UNIVERSITÀ COMMERCIALE LUIGI BOCCONI
Corso di Laurea Triennale / Bachelor of Science
in Economics, Management and Computer Science

BAYESIAN TECHNIQUES FOR RACING DYNAMICS: MONZA 2023 F1 GP APPLIED CASE

Relatore / Advisor:
Prof. Omiros Papaspiliopoulos

Tesi di Laurea Triennale di /
Bachelor of Science thesis by:
LUDOVICO AMEDEO PANARIELLO
matricola n. / student ID no. 3192212

Anno Accademico / Academic Year 2024-2025

Università Bocconi
MILANO

To my family — for giving me the space to figure things out on my own.

To my brothers — for teaching me with their courage not to settle.

To Giulia — for supporting me with love, patience, and understanding. To our future together.

To my friends and classmates — for turning pressure into motivation and for making even the hardest days part of what I will carry with me forever.

To my supervisor, Om — for his openness to share with me choices about my future; and to Max, for turning a passion into a thesis topic.

To myself — for keeping the enthusiasm alive and for pushing forward, even when it would have been easier to give up.

To every struggle — for teaching me more than any win ever could.

Just as Formula 1 forces drivers to redefine limits, so has my experience at Bocconi pushed me further than I thought possible.

This is where I start again, determined to pursue new paths.

Alla mia famiglia — per avermi lasciato il tempo e lo spazio per trovare la mia strada da solo.

Ai miei fratelli — perché con il loro coraggio mi hanno insegnato a non accontentarmi.

A Giulia — per avermi sostenuto con Amore, pazienza e comprensione. Al nostro futuro insieme.

Agli amici e ai compagni di corso — perché insieme la pressione è diventata stimolo, e anche le giornate più difficili sono entrate a far parte di un capitolo che porterò con me per sempre.

Al mio relatore, Om — per la sua disponibilità a condividere con me scelte del mio futuro; e a Max per aver trasformato una passione in un tema di tesi.

A me stesso — per aver tenuto vivo l'entusiasmo e aver proseguito, anche quando sarebbe stato più semplice fermarsi.

Alle difficoltà — perché mi hanno forgiato più di qualsiasi traguardo raggiunto.

Così come la Formula 1 porta i piloti a ridefinire i propri limiti, anche il percorso in Bocconi mi ha spinto più lontano di quanto immaginassi.

È da qui che riparto, deciso a percorrere nuove strade.

Contents

1. Introduction and Motivation

1.1 Purpose

In modern Formula 1, managing tire degradation is crucial for race performance. The actual degradation—the compound's loss of properties affecting lap times—is unobservable during a race, unlike indicators such as surface temperature or internal pressure.

Estimating this degradation is challenging due to multiple interacting factors, including track conditions, driving style, tire compound, and stint length. The degradation state is a latent, evolving quantity that must be inferred from noisy and indirect data, framing this as an inherently statistical problem. The significance of tire management in Formula 1 is becoming increasingly central to performance, and its complexity often leads to it being described as a "black art." One clear example of this sentiment comes from McLaren's Team Principal, Andrea Stella, who, while his team is currently dominating the championship, praised his engineering department for mastering one of the sport's most elusive challenges: "I just want to take the opportunity to praise the work that has been done by the engineers at McLaren … and then master one of the matters that still in Formula 1 looks like it's a little bit of a black art, which is dealing with tyres" [14].

This thesis aims to develop and evaluate a statistically principled framework for estimating in-race tire degradation, seeking to shed light on this "black art" through rigorous statistical modeling. The focus is on methods that are both interpretable, allowing for insights into the degradation process, and offer potential for online adaptability in a race context.

A key aspect of this work involves structuring data handling and preprocessing to effectively train and validate the proposed models—specifically, the linear Gaussian and natural spline models detailed in Chapters 3 and 4. This ensures that the models can learn meaningful tire degradation patterns from observable race data.

