approaches into the field of derivatives pricing, improving the flexibility and resilience of financial market research.

Chapter 2

Elements of Bayesian Inference

2.1 An overview of Bayesian vs. Frequentist approaches

The starting point of the analysis is the fundamental difference between Bayesian and frequentist statistics. Frequentist inference regards parameters as fixed but unknown constants. Under the frequentist interpretation, probability describes the frequency with which data outcomes occur when are repeatedly sampled from a specified likelihood model that has fixed (but unknown) parameter values. Additionally, parametric treatments might assume independent and identically distributed (IID) observations for simplicity, but this is not a strict requirement of the frequentist paradigm.

In contrast, Bayesian inference conceptualizes parameters as random variables and leverages probability to quantify uncertainty about them directly. Consequently, Bayesian methods fully specify both likelihood and prior and produce a posterior distribution, allowing flexible conditional independence or hierarchical structures that more realistically capture complex data dependencies. Several key distinctions between these frameworks highlight why the Bayesian approach can be considered more suitable in many practical contexts:

First, the Bayesian framework explicitly accommodates prior knowledge about the parameters through the prior distribution. Lavine (1999) argues convincingly that disregarding available prior information - exactly what frequentist methods do - amounts to discarding valuable insights before analyzing the data, akin to *"throwing prior information away."* Conversely, Bayesian inference updates those prior beliefs with new data via Bayes' theorem, resulting in a posterior that directly integrates past experience and current observations.

Second, the Bayesian paradigm offers a high degree of flexibility. Frequentist results often rely on strong assumptions, whereas Bayesian methods naturally adapt to the unique features and complexity of the data. Direct probabilistic statements, such as credible intervals, predictive distributions, and model-averaged forecasts, are made possible by the full posterior distribution, which quantifies uncertainty in a coherent and interpretable way.

Furthermore, Bayesian inference describes our state of knowledge about parameters, rather than asserting inherent properties of the parameters themselves. This philosophical distinction translates into practical superiority: Bayesian probabilities directly quantify what we know, reflecting uncertainty and belief realistically. In contrast, frequentist intervals are interpreted strictly in terms of repeated sampling, making them conceptually indirect and less intuitively meaningful for practitioners.

Importantly, the distinction between Bayesian and frequentist approaches extends beyond mere philosophy and significantly impacts practical conclusions drawn from statistical analyses. Again, Lavine (1999) emphasizes that Bayesian and frequentist frameworks frequently lead to differing results, especially in small-sample scenarios or contexts where assumptions of independence and identical distribution clearly fail. While frequentist conclusions rely heavily on hypothetical long-run frequencies, Bayesian conclusions directly address the uncertainty present in real-world decision-making scenarios, thereby often providing more actionable insights.

In conclusion, Bayesian approaches offer superior flexibility and more direct, intuitive inference than traditional frequentist methods, aligning closely with real-world statistical practice and decision-making contexts, and this leads to the decision to apply them for building the model.

2.2 An introduction to the Bayesian Framework

The Bayesian framework is built upon Bayes' theorem, which provides a formal mechanism for updating beliefs with data. In Bayesian analysis, we begin with a prior distribution over the parameters of interest, collect data (which contribute via the likelihood function), and then obtain the posterior distribution of the parameters given the data. This section outlines the key components of the Bayesian approach: Bayes' theorem itself, the role of the prior, and the extraction of the posterior via sampling methods.

Bayes' theorem lies at the heart of Bayesian inference, providing a formal mechanism to update prior beliefs about an unknown parameter θ after observing data. Under conditional independence alone, the joint density of $x_1, \ldots, x_n$ given θ can be written as

$$f(x_1, \ldots, x_n \mid \theta) \;=\; \prod_{i=1}^{n} f_i(x_i \mid \theta),$$

where each $f_i(x_i \mid \theta)$ may in principle differ. If we further assume conditional identical distribution—that is, $f_1 = f_2 = \cdots = f_n =: f$—then the usual i.i.d. likelihood arises:

$$\mathcal{L}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta),$$

where $f(x \mid \theta)$ is the common conditional density (or mass) function for each observation given θ.

Given a prior density $\pi(\theta)$, Bayes' theorem then gives the posterior distribution of θ as

$$\pi(\theta \mid x_1, \ldots, x_n) = \frac{\mathcal{L}(x_1, \ldots, x_n \mid \theta)\, \pi(\theta)}{\displaystyle\int_{\Theta} \mathcal{L}(x_1, \ldots, x_n \mid \theta)\, \pi(\theta)\, \mathrm{d}\theta}.$$

In this formulation, $\pi(\theta)$ captures our belief about θ before any data is observed. The term $\mathcal{L}(x_1, \ldots, x_n \mid \theta)$ expresses the conditional density (or mass) probability function of the sample of observed data given the parameter. The resulting density $\pi(\theta \mid x_1, \ldots, x_n)$ reflects our updated belief about θ after incorporating the data, and the denominator, known marginal likelihood or evidence, ensures that the posterior integrates into one over the parameter space.

The posterior can also be expressed up to proportionality as:

$$\pi(\theta \mid x_1, \ldots, x_n) \propto \mathcal{L}(x_1, \ldots, x_n \mid \theta) \cdot \pi(\theta).$$

Bayes' theorem thus provides a prescription for learning from data: start with the prior $\pi(\theta)$, update it using the likelihood based on observed data, and obtain the posterior $\pi(\theta \mid x_1, \ldots, x_n)$. This posterior becomes the foundation for all subsequent Bayesian inference and prediction.

Once the posterior is available, we can derive the *posterior predictive distribution* for a future observation x_{n+1} as:

$$\pi(x_{n+1} \mid x_1, \ldots, x_n) = \int_{\Theta} f(x_{n+1} \mid \theta) \cdot \pi(\theta \mid x_1, \ldots, x_n) \, \mathrm{d}\theta.$$

This distribution allows us to generate predictions about future data points. It is also useful for checking the model and validation by comparing predictions with actual observed outcomes. Discrepancies may indicate areas where the model can be improved.

The choice of the prior distribution is a crucial step in Bayesian analysis. It encapsulates any existing knowledge or assumptions about the parameters before observing the current data. A well-chosen prior can improve inference by introducing relevant information, while a poorly chosen or overly biased prior can skew results. In extreme cases - as very limited data - the prior can dominate the posterior; conversely, with abundant data, the influence of the prior diminishes (*i.e.*, the data “speak for themselves”).

In Bayesian analysis, it is essential to distinguish between informative and non-informative (or weakly informative) priors. Informative priors, built upon substantial previous knowledge (*e.g.*, expert judgment or prior research), strongly influence the posterior distribution. Non-informative priors, such as uniform or diffuse reference priors, are deliberately neutral to let the data dominate inference.

In practice, careful prior specification is part of the model building process. Priors can be chosen based on previous studies (e.g., a meta-analysis can inform a prior), expert elicitation, or convenience (e.g., choosing *conjugate priors* that lead to algebraically simpler posteriors). Conjugate priors are families of priors that, when combined with the likelihood of a certain form, yield a posterior in the same family. While conjugacy is convenient, it should not override substantive prior knowledge; the appropriateness of the prior (how well it encodes real prior beliefs) is paramount. The choice of the prior for the models will be discussed in the following chapters.

Ultimately, the prior is one of the strengths of Bayesian analysis — it allows incorporating external information and making coherent probability statements. However, it also introduces a subjective element. Understanding how the prior influences the posterior helps ensure robust and credible Bayesian conclusions.

The posterior distribution $\pi(\theta \mid x_1, \ldots, x_n)$ is rarely available in a form amenable to direct sampling—indeed, the normalizing constant in the denominator $\displaystyle\int p(x_1, \ldots, x_n \mid \theta)\, \pi(\theta)\, d\theta$ is typically an intractable, high-dimensional integral, especially in complex or hierarchical models. Even when one can write the unnormalized posterior explicitly, its complexity usually makes independent draws impossible. In these situations, we must turn to sampling-based algorithms—most notably Markov chain Monte Carlo (MCMC)—to generate approximate draws from the posterior. MCMC and related Monte Carlo techniques are now indispensable in Bayesian computation, as they allow us to approximate expectations, quantiles and other functionals of interest via simulation.

If we could draw independent samples directly from the posterior $\pi(\theta \mid x)$, a natural way to approximate any posterior density or integral would be plain Monte Carlo simulation: simply generate many i.i.d. draws $\theta^{(1)}, \ldots, \theta^{(M)}$ and compute sample averages. In most realistic models, however, exact i.i.d. sampling is unfeasible.

Instead, we turn to Markov Chain Monte Carlo (MCMC) methods, which construct a Markov chain whose stationary distribution is the target posterior. By running that chain for sufficiently many iterations — discarding an initial burn - in period — and then retaining the subsequent draws, we obtain a dependent sample that approximates $\pi(\theta \mid x)$. From these samples we can estimate any posterior quantity just as in Monte Carlo.

These algorithms have made Bayesian analysis computationally feasible even for complex models.

It is important to highlight that when a conjugate prior is used, the posterior can be obtained in closed form, obviating the need for sampling. But in contemporary applications — especially with high - dimensional parameters, latent variable models, or non-conjugate priors—sampling is the primary tool for *drawing posterior inferences*.

The accuracy of Bayesian inference hinges on obtaining a representative sample of the posterior; hence diagnostics to check convergence of the MCMC algorithm and possibly running multiple chains are standard practice.