1.2 Methods

The modeling approach is grounded in the theory of Bayesian state-space models [5]. We define a latent degradation state $\theta_t \in \mathbb{R}$ that evolves over time according to a stochastic process, and we specify a measurement equation that relates θ_t to observable telemetry data, such as lap times or sector deltas. The general formulation is:

$$\theta_t = f(\theta_{t-1}) + \eta_t,$$
$$y_t = g(\theta_t) + \epsilon_t,$$

where $\eta_t \sim \mathcal{N}(0, \tau^2)$ is process noise and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is observation noise. The functions f and g determine the dynamic and measurement models, respectively. Depending on the assumptions made on these functions, the resulting inference techniques vary in complexity and tractability.

Our methodological journey commences with a foundational linear Gaussian model, where $f(\theta_{t-1}) = \theta_{t-1} + \beta$ and $g(\theta_t) = \theta_t$. This initial framework, often solvable with techniques like the Kalman filter [10], provides a robust baseline and allows for the clear interpretation of degradation as a linear trend. It serves as an essential first step in understanding the basic dynamics and the impact of key covariates.

To capture more nuanced and realistic non-linear degradation patterns, which are frequently observed in tire performance data, we then progress to semi-parametric techniques. Specifically by employing Bayesian natural splines. This approach models the degradation trajectory with significant flexibility by representing it as a linear combination of natural spline basis functions (a specific type of B-spline basis, constrained to be linear beyond the boundary knots, and generated in R using ‘ns()‘). The smoothness of the resulting curve is primarily achieved through the Bayesian prior placed on the spline coefficients. This allows the model to adapt to complex shapes in the degradation profile, such as initial conditioning phases, periods of stable wear, and accelerated degradation towards the end of a tire's life, without imposing rigid parametric assumptions. The Bayesian formulation provides a natural way to quantify uncertainty in the estimated non-linear degradation function.

The core analytical framework in this thesis relies on two main approaches: linear

Gaussian models and Bayesian natural spline models. Each modeling approach is evaluated in terms of its predictive performance, ability to capture known phenomena, uncertainty quantification, and computational considerations. The goal is not only to estimate the latent state θ_t accurately but also to provide credible intervals that reflect the epistemic and aleatoric uncertainty inherent to the problem. While more advanced techniques such as Gaussian Processes and Particle Filters offer promising avenues for future research, particularly for enhanced non-parametric modeling and real-time online inference respectively, for now this thesis concentrates on establishing a solid bayesian modeling foundation.

1.3 Data

The empirical illustrations in this thesis are grounded in Formula 1 telemetry and timing data, which is programmatically accessed and processed. For this purpose, the FastF1 Python library [15] serves as a crucial tool. FastF1 is an unofficial, open-source package providing comprehensive access to F1 data, encompassing live timing information, historical race data, car telemetry, and session results. It interfaces with the Ergast Developer API and other data sources, parsing raw information into structured formats suitable for analysis. The capabilities of FastF1 leveraged in this work include its functions for loading lap-by-lap data, accessing detailed telemetry channels (e.g., speed, RPM, gear, throttle, brake), and retrieving information about tire compounds, stint lengths, and pit stops for individual drivers across various race events.

The models developed herein are designed to operate on typical data structures obtainable from such sources. The primary types of data relevant to the modeling of tire degradation, and their conceptual roles within the statistical frameworks presented, include:

- **Lap-Level Performance Metrics:**
  - *Lap Time:* This scalar value, representing the total time taken to complete a lap, serves as the primary response variable in our models. The objective is to predict and understand the systematic changes in lap time as a function of tire wear and other factors.

- *Tire Compound:* A categorical variable (e.g., Soft, Medium, Hard) indicating the specific tire specification used. This is a critical predictor, as different compounds exhibit distinct baseline performance and degradation characteristics. Models incorporate compound-specific intercepts or coefficients to capture these variations, as seen in both the linear Gaussian (Chapter 3) and natural spline model (Chapter 4) formulations.
- *Lap Number:* While primarily an ordinal counter, lap number can implicitly capture effects like track evolution or changing fuel loads if these are not explicitly modeled as separate covariates. It also helps sequence the observations correctly.
- *Pit Flags:* Boolean indicators for pit-in and pit-out laps are crucial for data segmentation, specifically for identifying the start and end of a stint and for excluding laps that are not representative of continuous racing performance under normal conditions.