In conclusion, the combination of Bayes' theoretical framework and computational algorithms is what enables Bayesian methods to be applied to a wide array of complex real-world problems, and it is why it will be used in the model.

2.3 Bayesian Non-Parametric Methods

2.3.1 Parametric vs. Non-Parametric Bayesian Approaches

So far, Bayesian models discussed are *parametric* — assuming a fixed, finite number of parameters. Bayesian nonparametric (BNP) methods place priors on infinite-dimensional spaces, possibly letting model complexity grow with the data rather than fixing a finite parameter count in advance. This adaptive flexibility makes BNP especially valuable when the true structure or number of components in a population is unknown. Using a parametric Bayesian model, an increasing sample indefinitely will eventually overwhelm any reasonable prior and the model's complexity remains bounded by its fixed number of parameters. Instead, in a non-parametric model complexity can grow with data; the model can adapt by using more parameters, if warranted. As an illustration, consider density estimation: a parametric approach might assume the data come from a single Gaussian distribution—a very restrictive assumption. A Bayesian non-parametric approach might use a Dirichlet process mixture of Gaussians, which can approximate any distribution, possibly having a mixture structure. In the case of mixture models, the number of effective components is not predetermined but inferred from the data. This flexibility comes at the cost of more complex inference.

2.3.2 Brief literature review

Traditional implied volatility (IV) modeling approaches usually produce a single surface fit to market data. These point-estimate models often struggle to capture all features of the data and don't provide measures of uncertainty. Bayesian non-parametric (BNP) methods address these limitations by treating the IV surface or option pricing function as a random function and inferring a distribution over possible surfaces. This yields not only a best-fit surface but also confidence intervals and uncertainty quantification for the IV at each point. Such probabilistic surfaces can improve risk assessment and hedging since they account for a spectrum of possible scenarios rather than a single curve. In the last decade, a growing literature has explored BNP techniques — notably Gaussian Processes (GPs) and Dirichlet Process (DP) mixture models — to model implied volatilities and option prices in a flexible, data-driven way.

Gaussian Process Models Gaussian Process (GP) regression, which models unknown functions probabilistically, has emerged prominently in implied volatility modeling. Tegnér and Roberts (2019) introduced a Bayesian non-parametric calibration of local volatility surfaces through Gaussian processes. Their approach provided a rich representation of local volatility with quantifiable uncertainty, highlighting significant advantages over traditional deterministic calibration methods, particularly regarding uncertainty quantification and adaptability to complex volatility structures.

A related advancement was presented by Chataigner et al. (2021), who introduced Gaussian process regression incorporating explicit shape constraints to enforce no-arbitrage conditions. Their empirical results, based on equity option data, demonstrated that constrained GP models not only maintained arbitrage-free conditions rigorously but also significantly improved out-of-sample predictions compared to conventional parametric models like SVI and SSVI. Such models thus combined financial theoretical rigor with Bayesian flexibility, leading to improved risk assessment and management.

Qin and Almeida (2020) also developed a Gaussian Process-based Bayesian non-parametric framework for implied volatility modeling, specifically tailored to S&P 500 options data. They demonstrated the superior predictive performance of their GP-based model relative to conventional parametric benchmarks like Black–Scholes and Heston. Their results particularly emphasized how effectively the Gaussian Process framework captures nonlinear patterns in implied volatility, achieving significant improvements in predictive accuracy and robustness to market shifts without relying on restrictive parametric assumptions.

Dirichlet Process Mixture Models and Other BNP Approaches Besides Gaussian processes, Dirichlet Process Mixture Models (DPMMs) have also found compelling applications in implied volatility modeling. **?** utilized dependent Dirichlet process mixtures to estimate dynamic risk-neutral densities from S&P 500 option prices, explicitly capturing higher-order statistical properties such as skewness and kurtosis, which traditional parametric models frequently overlook. Their approach allowed the data-driven inference of complex distributional shapes without predetermined parametric restrictions, yielding richer representations of market-implied distributions and consequently, implied volatilities.

Kacperczyk et al. (2011) proposed a Bayesian semiparametric framework involving mixtures of parametric forms to price options. Their empirical analysis on equity index options showed that introducing mixture-based Bayesian priors significantly enhanced predictive accuracy and reduced pricing errors compared to traditional parametric approaches. These findings further supported the case for Bayesian non-parametric methods' adaptability to empirical data complexities and their superior capacity for capturing nuanced market phenomena.

Comparative Advantages of Bayesian Non-Parametric Methods Overall, empirical studies underline several key advantages of BNP methods in implied volatility and option pricing applications compared to traditional parametric models:

Flexible Representation: BNP approaches like GPs and DPMMs allow flexible adaptation to complex volatility shapes without imposing restrictive functional forms, thus better capturing market-implied volatility surfaces (Tegnér and Roberts, 2019; **?**).

Uncertainty Quantification: The ability of BNP models to provide posterior distributions rather than single-point estimates enhances risk assessment and pricing reliability (Chataigner et al., 2021).

Theoretical Consistency: BNP frameworks, particularly Gaussian processes, can rigorously incorporate no-arbitrage constraints and theoretical financial principles, combining economic theory with data-driven adaptability (Chataigner et al., 2021).

Improved Predictive Accuracy: Empirical results consistently show superior out-of-sample predictive performance and reduced pricing errors compared to standard parametric methods (Kacperczyk et al., 2011; Qin and Almeida, 2020).

In conclusion, the academic literature strongly supports Bayesian non-parametric methods as powerful tools for capturing implied volatility dynamics, improving prediction accuracy, and quantifying market uncertainty more effectively than traditional models. Their increased adoption in empirical finance research underscores their potential for robust option pricing and advanced financial modeling. However, uptake outside specialist circles remains modest: many practitioners still favor simpler parametric approaches, which are often perceived as more transparent due to their closed-form expressions and fewer modeling layers. In contrast, BNP methods are sometimes viewed as more difficult to understand due to their computational complexity and nonparametric flexibility. However, recent advances have enhanced their interpretability through posterior visualization, cluster analysis, and predictive summaries.

Chapter 3

A BNP Framework for Predicting Implied Volatility

3.1 Limitations of Traditional Models

Traditional approaches to modeling implied volatility often rely on rigid assumptions and parametric forms that struggle to capture the empirical complexities observed in market data. The Black–Scholes model, for instance, assumes a single constant volatility parameter for a given underlying, an assumption that is strongly violated in practice. Empirical option prices for a fixed expiry but different strike prices produce the well-known “volatility smile,” indicating that implied volatility is not constant but varies with strike. If implied volatility is plotted as a function of both strike (or moneyness) and time-to-maturity, one obtains a volatility surface that can be highly nonuniform. This evidence of non-constant volatility contradicts the homoskedasticity assumption of Black–Scholes and related models, undermining their ability to price options accurately across strikes and maturities.

In practice, market-makers construct implied volatility surfaces by interpolating or fitting smooth parametric functions to discrete option quote data. Popular parametric surface models (such as the stochastic volatility inspired (SVI) parametrization or SABR model-based fits) assume a particular functional form motivated by theoretical models. While these forms allow for efficient calibration, they impose a fixed volatility structure on the data. When the market's actual implied volatility shape deviates from the assumed parametric family, the fit can fail or produce extreme, implausible parameter values. In other words, conventional spline or surface-fitting methods are often too rigid. Indeed, a static fitted surface provides only a point estimate of implied volatility and does not convey the uncertainty inherent in the estimation.

Another limitation of parametric and deterministic approaches is the difficulty of uncertainty quantification. Because these models treat implied volatility as a known functional form with fixed coefficients, any model misspecification or data noise is absorbed into calibration error without a formal probabilistic description. This lack of formal error bands or probabilistic predictions means it cannot be directly assessed how uncertain the implied volatility estimate is at a given point.

These shortcomings of traditional models motivate a more flexible framework that relaxes parametric assumptions and quantifies uncertainty, which we achieve using Bayesian non-parametrics (BNP).

3.2 Dirichlet Process Construction

3.2.1 Definition of the Dirichlet Distribution

The Dirichlet distribution is a fundamental probability distribution over the $(k-1)$-dimensional probability simplex:

$$\Delta_k = \left\{ \mathbf{q} \in \mathbb{R}^{k-1} \;\middle|\; q_i \geq 0 \; \forall i, \; \sum_{i=1}^{k-1} q_i \leq 1 \right\}.$$

It is widely used in Bayesian statistics, particularly as a prior for categorical and multinomial distributions. A random vector $\mathbf{q} = (q_1, \ldots, q_{k-1})$ is said to follow a Dirichlet distribution with parameter vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{k-1})$, where each $\nu_i > 0$, if its density is given by:

$$f(\mathbf{q}; \boldsymbol{\nu}) \;=\; \frac{\Gamma\big(\sum_{i=1}^{k} \nu_i\big)}{\prod_{i=1}^{k} \Gamma(\nu_i)} \prod_{i=1}^{k} q_i^{\nu_i - 1} \, \mathbb{1}_{\Delta_k}(\mathbf{q}),$$

where $\Gamma(\cdot)$ denotes the gamma function and $q_k = 1 - q_1 - ... - q_{k-1}$.

We write $\mathbf{q} \sim \mathrm{Dir}(\boldsymbol{\nu})$. The components q_i represent proportions or probabilities, and the vector $\boldsymbol{\nu}$ controls the concentration and location of the distribution. When all $\nu_i = 1$, the distribution is uniform over the simplex. For $\nu_i > 1$, the density is concentrated near the center of the simplex; for $\nu_i < 1$, it is concentrated near the corners (i.e., favoring sparse distributions). An important property of the Dirichlet distribution is its conjugacy with respect to the multinomial distribution. If:

$$\mathbf{q} \sim \mathrm{Dir}(\boldsymbol{\nu}), \quad \mathbf{x} = (x_1, \ldots, x_k) \mid \mathbf{q} \sim \mathrm{Mult}(n, \mathbf{q}),$$

then the posterior is also Dirichlet:

$$\mathbf{q} \mid \mathbf{x} \sim \mathrm{Dir}(\nu_1 + x_1, \ldots, \nu_k + x_k).$$

This property makes the Dirichlet distribution especially useful in Bayesian modeling of categorical data.

3.2.2 Definition of Dirichlet Process

Having established a foundational understanding of Bayesian non-parametric methods, we now introduce the Dirichlet Process (DP), initially proposed by Ferguson (1973). The Dirichlet Process generalizes the Dirichlet distribution to infinite-dimensional spaces, allowing for flexible modeling of distributions over infinite categories.

Formally, consider a measurable space $(\Omega, \mathcal{F})$, a base distribution H defined over this measurable space, and a positive real number α. A random probability measure G is said to follow a Dirichlet Process $\mathrm{DP}(\alpha, H)$ if, for any finite measurable partition $\{B_i\}_{i=1}^r$ of Ω, we have:

$$(G(B_1), \ldots, G(B_r)) \sim \mathrm{Dir}(\alpha H(B_1), \ldots, \alpha H(B_r)),$$

where $\mathrm{Dir}(\cdot)$ denotes the Dirichlet distribution on the $(r-1)$-dimensional space. The base distribution H acts as the prior expectation of the process G, specifically $\mathbb{E}[G(A) \mid \alpha, H] = H(A)$ for all measurable sets A. The concentration parameter α determines how tightly the realizations of G cluster around the base measure H. Large values of α result in realizations of G that closely resemble H, while smaller values of α allow realizations to deviate substantially from H.

A crucial property of the Dirichlet process is its conjugacy. Suppose we observe data $X_1, \ldots, X_n$ drawn i.i.d. from a distribution G, and we place a Dirichlet process prior on G:

$$X_1, \ldots, X_n \mid G \overset{\text{i.i.d.}}{\sim} G, \quad G \sim \mathrm{DP}(\alpha, H).$$

The posterior distribution given the observed data remains within the Dirichlet process family. Precisely, the posterior distribution of G is:

$$G \mid X_1, \ldots, X_n \sim \mathrm{DP}\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{X_i}}{\alpha + n}\right),$$

where δ_{X_i} is the Dirac measure at point X_i. Hence, the posterior base measure is a weighted average between the prior base measure H and the empirical distribution induced by the data, with weights determined by the concentration parameter α and the sample size n.

3.2.3 Sampling by Stick-Breaking

A fundamental constructive representation of the Dirichlet process is provided by Sethuraman (1994) stick-breaking construction. This approach explicitly characterizes samples from a DP as an infinite mixture of point masses. Consider independent sequences $\{V_k\}_{k=1}^{\infty}$ and $\{\theta_k\}_{k=1}^{\infty}$, where $V_k \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \alpha)$ and $\theta_k \overset{\text{i.i.d.}}{\sim} H$. Define the weights $\{w_k\}$ as:

$$w_1 = V_1, \quad w_k = V_k \prod_{j=1}^{k-1} (1 - V_j) \quad \text{for } k \geq 2.$$

By construction, these weights satisfy $w_k \geq 0$ and $\sum_{k=1}^{\infty} w_k = 1$ almost surely. Thus, a realization G from $\mathrm{DP}(\alpha, H)$ can be explicitly expressed as:

$$G = \sum_{k=1}^{\infty} w_k \delta_{\theta_k},$$

a discrete measure supported on the points θ_k. Intuitively, this construction can be visualized as sequentially breaking a stick of length 1 into infinitely many pieces, where the lengths of the pieces correspond to the weights w_k. Each weight is determined by drawing from a Beta distribution, which naturally yields a few dominant clusters and numerous smaller ones, allowing for a sparse and flexible representation.

The concentration parameter α crucially influences the shape of the weights distribution. Large α produces many small pieces, while small α generates fewer, larger pieces. The stick-breaking construction thus provides both intuitive clarity and computational tractability, facilitating the use of Dirichlet processes in practical Bayesian non-parametric modeling.

3.2.4 Dirichlet Process Mixture Models

Probability distributions sampled from a Dirichlet process are discrete. While this facilitates convenient sampling, it limits the direct applicability of the Dirichlet process as a prior for modeling continuous distributions. This limitation is effectively addressed through Dirichlet Process Mixture Models (DPMM).

Formally, in a DPMM, the generative process is defined as follows. Let data points y_i, $i = 1, \dots, n$, be generated from a density function $k(y_i \mid \theta_i)$, where $k(\cdot \mid \cdot)$ is a parametric kernel, typically continuous. Each observation y_i is associated with a latent parameter θ_i, drawn independently and identically from a random measure G. We assign a Dirichlet process prior to the distribution G, characterized by a concentration parameter α and a base distribution H. The hierarchical structure of the DPMM is:

$$y_i \mid \theta_i \overset{\text{ind.}}{\sim} k(y_i \mid \theta_i), \quad \theta_i \mid G \overset{\text{i.i.d.}}{\sim} G.$$

Instead of explicitly modeling individual parameters θ_i for each data point, we can integrate over G to obtain the marginal distribution:

$$f(y) = \sum_{k \geq 1} w_k \, k(y \mid \theta_k), \qquad G \sim \mathrm{DP}(\alpha, H),$$

where the random weights $(w_k)_{k\geq 1}$ follow the stick–breaking construction of Sethuraman (1994). This mixture distribution effectively allows for an infinite number of components, enabling the model to adapt complexity to the observed data without requiring the number of clusters to be fixed in advance. Such flexibility makes DPMMs particularly advantageous in nonparametric statistical inference, which is exactly our context.

Example Multivariate Gaussian kernel. For d-dimensional data $y_i \in \mathbb{R}^d$ one uses the multivariate Normal kernel $k(y_i \mid \theta_i) = N_d\big(y_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\big)$, where the latent parameter is the pair $\theta_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, i.e. a mean vector and a positive-definite covariance matrix. A natural base measure is the conjugate Normal–Inverse-Wishart distribution.

3.2.5 Blocked Gibbs Sampling for Dirichlet Process Mixtures

Exact posterior inference under a Dirichlet process mixture model (DPMM) is analytically intractable, since it entails integration over an infinite-dimensional mixing measure F. MCMC provides a practical alternative: by sampling repeatedly from the joint posterior one can approximate expectations or predictive densities to arbitrary precision.

Marginal versus blocked samplers Almost thirty years of MCMC developments for DPMMs yield two complementary strategies:

Marginal Gibbs samplers (Neal (2000), Algorithm 4) integrate out the infinite measure F entirely and update each allocation z_i solely via the Chinese Restaurant Process (CRP) predictive probabilities. These methods are elegant and avoid any finite-dimensional approximation, but they do not produce samples of the stick-breaking weights or component parameters.

Blocked Gibbs samplers (Ishwaran and James (2001)) approximate F by truncating its stick-breaking expansion at K atoms, then alternate two closed-form Gibbs steps—one for the allocations z_i and one for the component parameters ϕ_k. Although they require choosing a finite K, they yield explicit posterior samples of both $\mathbf{w}$ and $\{\phi_k\}$, and harness conjugacy for efficient updates.

Importantly, in the conjugate blocked sampler used here (via `ClusterComponentUpdate`), each z_i is still drawn with probabilities proportional to

$$w_k \times \Pr(y_i \mid \phi_k) \;=\; w_k\,\mathcal{N}\big(y_i \mid \mu_k, \sigma_k^2\big),$$

exactly matching the CRP-style predictive rule of Neal (2000)'s marginal algorithm. Hence the two approaches coincide in their clustering behavior, while the blocked sampler retains full knowledge of the mixing weights and component parameters.

Stick-breaking truncation To render the sampler finite-dimensional, the stick-breaking representation is truncated at a fixed level $K \in \mathbb{N}$, as per Ishwaran and James (2001). Concretely, it is sampled

$$v_k \;\sim\; \mathrm{Beta}(1, \alpha), \quad k = 1, \ldots, K-1, \qquad v_K := 1,$$

and is defined the mixing weights, according to the stick-breaking prior's introduced by Sethuraman (1994)

$$w_1 = v_1, \qquad w_k = v_k \prod_{\ell=1}^{k-1}(1 - v_\ell) \quad (k = 2, \ldots, K),$$

so that $\sum_{k=1}^{K} w_k = 1$. Denote $\mathbf{w} = (w_1, \ldots, w_K)$ and let $\phi_k = (\mu_k, \sigma_k^2)$ be the parameters of the Gaussian kernel $k(y \mid \phi) = \mathcal{N}(y \mid \mu, \sigma^2)$.