- **Contextual and Derived Variables:**
  - *Driver Identification:* A categorical variable used to account for systematic variations in performance that can be attributed to individual driver characteristics (e.g., driving style, inherent pace). Models may include driver-specific random or fixed effects, as implemented in the provided R scripts for both linear and spline approaches.
  - *Stint Identifier:* A grouping variable that uniquely identifies a sequence of consecutive laps run on a single set of tires by a specific driver. This is fundamental for calculating tire age accurately.
  - *Tire Age (`TireLife`):* This is a key numerical predictor, derived by counting the number of laps a specific tire set has completed within its current stint. It is the primary variable through which tire degradation is modeled.
    - In linear Gaussian models (Chapter 3), `TireLife` enters as a direct predictor with interactions with `TireCompound`, to estimate a rate of lap time increase per lap of tire use. The R code for the linear model (`linear_code.R`) exemplifies this by using `TireLife` as a covariate.

    - In semi-parametric spline models (Chapter 4), `TireLife` serves as the independent variable for the basis functions (e.g., B-splines). The model then estimates coefficients for these basis functions, allowing for a flexible, non-linear relationship between tire age and lap time degradation. The Stan code (`splines_model.stan`) and its accompanying R script (`splines_code.R`) demonstrate this by constructing a B-spline basis matrix from `TireLife` values, which then forms part of the design matrix for predicting lap times.

- **Telemetry Data (Secondary Focus for this Thesis):** While the primary models in this thesis focus on lap-level data for broader applicability, more granular telemetry (e.g., speed, distance, RPM, gear, throttle, brake, DRS status) sampled at high frequency can provide richer information for more detailed or alternative modeling approaches, such as those involving Gaussian Processes or Particle Filters mentioned as future avenues.

- **Session and Environmental Information:** Contextual data such as the event name, session type (Practice, Qualifying, Race), track status (e.g., safety car periods), and weather conditions (ambient temperature, track temperature) are important covariates that can influence tire behavior. While often considered in comprehensive analyses, they are treated as controlled or averaged out in the primary models of this thesis for simplicity, to maintain focus on the core degradation dynamics influenced by tire age and compound.

Although this thesis is primarily theoretical, focusing on the development and Bayesian treatment of statistical models for tire degradation, the data serves an illustrative purpose. It allows for the demonstration of model application, the exploration of inferential outcomes, and the comparison of different modeling strategies using realistic scenarios. It is important to note that the data used in the following chapters is drawn from the 2023 Monza Grand Prix, specifically the race stints of drivers Leclerc, Verstappen, Gasly, and Norris.

1.4 Theoretical Framework

The theoretical framework of Bayesian statistics provides a robust foundation for the objective. The work by [9] offers a comprehensive and practically oriented overview of Bayesian methods, which has guided the modeling choices adopted in this thesis. Bayesian inference allows for the incorporation of prior knowledge, the quantification of uncertainty in a principled manner, and the flexible modeling of complex dependencies. Core concepts such as prior distributions, likelihood functions, and posterior distributions are central to this approach, as detailed by [13].

At the heart of Bayesian inference is the updating of beliefs in light of new evidence, formalized through Bayes' theorem. In the context of this thesis, Bayesian methods are particularly suited for modeling tire degradation due to their ability to handle complex, non-linear relationships and to update predictions as more data becomes available during a race. The incorporation of prior distributions reflects any existing knowledge about tire degradation patterns, which can be updated with race data to yield posterior distributions representing updated beliefs about the degradation state.

The flexibility of Bayesian models allows for the accommodation of various sources of uncertainty, including the inherent variability in tire performance, measurement errors in lap times, and the stochastic nature of tire degradation processes. By employing hierarchical modeling approaches, the models can also account for variations across different races, drivers, and tire compounds, capturing the multi-level structure of the data.