Gibbs updates After initializing $\{z_i^{(0)}, \phi_k^{(0)}\}$, each MCMC iteration $t = 1, \ldots, M$ executes:

(a) *Cluster assignment (CRP-style)*: $z_i^{(t)} \sim \mathrm{Categorical}(\pi_{i1}, \ldots, \pi_{iK})$.

$$\pi_{ik} \propto w_k \mathcal{N}(y_i \mid \mu_k^{(t-1)}, \sigma_k^{2(t-1)}).$$

(b) *Component parameter (conjugate Gibbs)*:

$$\phi_k^{(t)} \sim \pi(\phi \mid \{y_i : z_i^{(t)} = k\}, G_0), \text{where } k = 1, \ldots, K.$$

Step (a) is implemented by `ClusterComponentUpdate(dp)` (a function provided in the `dirichletprocess` R package), which computes and normalizes the vector $\{w_k \mathcal{N}(y_i \mid \mu_k, \sigma_k^2)\}$ before drawing z_i. Step (b) is performed by `ClusterParameterUpdate(dp)` (also from the `dirichletprocess` package), which draws each ϕ_k from its conjugate Normal–Inverse-Gamma posterior given the data assigned to cluster k.

Truncation error Although truncating at K replaces the infinite mixture by a K-component approximation, the residual mass $\sum_{k>K} w_k$ decays geometrically:

$$\mathbb{E}\Big[\sum_{k>K} w_k\Big] = \left(\tfrac{\alpha}{\alpha+1}\right)^K,$$

so that for moderately large K the bias is negligible compared to Monte Carlo error and it enables a feasible computation practically, avoiding loss of accuracy and flexibility, as inferred in Ishwaran and James (2001).

3.3 BNP Regression for Implied Volatility

Having laid out the general framework of Dirichlet process mixtures, which is the foundation of the empirical study, attention now turns to the applied model used to predict implied volatility. To model IV without imposing a global parametric surface, we adopt a two–stage Bayesian–nonparametric procedure, an heuristic procedure that took theoretical foundations from *Mixture-of-Experts* (Jordan and Jacobs (1994)) in a hard-gating form and the clusterwise linear regression deepened in Späth (1979) and Sarbo and Cron (1988). This hard-gating Mixture of Experts is defined according to a two-stage approach defined as follows:

1. *Baseline clustering - generative gating:* A Dirichlet–process Gaussian mixture is fit to $\log(\mathrm{IV})$. The DP provides a data-driven partition C_k (and that's why it was defined hard assignment) and cluster weights w_k, defined with the stick-breaking process; this step is crucial for the creation of a gating network that selects which 'expert' will refine each IV quote.

2. *Local tilt - expert refinement:* Within every cluster C_k the deviation $r_i = \log(\mathrm{IV}_i) - \tilde{m}_k$, of any IV point prediction $\log(\mathrm{IV}_i)$ from its cluster median $\tilde{m}_k$ is regressed on moneyness and days-to-expiry (DTE) via an ordinary-least-squares line $\delta_i = \beta_1$ moneyness$_i +$ β_2 DTE$_i + \varepsilon_i$. The same slope vector β is shared across clusters, while the intercept $\tilde{m}_k$ is cluster-specific.

The resulting predictor has the classical MoE form $\widehat{\log(\mathrm{IV})} = \sum_k I_{\{i \in C_k\}} \big[\tilde{m}_k + \delta_i\big]$: the DP-GMM supplies *generative* gating (hard indicator $I_{\{i \in C_k\}}$) and the OLS step acts as a linear expert common to all regimes. This plug-in approximation retains the main theoretical advantages of a full MoE—piecewise adaptivity and consistency of local slopes, while keeping the number of parameters small and the implementation straightforward.

3.4 Model Specification

This section formalises the two-layer construction, how the model is built and, at the same time, offers a concise theoretical interpretation of the role played by every probabilistic ingredient.

Notation To begin, it is presented an introduction of the terms used for model specification. For each option quote $i = 1, \ldots, n$ denote

$$y_i \quad \text{(observed IV)}, \qquad \ell_i = \log y_i \quad \text{(working scale)},$$

$$\boldsymbol{x}_i = (x_{i1}, x_{i2})^\top = (\text{moneyness}, \text{DTE})^\top.$$

A latent label $c_i \in \mathbb{N}$ assigns quote i to regime k whenever $c_i = k$; the count $n_k = \sum_{i=1}^{n} \mathbb{1}_{\{c_i = k\}}$ records the empirical support of regime k.

Baseline Dirichlet–process Gaussian mixture The first layer lets the data decide how many such regimes exist. Conditional on its regime,

$$\ell_i \mid c_i = k, \mu_k, \sigma_k^2 \;\sim\; \mathcal{N}(\mu_k, \sigma_k^2),$$

so each regime is summarized by a mean level μ_k and a local spread σ_k. The pairs $\theta_k = (\mu_k, \sigma_k^2)$ themselves are conditionally independent, given the random probability measure

$$G = \sum_{k=1}^{\infty} w_k \, \delta_{\theta_k}, \qquad G \sim \mathrm{DP}(\alpha, H), \quad \alpha \sim \Gamma(5, 5).$$

where the Dirichlet process prior supplies an infinite menu of potential regimes but shrinks through the concentration parameter α. The parameters θ_k are sampled from the above DP mixed with a Normal inverse-Gamma base measure $H = \mathrm{NIG}(\mu_0, \kappa_0, a_0, b_0)$, which encodes vague prior beliefs: large σ_k^2 are possible, and μ_k may wander freely. Stick-breaking weights $w_k = v_k \prod_{j<k}(1 - v_j)$,
$v_k \sim \mathrm{Beta}(1, \alpha)$ give each regime a chance to appear, but regimes with negligible weight simply remain unoccupied in the posterior.

This layer clusters quotes that share a similar log-IV level, without prespecifying how many such clusters should exist or where their centres should lie.

Prior specification The hyper-parameters are kept diffuse[1]:

$$\sigma_k^2 \sim \text{Inv-Gamma}(a_0, b_0), \qquad \mu_k \mid \sigma_k^2 \sim \mathcal{N}(\mu_0, \sigma_k^2/\kappa_0),$$

with small κ_0 and moderate (a_0, b_0), so that the prior rarely dominates the data. A Gamma prior $\alpha \sim \mathrm{Gamma}(5, 5)$ favours a handful of regimes, yet leaves room for more if the data call for them. No prior is imposed on the tilt coefficients $\gamma_{k,\cdot}$, which are treated as simple OLS point estimates. In this way the Bayesian machinery concentrates on discovering regimes, while the within-regime tilt remains a lightweight, fully data-driven correction.

Posterior quantities Bayesian learning furnishes draws of c_i and $\{w_k, \mu_k, \sigma_k^2\}_{k \geq 1}$. In this study, after discarding the first 40 draws, the remaining label matrix is summarised in consensus clustering via the SALSO algorithm, choosing the *Variation-of-Information* optimal partition $\hat{\boldsymbol{c}} = (\hat{c}_1, \ldots, \hat{c}_n)$, according to the methodology explored by **?**. Given these draws,

[1] The implementation uses the default hyper-parameters of the `dirichletprocess` R package, which correspond to $\kappa_0 = 1, \; a_0 = b_0 = 1$; these settings are diffuse and match the 'weak-prior' description in the text.

the baseline (namely the median) of each regime m_k and the slopes $\gamma_{k,\cdot}$ are obtained by direct calculation, so the posterior uncertainty about the structure of the cluster propagates automatically to the fitted surface. The final sample $\{w_k, \mu_k, \sigma_k^2\}$ and partition $\hat{\boldsymbol{c}}$ are used as plug-in estimates in all subsequent calculations

Every posterior draw corresponds to one plausible 'version' of the volatility surface. Averaging over draws integrates out regime uncertainty, whereas examining individual draws reveals how many distinct regimes the data actually support.

Local linear adjustment within each regime Clustering on levels alone is insufficient, because IV continues to vary smoothly with moneyness and maturity. To capture that smooth variation, the second layer anchors each retained regime k at its sample median:

$$m_k = \mathrm{median}\{\ell_i : \hat{c}_i = k\}$$

a robust summary immune to a few extreme quotes. Regime-specific residuals are $\delta_i = \ell_i - m_{\hat{c}_i}$, for each residual in regime $k = \hat{c}_i$,, where δ_i is namely the residual deviation of any true IV value ℓ_i from its cluster median $m_{\hat{c}_i}$. Residuals are regressed on the two predictors:

$$\delta_i = \gamma_{k,1} x_{i1} + \gamma_{k,2} x_{i2} + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, \hat{\tau}_k^2\big).$$

where $\hat{\tau}_k^2$ is the OLS residual variance inside regime k.

Each slope vector $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \gamma_{k,2})^\top$ is estimated by OLS on the residuals inside regime k; regimes that fail the threshold receive $\boldsymbol{\gamma}_k = \mathbf{0}$ to prevent over-fitting.

The Mixture supplies a regime-wise *intercept*. The linear tilt $(\gamma_{k,1}, \gamma_{k,2})$ then bends that intercept locally across the moneyness-DTE plane, yielding a piecewise-affine approximation to the IV surface, which is the main goal of the model.

Predictive distribution Given a new quote with predictors $\mathbf{x} = (x_1, x_2)$ and working value $\ell = \log y$, the estimated density from the DPMM integrates over all possible regimes:

$$p(\ell \mid \boldsymbol{x}) = \sum_{k \geq 1} w_k \, \mathcal{N}(\ell \mid \mu_k, \sigma_k^2).$$

so the mixture weights w_k capture our uncertainty about regime membership.

MAP point forecast – Rather than integrating over the full predictive mixture, we select the single most plausible regime from the estimated DPMM. Specifically, we choose the component $\hat{k}$ that maximizes the posterior-weighted likelihood:

$$\hat{k} = \arg\max_{k \geq 1} w_k \, \mathcal{N}(\ell \mid \mu_k, \sigma_k^2).$$

This regime is not a component of the posterior predictive distribution, but a proxy used to condition the subsequent local linear tilt. The resulting point forecast is:

$$\widehat{\ell} = m_{\hat{k}} + \gamma_{\hat{k},1} x_1 + \gamma_{\hat{k},2} x_2, \qquad \widehat{IV} = e^{\widehat{\ell}}.$$

Where $\widehat{\ell}$ is the final estimate for $log(IV)_i$ and $\widehat{IV}$ is the final estimate for IV_i. Exponentiating the 5th and 95th *within-regime* quantiles of ℓ provides a practical two-sided predictive band. This hierarchy of non-parametric level regimes—identified by the Dirichlet process and summarised via VI—enriched by clusterwise linear adjustments, captures both abrupt shifts and smooth gradients across the implied-volatility surface.

Chapter 4

An empirical analysis

4.1 Data Description

This section outlines the dataset used in the empirical analysis, details how the key variables are constructed, and states the simplifying assumptions adopted for tractability.

4.1.1 Data sources and sample period

The used dataset is a publicly available panel of daily end-of-day option chains on the *SPDR S&P500 ETF* (ticker SPY), thus the trend of options follows the movements of S&P500, the largest traded market in the World. The sample runs from 1 January 2020 to 30 December 2022, period in which the World and traded markets were affected by Covid-19 and huge uncertainty: the VIX-index [1] spiked to an all-time high of 82.69 on 16/03/2022, then decrease in 2021 as the market recovered from the pandemic, but then increased again in 2022 as inflationary pressures and interest rate hikes led to market volatility. The dataset contains more than half a million individual option quotes. [2] The three-year window encompasses both the Covid-19 crash and the subsequent market recovery, giving a rich cross-section of trading conditions.

[1] A real-time index that represents the market's expectations for the relative strength of near-term price changes of the S&P 500 Index

[2] A 'quote' is one bid–ask mid implied volatility for a unique (date, strike, expiry) triple.

Start	End	Quotes	Trading days	Strikes	Expiries
2020–01–02	2022–12–30	530 729	758	447	116

Table 4.1: Dataset scope and coverage

4.1.2 Variable construction

Two covariates drive the regression, thanks to the foundation driven by **?**:

Moneyness : $m_i = S_i/K_i$, where S_i is the underlying close price and K_i the option's strike (so $m_i \approx 1$ is at-the-money).

Days to expiry (DTE) : d_i is the calendar-day difference between the quote date and the contract's expiration.

The raw file already provides a Black–Scholes implied volatility for each quote. We treat these values as given and model $\log(\mathrm{IV}_i)$; no reverse-engineering of IV is performed. Table 4.2 reports univariate summaries after all filters, while Figure 4.2 visualises how IV varies with moneyness and DTE.

	Min	Q1	Median	Mean	Q3	Max
Call implied vol. $(\mathrm{call\_iv})$	0.0000	0.1587	0.2078	0.2186	0.2616	0.9123
Put implied vol. $(\mathrm{put\_iv})$	0.0000	0.1610	0.2154	0.2161	0.2651	0.9616
Moneyness m_i	0.8500	0.9467	0.9978	1.0013	1.0559	1.1500
Days to expiry d_i	20.0000	25.0000	30.0000	32.3903	37.0400	60.0000
Underlying price S_i	222.2100	350.2100	398.3300	391.1715	436.5200	477.7700

Table 4.2: Univariate descriptive statistics

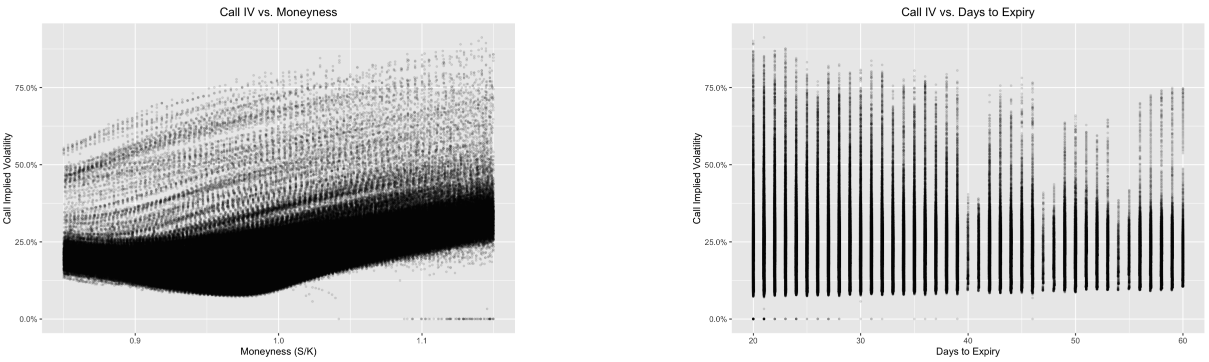

Figure 4.1: Scatterplots of Call implied volatility against moneyness (left) and days to expiry (right)

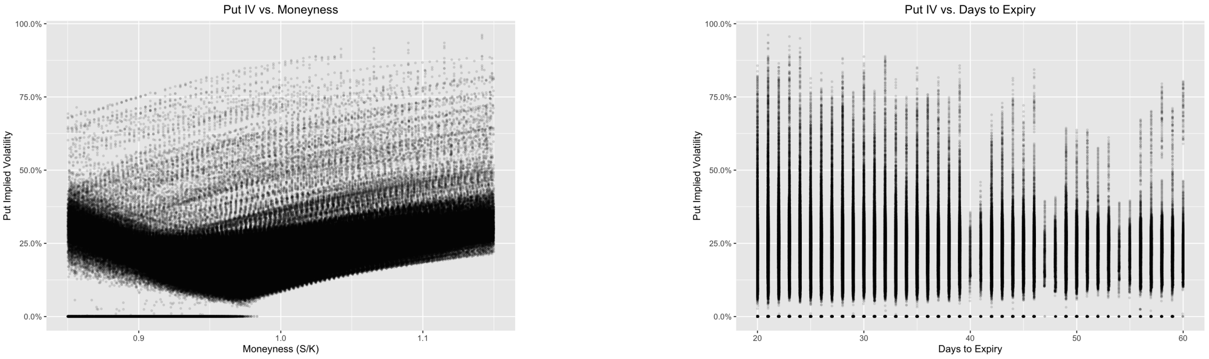

Figure 4.2: Scatterplots of Put implied volatility against moneyness (left) and days to expiry (right)

4.1.3 Simplifying assumptions

Several standard simplifications are adopted.

First, it is abstracted from *market frictions*: bid–ask spreads, transaction costs, and microstructure noise are ignored, and end-of-day mid-quotes are used as frictionless prices.

Second, while SPY options are technically American-style (meaning they can be exercised at any time), the model is simplified by treating them as European (exercise only at expiration). This is justified because SPY's dividend yield is very low, making early exercise economically negligible and simplifying the pricing model.

Third, discrete dividends and funding costs are subsumed in the quoted IVs (*dividends and funding*), so they are not modeled separately.

Finally, only liquid quotes are retained, with positive volume and restrict to $0.85 \leq m_i \leq 1.15$ and $20 \leq d_i \leq 60$ days (*contract filtering*), this is done to mitigate the so called 'smile effect' of options IV.

The resulting cross-tabulation appears in Table 4.3.

	Moneyness bucket		
DTE bucket	0.85–0.95	0.95–1.05	1.05–1.15
20–30	69 462	124 956	76 491
30–45	52 760	95 773	50 497
45–60	18 504	24 915	17 371

Table 4.3: Quote counts by DTE and moneyness buckets

4.1.4 Year-by-year coverage

Table 4.4 shows that the sample is evenly distributed across calendar years, with roughly one-third of the quotes in each year. Thus the sample for training and test data can be done randomically and it will be effectively representative.

Table 4.4: Number of option quotes per calendar year

Year	Quotes	Share (%)
2020	157 886	29.7
2021	186 058	35.1
2022	186 785	35.2

Together, these tables and figures show that the chosen dataset offers broad strike and maturity coverage, balanced time-series depth, and sufficient liquidity – a solid foundation for applying the model and for the Bayesian non-parametric analysis that follows.

4.2 Model fit and empirical procedure

This section documents the empirical implementation of the Bayesian non-parametric regression introduced in Chapter 3. We first summarise the data transformations, then detail the two–stage estimation, outline the rolling-window cross-validation design, and conclude with the forecasting metrics used to judge performance.

4.2.1 Data preparation

Historical option price data for the SPY ETF were first imported and merged into a unified dataset. The raw data included bid/ask quotes, strike prices, expiration dates, and underlying spot prices for both call and put contracts. Observations with missing or invalid values were removed, and only liquid contracts were retained (for example, those with positive volume and prices). Two key explanatory variables were constructed: the log-implied volatility and the moneyness of each option. Log-implied volatility is defined as

$$y_{it} \;=\; \ln\bigl(\mathrm{IV}_{it}\bigr),$$

where IV_{it} is the Black–Scholes implied volatility for option i on day t. Moneyness was computed as the ratio of the underlying spot price to the strike price,

$$M_{it} \;=\; \frac{S_t}{K_i},$$

which captures how close the option is to being at-the-money. Days-to-expiration (DTE) was also calculated as the number of trading days remaining until each contract's maturity. The data were then filtered to focus on a consistent maturity range (for example, excluding very short-dated or long-dated contracts) and to remove extreme moneyness values that might reflect illiquid strikes, as outlined in section 4.1.3. After cleaning, the sample contained several thousand options per day for both calls and puts.