Computationally, the implementation of Bayesian models in this thesis utilizes Markov-Chain Monte Carlo (MCMC) methods for posterior inference, specifically the No-U-Turn Sampler (NUTS) algorithm, as implemented in the Stan modeling language [2]. This approach enables the fitting of complex models with high-dimensional parameter spaces, typical in Bayesian hierarchical modeling. The use of informative priors, based on historical data and expert knowledge, further aids in stabilizing the estimates and improving the convergence of the MCMC algorithms.

In summary, the theoretical underpinnings of Bayesian statistics, coupled with its computational implementation through advanced MCMC methods, provide a powerful toolkit for tackling the challenging problem of tire degradation modeling in Formula 1. The following chapters will detail the specific modeling approaches and their applications

to real race data, illustrating the practical utility of this theoretical framework.

2. Bayesian Framework

This chapter establishes the Bayesian modeling foundations essential for studying tire degradation in Formula 1. It covers the Bayesian interpretation of probability, posterior inference, predictive distributions, and computational methods, with a focus on Hamiltonian Monte Carlo (HMC). This foundational toolkit underpins the analytical approach of the subsequent modeling chapters.

2.1 Philosophical and Mathematical Foundations of Bayesian Inference

In Bayesian statistics, probability is interpreted as a degree of belief or uncertainty. This framework allows updating beliefs about a parameter θ after observing data y by combining prior beliefs with observed information. Bayes' theorem, derived from conditional probability, is central:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{2.1}$$

This extends to statistical modeling with continuous parameters. For an unknown parameter $\theta \in \Theta$ and observed data y, Bayes' theorem is:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}, \tag{2.2}$$

where:

- $p(\theta)$ is the **prior distribution**, representing beliefs about θ before seeing data;
- $p(y \mid \theta)$ is the **likelihood**, modeling how likely the data is given θ;
- $p(\theta \mid y)$ is the **posterior distribution**, our updated belief after seeing y;
- $p(y) = \int p(y \mid \theta)p(\theta)\, d\theta$ is the **marginal likelihood**, ensuring normalization.

This formulation of Bayes' theorem is the cornerstone of Bayesian learning, providing a robust mechanism to update beliefs and quantify uncertainty in the presence of data.

2.1.1 Prior Distributions in Practice

The choice of prior distribution, $p(\theta)$, is a defining characteristic of Bayesian inference. Priors can range from being highly informative, encoding specific expert knowledge or results from previous studies, to weakly informative or non-informative, intended to let the data speak for itself as much as possible. In the context of regression models, such as the linear and natural spline models explored in Chapters 3 and 4, common choices include:

- **Gaussian (Normal) priors** for regression coefficients (e.g., β_j). These are often centered at zero, $\mathcal{N}(0, \sigma_\beta^2)$, implying that large deviations from zero are less likely a priori. The variance σ_β^2 controls the degree of regularization: smaller variances lead to stronger shrinkage of coefficients towards zero.
- **Half-Cauchy or Inverse-Gamma priors** for variance parameters (e.g., σ^2, the residual variance, or variances of random effects). These priors are defined on positive real numbers and are often chosen for their flexibility and good frequentist properties in some settings. For example, a Half-Cauchy prior is often recommended as a weakly informative prior for scale parameters.

Careful consideration of priors is important, as they can influence the posterior, especially with limited data. Sensitivity analysis, exploring how posterior inferences change under different prior specifications, is a good practice.

2.1.2 Likelihood Functions for Regression

The likelihood function, $p(y \mid \theta)$, quantifies how probable the observed data y are for a given set of parameters θ. For models aiming to predict a continuous outcome, such as the tire degradation metrics in this thesis, a common choice is the Gaussian (Normal) likelihood. If y_i is an individual observation and μ_i is its predicted mean from the model (e.g., $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ in a linear model), the likelihood for y_i is often assumed to be:

$$p(y_i \mid \theta, \sigma^2) = \mathcal{N}(y_i \mid \mu_i(\theta), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i(\theta))^2}{2\sigma^2}\right), \tag{2.3}$$

where σ^2 is the observation variance. The full likelihood for N independent observations is then the product $\prod_{i=1}^{N} p(y_i \mid \theta, \sigma^2)$. This choice implies that the errors $(y_i - \mu_i(\theta))$ are

normally distributed.