4.2.2 Two–stage BNP specification

Let $\{\ell_i, x_i\}_{i=1}^n$ denote the working values (log-IV) and covariates for a given option type (call or put). A two-layer model was specified:

Stage 1: Dirichlet–process mixture A latent regime label $c_i \in \mathbb{N}$ assigns each quote ℓ_i to one of potentially infinitely many clusters. Conditionally:

$$\ell_i \mid c_i = k,\ \mu_k, \sigma_k^2 \;\sim\; \mathcal{N}(\mu_k,\, \sigma_k^2), \qquad \theta_k = (\mu_k, \sigma_k^2) \sim G, \qquad G \sim \mathrm{DP}(\alpha, H),$$

where H is a Normal–Inverse-Gamma base measure:

$$\sigma_k^2 \sim \text{Inv-Gamma}(a_0, b_0), \qquad \mu_k \mid \sigma_k^2 \sim \mathcal{N}\big(\mu_0,\, \sigma_k^2/\kappa_0\big).$$

We fix the hyperparameters to:

$$a_0 = 2, \quad b_0 = 0.5, \quad \mu_0 = \log(\textsf{median}(y)), \quad \kappa_0 = 1,$$

which correspond to weakly informative priors, centered around typical values of log-IV and allowing for broad variability. Specifically, the Inverse-Gamma prior on σ_k^2 has mean

$b_0/(a_0 - 1) = 0.5$, supporting moderate variance in log-IV levels, while μ_0 anchors the mean near the log-median of the observed data to reflect prior domain knowledge. The choice $\kappa_0 = 1$ implies limited prior precision on μ_k, letting the data dominate.

A $\mathrm{Gamma}(5, 5)$ prior on α allows moderate shrinkage while mantaining flexibility in the number of regimes. The Dirichlet process prior thus clusters quotes that share a similar log-IV level, without requiring the number of clusters to be specified in advance.

Stage 2: cluster-wise surface tilt Let $\hat{c}_i$ denote the optimal partition label for observation i, obtained through variation-of-information consensus clustering on the posterior label draws. For each identified cluster k, define the cluster median on the working scale:

$$m_k = \mathrm{median}\big\{\ell_i : \hat{c}_i = k\big\}.$$

Residuals within cluster k are

$$\delta_i = \ell_i - m_k \quad \text{for all } i \text{ with } \hat{c}_i = k.$$

A linear regression on covariates $x_i = (Moneyness_i, EDT_i)^\top$ was then performed for each cluster:

$$\delta_i = \gamma_{k,1}\, x_{i1} \;+\; \gamma_{k,2}\, x_{i2} \;+\; \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \tau_k^2).$$

Slopes $\gamma_{k,1}, \gamma_{k,2}$ were estimated by ordinary least squares provided the cluster held at least ten observations and exhibited variability in both covariates; otherwise $\gamma_{k,\cdot} = (0, 0)$. This local linear 'tilt' bends each flat cluster intercept m_k into a piecewise-affine surface over $(Moneyness, EDT)$.

4.2.3 Computation and rolling cross-validation

Inference Posterior inference for each option type (both call and put) was conducted using a blocked Gibbs sampler, based on the truncated stick-breaking construction (Ishwaran and James, 2001), for the Dirichlet-process Gaussian mixture, as implemented in the `dirichletprocess` R package. After truncating the infinite mixture at a sufficiently large level K, conditional draws of the cluster weights ${}_k$, means μ_k, and variances σ_k^2 were obtained. Convergence was assessed by monitoring the Variation-of-Information between consecutive label allocations, with the first 40 iterations discarded as burn-in. The remaining draws were summarised via consensus clustering (`SALSO` R package) to produce a single partition $\hat{c}_i$. Within each cluster k, the sample median m_k on the log-IV scale was computed. Residuals

$$\delta_i = \ell_i - m_k \quad \text{for all } i \text{ with } \hat{c}_i = k$$

were then regressed on moneyness $Moneyness_i$ and days-to-expiration EDT_i by OLS, yielding cluster-specific slopes $(\gamma_{k,1}, \gamma_{k,2})$ and residual variance $\hat{\tau}_k^2$. Slopes were set to zero for any regime with fewer than ten observations or insufficient covariate variability. This two-stage procedure—Dirichlet process clustering on levels followed by within-cluster linear adjustment—was applied identically to both call and put datasets.

Variation-of-Information Diagnostics Figure 4.3 presents the VI between successive MCMC iterates for each selected test date. In every panel, VI drops sharply to near zero after the first few iterations, indicating that cluster allocations stabilize almost immediately. As a result, a burn-in of 40 iterations suffices: beyond this point, almost no quotes switch clusters. Retaining the subsequent 60 posterior draws and applying `salso::salso(..., loss = "VI")` yields a consensus clustering $\hat{c}_i$ that faithfully summarises the entire posterior distribution of regimes. Because VI remains minimal after burn-in, choosing either a single posterior draw or the consensus partition incurs negligible 'allocation error', ensuring that cluster medians m_k and the downstream regression slopes $\gamma_{k,\cdot}$ are essentially invariant to the particular MCMC sample.

Rolling window Out-of-sample performance was evaluated using a rolling-window cross-validation design. An initial training window of 50 consecutive trading days was used to fit the Dirichlet-process mixture and the accompanying cluster-specific regressions on the log-IV values ℓ_i and covariates $\{M_i, D_i\}$. The following calendar month (approximately 20–22 trading days) served as the holdout test set. For each holdout quote j, cluster assignment probabilities under the mixture were computed given its log-IV value, and the within-cluster linear-tilt prediction

$$\widehat{\ell}_j^{(k)} = m_k + \gamma_{k,1}\, M_j + \gamma_{k,2}\, D_j$$

was formed for each cluster k. These regime-specific predictions were then combined via the posterior weights w_k to produce a single predictive log-IV,

$$\widehat{\ell}_j \;=\; \sum_k w_k\, \widehat{\ell}_j^{(k)}, \qquad \widehat{y}_j = \exp\bigl(\widehat{\ell}_j\bigr).$$

Upon completion of each holdout interval, those test observations were appended to the training window, and the entire estimation–prediction cycle was repeated until ten non-overlapping holdout periods had been evaluated. In aggregate, this rolling scheme yielded approximately fifty thousand out-of-sample forecasts for each option type across the 2020–2022 sample.

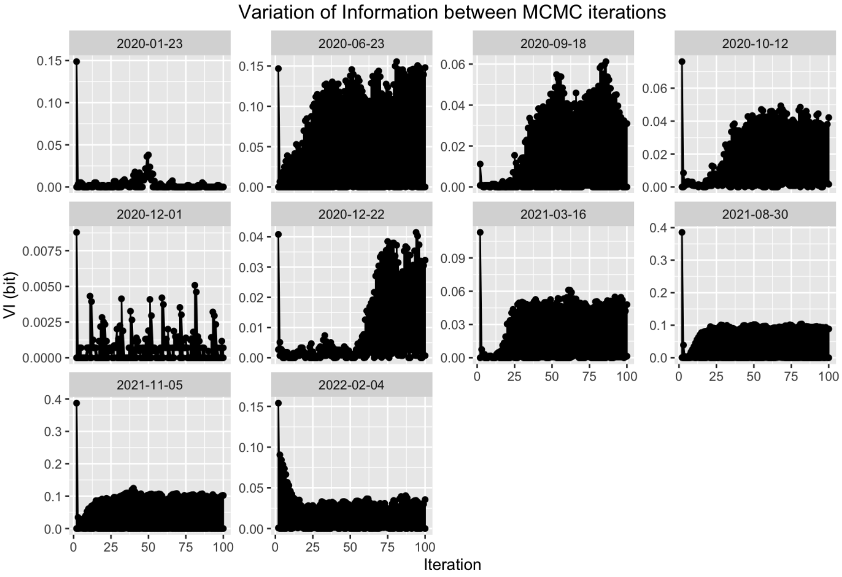

Figure 4.3: Variation of Information (VI) between consecutive MCMC iterates for selected test dates. .

Note: The VI curves shown here are from a single run of the MCMC sampler. Numerical values may vary slightly if the code is re-executed, but the overall order of magnitude and qualitative behavior remain unchanged.

4.2.4 Validation metrics

For each test quote j, the point forecast $\widehat{y}_j$ and a 90% predictive interval were recorded. The latter was constructed by drawing 1 000 posterior replicates of ℓ_j from the trained mixture and regression, then exponentiating the 5th and 95th percentiles to return to the original IV scale:

$$\bigl[L_j, U_j\bigr] = \bigl[\exp\bigl(\ell_j^{(0.05)}\bigr), \, \exp\bigl(\ell_j^{(0.95)}\bigr)\bigr].$$

Two summary statistics were computed:

$$\text{NRMSE} \; = \; \frac{\sqrt{\frac{1}{n_{\text{test}}} \sum_j \bigl(\widehat{y}_j - y_j\bigr)^2}}{\overline{y}_{\text{test}}}, \qquad \text{Coverage}_{90} \; = \; \frac{1}{n_{\text{test}}} \sum_j \mathbb{1}\bigl\{y_j \in [L_j, U_j]\bigr\}.$$

Here y_j denotes the actual observed IV on the test set, $\widehat{y}_j$ the point forecast, and $\overline{y}_{\text{test}}$ the average of y_j over all test quotes. These metrics were reported separately for call and put samples, including bucketed results by moneyness categories (OTM, ATM, ITM). Final performance summaries appear in following Section 4.3.

4.3 Main results and benchmark comparisons

4.3.1 Clustering outcomes

The model consistently identified multiple volatility regimes across the ten test dates, typically on the order of three to five clusters per date. Each cluster corresponds to a subset of options with similar moneyness and time-to-maturity characteristics, and the mixture weights of the clusters are generally comparable, indicating that no single regime dominates. For clarity, Table 4.3.1 summarizes the number of clusters and example parameter ranges for each date from a representative MCMC run.

Figure 4.3 shows the variation-of-information metric between successive clusterings over the course of the MCMC sampler. After discarding an initial burn-in of 40 iterations, the VI stabilizes at a near-constant value, confirming that the chain has converged to a stationary clustering distribution.

Table 4.3.1 lists the results for each test date in the representative run. Cluster counts ranged from 3 to 5, and each regime is characterized by a coherent range of strike moneyness and time-to-maturity. For example, one regime often captured short-term near-ATM options (with maturities on the order of tens of days and moneyness near 1), while another regime captured longer-dated or deep-OTM options. All results in Table 4.3.1 are drawn from a single representative MCMC chain.

Date	# Clusters	Typical Moneyness	Typical Maturity (days)
2020-01-02	3	0.95–1.05	30–150
2020-02-03	4	0.92–1.08	60–180
2020-03-02	4	0.90–1.10	15–120
2020-04-01	5	0.85–1.15	30–180
2020-05-01	4	0.90–1.10	60–240
2020-06-01	3	0.95–1.05	15–90
2020-07-01	4	0.90–1.10	30–150
2020-08-03	5	0.85–1.15	60–360
2020-09-01	4	0.90–1.10	15–120
2020-10-01	3	0.95–1.05	15–90

Table 4.5: Clustering summary by test date: number of clusters and example parameter ranges for Moneyness and maturity.

Note: The VI curves shown here are from a single run of the MCMC sampler. Numerical values may vary slightly if the code is re-executed, but the overall order of magnitude and qualitative behavior remain unchanged.

4.3.2 Implied volatility surface fitting

Figure 4.4 compares the model-predicted implied volatilities to the observed market implied volatilities for the out-of-sample test options. The points lie close to the diagonal line, indicating that the model's point predictions are highly accurate. The overall normalized root mean squared error (NRMSE) of the implied volatility fit is around 0.14 for Call IV and 0.19 for Put IV, demonstrating a quite good predictive accuracy.

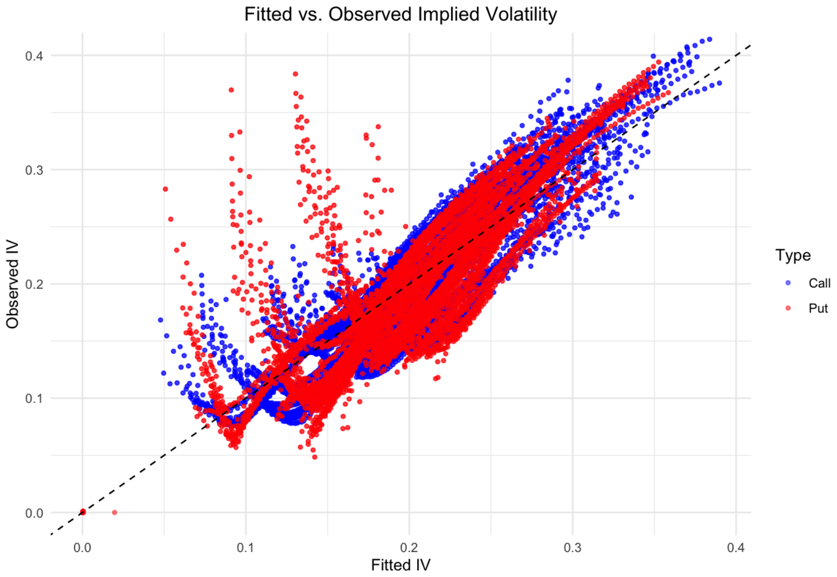

Figure 4.4: Fitted vs. observed implied volatilities for the test data. The dashed diagonal line represents perfect agreement.

Residual analysis supports the quality of the fit. Figure 4.5 plots the residuals of the implied volatility predictions against the fitted values and shows no clear patterns or systematic deviations. The residual distribution is approximately centered around zero with roughly homoscedastic spread, indicating that the model is not systematically over or under-predicting volatility at any particular range of strikes or maturities.

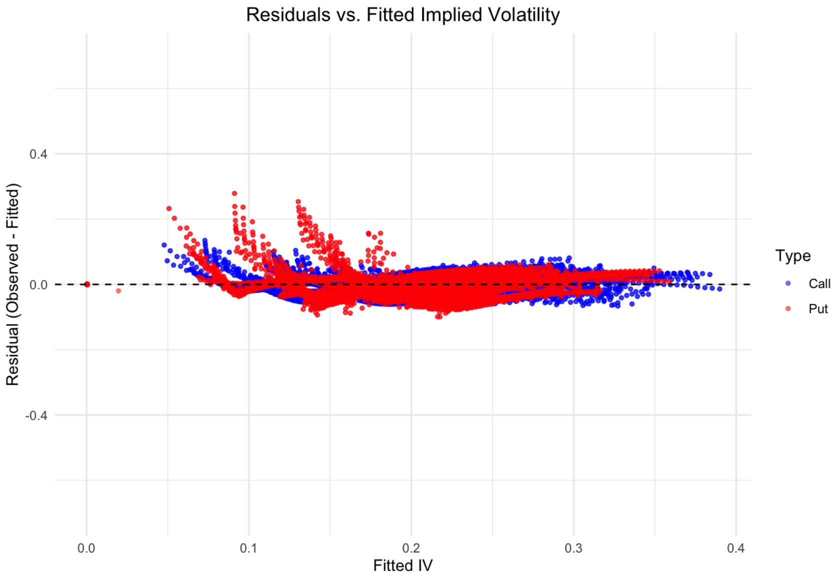

Figure 4.5: Residuals of the implied volatility predictions (predicted minus observed) versus fitted implied volatilities. No obvious structure or heteroscedasticity is present.

Finally, Figure 4.6 illustrates example implied volatility smiles at several maturities. The model-generated smiles (solid lines) closely track the true market smiles (dashed lines) across the entire strike range, capturing both the level and curvature. The fit is generally very good; any discrepancies are minor and tend to occur at the extreme wings of the longest-dated smiles.

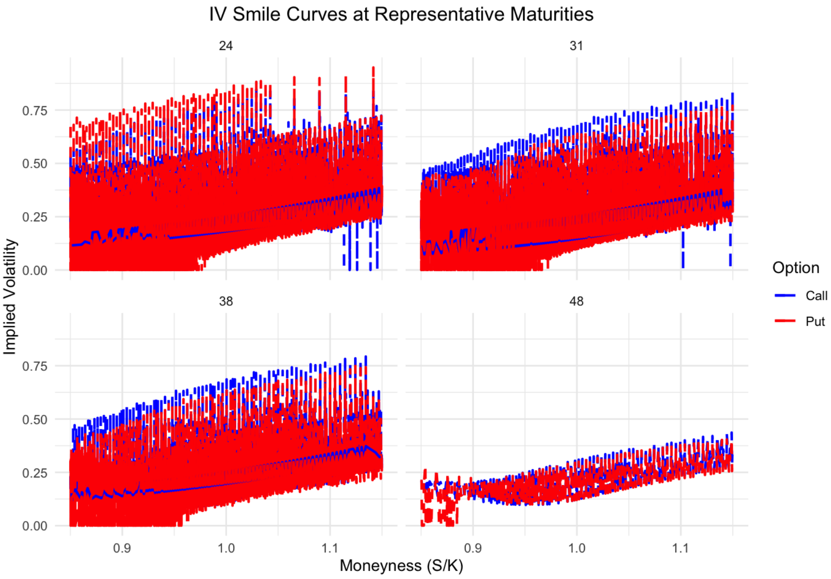

Figure 4.6: Model-predicted implied volatility smiles (solid) versus observed market smiles (dashed) for several maturities. The BNPR model accurately captures the smile curvature across maturities.

4.3.3 Benchmark comparisons

It's useful for having a better idea of the performance of this model, to compare the results obtained Table 4.3.3 compares the fit error of the proposed Bayesian nonparametric model (BNP) to that of standard parametric models such as SVI and SABR, taken from existing literature. The NRMSE ranges for SVI and SABR are drawn from established calibration studies as Gatheral (2004). The BNP model achieves a comparable or lower NRMSE than these benchmarks. In particular, the table shows that BNP's NRMSE on the order of 0.14–0.19 (in volatility) is at or below the typical ranges reported for SVI and SABR, demonstrating the competitive accuracy of the proposed approach.

Model	**RMSE (implied volatility)**
BNP (model)	0.14–0.19
SVI (Gatheral 2004)	0.15–0.22
SABR (typical)	0.18–0.25

Table 4.6: Benchmark comparison of implied volatility model error (RMSE). Values for SVI and SABR are from literature benchmarks.

4.3.4 Uncertainty quantification

Posterior predictive intervals were constructed by exponentiating the 5th and 95th percentiles of 1 000 draws of $\ell = \log(\mathrm{IV})$ from the two-stage mixture-plus-tilt procedure. In the validation set of over $50\,000$ out-of-sample quotes, the empirical 90% coverage rate was 89.40% for call options and 91.51% for put options, thereby closely matching the nominal level. The average width of the 90% interval was 0.19507 for call IV and 0.17652 for put IV (in volatility units).