2.1.3 Posterior and Predictive Inference

Once we have computed the posterior distribution $p(\theta \mid y)$, we can summarize it in various ways depending on the inferential goal. Common summaries include:

- The **posterior mean** $\mathbb{E}[\theta \mid y]$, which minimizes squared error loss.
- The **posterior mode** $\arg\max_\theta p(\theta \mid y)$, useful when the posterior is unimodal.
- **Credible intervals**, which provide intervals $[a, b]$ such that $\mathbb{P}(a < \theta < b \mid y) = 1 - \alpha$.

A key strength of Bayesian inference is its natural and coherent approach to prediction. Rather than conditioning on a point estimate of θ, Bayesian predictions account for our full uncertainty about θ. The **posterior predictive distribution** for a new data point $\tilde{y}$ is:

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta) p(\theta \mid y)\, d\theta. \tag{2.4}$$

This integral averages the likelihood of future data over the posterior distribution of the parameters, ensuring that all sources of uncertainty are integrated into predictions. The posterior predictive distribution is not only crucial for making predictions about future, unobserved data but also plays a vital role in **posterior predictive checking**. This involves generating replicated datasets $\tilde{y}^{\text{rep}}$ from $p(\tilde{y} \mid y)$ and comparing them to the observed dataset y. If the model is a good fit to the data, the replicated datasets should look similar to the observed data in terms of relevant summary statistics or graphical features. Discrepancies can highlight aspects where the model fails to capture the data generating process, guiding model refinement.

2.1.4 Model Comparison and Adequacy

In practice, multiple candidate models may be proposed to explain a given phenomenon. Bayesian statistics offers a principled framework for comparing these models and assessing their adequacy.

Information Criteria

For complex models where the marginal likelihood $p(y)$ is difficult to compute directly, information criteria provide a practical way to estimate out-of-sample predictive accuracy. Two widely used criteria are:

- **Watanabe-Akaike Information Criterion (WAIC):** WAIC is a more fully Bayesian approach to estimating out-of-sample expectation and is asymptotically equal to Bayesian cross-validation. It is calculated as:

$$\text{WAIC} = -2\left(\text{lppd} - p_{\text{WAIC}}\right), \tag{2.5}$$

  where lppd is the log pointwise predictive density, and p_{WAIC} is an effective number of parameters penalty.

- **Leave-One-Out Cross-Validation (LOO-CV):** LOO-CV estimates the predictive accuracy by iteratively holding out one data point, fitting the model to the remaining data, and predicting the held-out point. While computationally intensive if done naively, approximations like Pareto Smoothed Importance Sampling LOO (PSIS-LOO) make it feasible.

Lower values of WAIC or LOO-CV generally indicate better expected predictive performance. These criteria will be employed in Chapter 5 to evaluate the relative performance of the linear and spline models for tire degradation.

Graphical Posterior Predictive Checks

Beyond numerical summaries, graphical checks are invaluable. As mentioned earlier, by simulating replicated data $\tilde{y}^{\text{rep}}$ from the posterior predictive distribution, we can compare various aspects of $\tilde{y}^{\text{rep}}$ to the observed data y. For instance, one might compare histograms, means, variances, or specific quantiles. If the model captures the data's characteristics well, these simulated summaries should be consistent with those from the observed data. Such checks provide qualitative insights into model fit and potential areas of mis-specification.

2.2 Bayesian Computation and MCMC

In simple models, the posterior distribution $p(\theta \mid y)$ may have a closed-form solution. However, for most realistic models — particularly hierarchical, non-linear, or high-dimensional ones — the posterior must be approximated using numerical methods. The most common class of such methods is **Markov Chain Monte Carlo** (MCMC), which generates samples from the posterior by simulating a Markov chain that has $p(\theta \mid y)$ as its stationary distribution.