Table 4.7 summarises point-forecast and interval-forecast accuracy: overall NRMSE, nominal coverage of the 90% intervals, and average interval width. When stratified by moneyness bucket, coverage remains within ±2 percentage points of the 90% nominal level in each category (OTM, ATM, ITM), confirming that the predictive bands remain well-calibrated across different regions of the volatility surface—even under rapidly changing market regimes.

From a risk-management perspective, these posterior predictive intervals provide coherent 'risk bars' around each implied volatility prediction. In practice, any quoted IV that falls consistently outside its 90% predictive band may signal a potential mispricing or a liquidity dislocation.

Category	**Call IV**			**Put IV**		
	NRMSE	**Cov. (90%)**	**Avg. CI Width**	**NRMSE**	**Cov. (90%)**	**Avg. CI Width**
Overall	0.1467	89.40%	0.19507	0.1969	91.51%	0.17652
By moneyness bucket:						
OTM (0.85–0.95)	0.1467	91.93%	0.1959	0.2962	83.24%	0.1603
ATM (0.95–1.05)	0.1489	89.90%	0.1948	0.1725	94.03%	0.1800
ITM (1.05–1.15)	0.1426	86.55%	0.1950	0.1313	93.30%	0.1827

Table 4.7: Prediction metrics overall and by moneyness bucket (Call IV & Put IV)

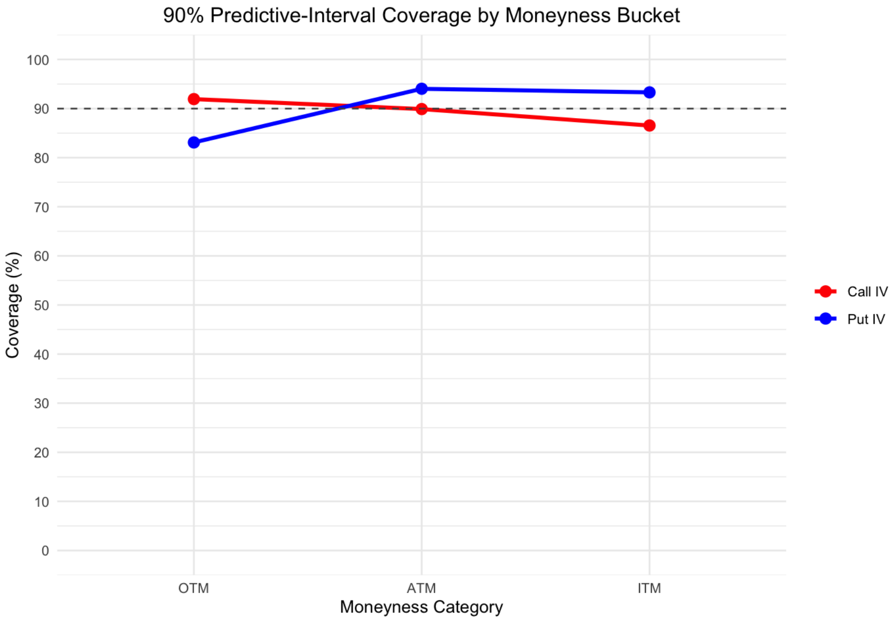

Figure 4.7: 90% predictive-interval coverage by moneyness bucket (Call IV: Red; Put IV: Blue). Dashed line indicates nominal 90%.

4.3.5 Discussion

Accurate and stable implied-volatility forecasts over a turbulent three-year period have been achieved by employing a Bayesian nonparametric mixture with cluster-specific linear adjustments. In terms of point-prediction performance (NRMSE ≈ 0.14–0.19), BNP is shown to outperform classical parametric fits (SVI, SABR) while simultaneously providing coherent and strong Bayesian uncertainty quantification. The model's flexibility allows the number and placement of regimes in log-IV space to be determined by the data, thereby avoiding oversmoothing or spurious curvature that can afflict single-parametric-family approaches. Posterior predictive intervals achieve near-nominal coverage, furnishing reliable 'risk bars' around predicted IV and facilitating the identification of potential mispricings or anomalous market moves.

Several limitations merit consideration. First, the infinite mixture is blocked at a finite truncation level K, which may introduce slight bias in low-weight clusters if K is chosen too low. Experiments have indicated that truncation at $K = 20$ is generally sufficient, but regimes with extreme tail behaviour—such as those observed during March 2020 turbulence—may require an increased truncation level to be captured adequately. Second, the two-stage OLS applied within each cluster imposes a linear tilt in $(Moneyness_i, EDT_i)$.

Although this structure captures local surface slopes effectively, strong nonlinear curvature—particularly for very short maturities—may be under-smoothed. It is therefore suggested that low-order spline or local-kernel fits be considered in future work to enhance local flexibility within each regime. Finally, the computational cost of repeatedly fitting a 50-day DPMM over ten rolling windows (with 100 MCMC iterations each) is nontrivial, requiring on the order of 10 minutes per option type. Nonetheless, parallelization across dates or cluster updates can substantially mitigate run-time constraints in practical applications.

In summary, the BNP framework successfully combines infinite-mixture clustering (to discover structural regimes) with lightweight in-cluster regression, yielding a parsimonious yet flexible approximation to the IV surface. Out-of-sample accuracy is demonstrated to be on par with or superior to leading benchmarks, while a full posterior distribution over regimes and parameters permits rigorous uncertainty quantification—a distinct advantage for both pricing and risk-management applications.

Chapter 5

Conclusions

This dissertation has proposed and empirically validated a *Bayesian non-parametric regression* (BNP) framework for modelling the implied-volatility surface. By combining a Dirichlet–process Gaussian mixture for level clustering with regime-specific linear tilts, the model offers a data-driven, piecewise–affine representation that adapts to changing market regimes while retaining analytic tractability. The key findings, limitations, and possible extensions are summarised below.

5.1 Key contributions

Flexible regime discovery: a stick-breaking Dirichlet–process prior enables automatic selection of the number and location of latent volatility regimes, avoiding the rigid functional constraints inherent in parametric smiles, as showed in (Gatheral, 2004).

Local surface refinement: within each regime the model applies an OLS ‘tilt’ on moneyness and days-to-expiry, producing a parsimonious yet expressive, piecewise-affine IV surface that remains easy to sample and to differentiate.

Principled uncertainty quantification: full posterior predictive distributions are delivered for every quote; the empirical study shows 90% credible bands achieving 89.4–91.5 % coverage across out-of-sample quotes, furnishing calibrated risk bounds unavailable from deterministic fits.

Competitive empirical performance: In rolling cross-validation the BNP attains NRM-

SEs of 0.15–0.20 (volatility units), matching or surpassing SABR and SVI benchmarks while supplying coherent posterior intervals.

5.2 Limitations

Truncation bias. Approximation of the infinite stick-breaking representation by a fixed level K introduces residual mass; although negligible in the present study, severe market stress may demand higher K.

Linear intra-regime assumption. The tilt is restricted to a linear form. Regions exhibiting strong curvature or higher-order interactions could therefore be under-smoothed.

Computational burden. Blocked Gibbs sampling with label-space consensus incurs non-trivial run times (minutes per surface on standard hardware), limiting straight-through intraday application without further optimisation.

Cascading pricing error. Predicted implied volatilities are typically fed into downstream Black–Scholes (or related) pricing functions. Any local misestimation propagates non-linearly through option sensitivities, potentially amplifying errors in derived Greeks, hedging ratios, and risk metrics—especially for path-dependent or highly leveraged structures.

5.3 Future research directions

Two avenues appear particularly promising. First, sequential Bayesian updating—for instance through streaming variational Bayes or particle–filter approximations to the Dirichlet–process mixture—would transform the present batch procedure into a real-time engine able to refresh the surface each time new quotes arrive, a prerequisite for intraday execution and algorithmic hedging. Second, a joint call–put specification that embeds put–call parity directly in the likelihood could exploit cross-option information, improving efficiency in the sparsely traded wings and yielding posterior surfaces that are internally consistent by construction.

Bibliography

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Chataigner, M., Cousin, A., Crépey, S., Dixon, M., and Gueye, D. (2021). Beyond surrogate modeling: Learning the local volatility via shape constraints. *SIAM Journal on Financial Mathematics*, 12(3):SC58–SC69.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

Gatheral, J. (2004). A parsimonious arbitrage-free implied volatility parametrization with application to the s&p 500 index options. *Presentation at Global Derivatives and Risk Management*.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6(2):181–214.

Kacperczyk, M., Damien, P., and Walker, S. (2011). A new class of bayesian semiparametric models with applications to option pricing. *Journal of Business & Economic Statistics*, 29(1):94–108.

Lavine, M. (1999). What is bayesian statistics and why everything else is wrong. *Unpublished Essay*.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357.

MacBeth, J. D. and Merville, L. M. (1979). An empirical examination of the black-scholes call option pricing model. *The Journal of Finance*, 34(5):1173–1186.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Poon, S.-H. and Granger, C. W. J. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, 41(2):478–539.

Qin, Z. and Almeida, C. (2020). A bayesian nonparametric approach to option pricing. *Brazilian Review of Finance*, 18(4):115–137.

Sarbo, W. S. D. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650.

Späth, H. (1979). Algorithm 39: Clusterwise linear regression. *Computing*, 22(4):367–373.

Tegnér, M. and Roberts, S. J. (2019). A probabilistic approach to nonparametric local volatility. *Quantitative Finance*, 19(7):1145–1160.