2.2.1 Markov Chains and Ergodicity

A **Markov chain** is a sequence of random variables $\{\theta^{(t)}\}_{t=1}^{\infty}$ with the property that the future depends only on the present, not on the past:

$$\mathbb{P}(\theta^{(t+1)} \in A \mid \theta^{(t)}, \ldots, \theta^{(0)}) = \mathbb{P}(\theta^{(t+1)} \in A \mid \theta^{(t)}). \tag{2.6}$$

The chain evolves according to a **transition kernel** $K(\theta' \mid \theta)$, which specifies the probability of moving from state θ to θ'. A distribution π is **stationary** for K if:

$$\pi(\theta') = \int K(\theta' \mid \theta)\pi(\theta)\, d\theta. \tag{2.7}$$

A Markov chain is **ergodic** if, regardless of the starting value, the distribution of $\theta^{(t)}$ converges to π as $t \to \infty$. The **ergodic theorem** guarantees that empirical averages converge to posterior expectations:

$$\frac{1}{T}\sum_{t=1}^{T} f(\theta^{(t)}) \to \mathbb{E}_{\pi}[f(\theta)] \quad \text{as } T \to \infty. \tag{2.8}$$

This theorem underpins the practice of approximating posterior expectations by averaging functions of the MCMC samples.

2.2.2 Metropolis–Hastings Algorithm

The most basic and widely used MCMC algorithm is Metropolis–Hastings. Given a current state θ, we propose a new state θ' from a **proposal distribution** $q(\theta' \mid \theta)$ and

accept it with probability:

$$\alpha(\theta, \theta') = \min\left(1, \frac{p(\theta')q(\theta \mid \theta')}{p(\theta)q(\theta' \mid \theta)}\right). \tag{2.9}$$

If the proposal is accepted, we set $\theta^{(t+1)} = \theta'$; otherwise, we retain the current state. Under mild conditions, this procedure generates a Markov chain whose stationary distribution is the desired posterior.

While broadly applicable, the Metropolis–Hastings algorithm can be inefficient in high-dimensional spaces. The random walk behavior often leads to slow mixing and high autocorrelation. This motivates more sophisticated methods such as Hamiltonian Monte Carlo.

2.3 Hamiltonian Monte Carlo (HMC)

Hamiltonian Monte Carlo is a gradient-based MCMC algorithm that reduces random walk behavior by introducing auxiliary momentum variables and simulating deterministic dynamics through parameter space. It achieves more efficient exploration by proposing distant moves with high acceptance probability.

2.3.1 Hamiltonian Dynamics

We begin by augmenting the parameter space $\theta \in \mathbb{R}^d$ with a momentum variable $\rho \in \mathbb{R}^d$. We define a joint distribution:

$$p(\theta, \rho) = p(\theta)p(\rho), \tag{2.10}$$

where $p(\rho) = \mathcal{N}(0, M)$ for some positive definite mass matrix M. The joint density defines a **Hamiltonian** function:

$$H(\theta, \rho) = U(\theta) + K(\rho), \tag{2.11}$$

where $U(\theta) = -\log p(\theta)$ is the potential energy, and $K(\rho) = \frac{1}{2}\rho^T M^{-1}\rho$ is the kinetic energy.

The dynamics of the system are governed by Hamilton's equations:

$$\frac{d\theta}{dt} = \nabla_\rho H = M^{-1}\rho, \tag{2.12}$$

$$\frac{d\rho}{dt} = -\nabla_\theta H = -\nabla_\theta U(\theta). \tag{2.13}$$

These equations describe a trajectory through (θ, ρ) space that preserves the total energy $H(\theta, \rho)$. If we could simulate this trajectory exactly, we could use it to propose new states with probability one. In practice, we simulate it approximately using the leapfrog method.

2.3.2 Leapfrog Integrator and Acceptance Step

The leapfrog integrator is a time-reversible and volume-preserving method that approximates the Hamiltonian dynamics in three steps:

1. Half-step momentum update: $\rho \leftarrow \rho - \frac{\epsilon}{2}\nabla_\theta U(\theta)$
2. Full-step position update: $\theta \leftarrow \theta + \epsilon M^{-1}\rho$
3. Half-step momentum update: $\rho \leftarrow \rho - \frac{\epsilon}{2}\nabla_\theta U(\theta)$

In this context, the symbol $\leftarrow$ denotes computational assignment, where the value on the right-hand side is computed and then used to update the variable on the left-hand side. This is a common notation in algorithmic descriptions to distinguish from a strict mathematical equality.

After L leapfrog steps of size ϵ, we obtain a proposal (θ^*, ρ^*). This is accepted with probability:

$$\min\left(1, \exp\left[H(\theta, \rho) - H(\theta^*, \rho^*)\right]\right). \tag{2.14}$$

If the leapfrog integration were exact, the Hamiltonian would be conserved and the proposal always accepted. In practice, small integration errors are corrected by the Metropolis step.

2.3.3 Benefits and Practical Use

HMC is particularly effective for models with continuous parameters and differentiable log-posteriors. It allows for larger, informed moves in parameter space and tends to exhibit

lower autocorrelation than random walk Metropolis. However, its performance depends on tuning parameters such as the step size ϵ, number of steps L, and the mass matrix M. The No-U-Turn Sampler (NUTS) is an adaptive variant that removes the need to manually set L and improves robustness.

In this thesis, HMC is the main computational tool used to sample from the posterior distributions of models of tire degradation, including those based on latent variable dynamics. Its use enables precise inference in complex models where traditional methods would be inefficient.

2.3.4 Model Checking and Evaluation

A crucial aspect of the Bayesian workflow is model checking, which involves assessing the fit of the model to the data and evaluating its predictive performance. This can be done through several techniques:

- **Posterior Predictive Checks (PPCs):** PPCs involve simulating replicated data from the posterior predictive distribution and comparing these simulations to the observed data. Discrepancies can indicate model misfit. Graphical checks, such as comparing histograms or density plots of observed versus replicated data, are common [9].
- **Information Criteria:** Metrics like the Watanabe-Akaike Information Criterion (WAIC) or the Leave-One-Out Cross-Validation (LOO-CV) provide estimates of a model's out-of-sample predictive accuracy, penalizing for model complexity. These are useful for comparing different models [16].
- **Sensitivity Analysis:** This involves examining how sensitive the posterior inferences are to changes in prior specifications or likelihood assumptions. Robust models should exhibit minimal sensitivity to reasonable variations in these components.

These methods collectively help in validating the model, understanding its limitations, and ensuring that the conclusions drawn are well-supported by the data. The process of building, fitting, and assessing model fit is iterative, often leading to model refinement.

3. Linear Gaussian Modeling

This chapter presents the first applied model for investigating Formula 1 tire degradation: a Bayesian latent variable model. It captures degradation evolution per stint, accounting for tire compound and driver effects, using a dynamic state-space structure. This allows differentiation of degradation patterns by compound while considering driver performance.

3.1 Model Specification

We model the standardized observed lap time y_n^{std} at each lap $n = 1, ..., N$ as

$$y_n^{\text{std}} \sim \mathcal{N}(\mu_{\text{driver}[n]} + \beta_{\text{compound}[n]} + \theta_n^{\text{std}}, \sigma^2), \tag{3.1}$$

where $y_n^{\text{std}} = (y_n - \bar{y})/s_y$ is the lap time y_n standardized using the overall mean $\bar{y}$ and standard deviation s_y. Consequently, all mean and degradation parameters are on this standardized scale.

- $\mu_{\text{driver}[n]}$ is the driver-specific baseline performance (standardized).
- $\beta_{\text{compound}[n]}$ is the additive adjustment for the compound in use (standardized). For identifiability, the effect of one compound (e.g., HARD) is fixed to zero.
- θ_n^{std} is a latent variable that represents cumulative tire degradation up to lap n (standardized).
- σ is the observation noise.

The model accommodates data from multiple drivers, various tire compounds, and numerous stints (defined as continuous sequences of laps on a given compound without pit stops). We denote the total number of drivers by D, stints by S, and tire compounds by C.

3.2 Hierarchical Priors and Latent Structure

We adopt hierarchical priors for driver baselines μ_d (represented as `mu_driver` in Stan, derived from `mu_0` and `z_mu_driver`).

These are defined as follows.

$$\begin{aligned}
\mu_0 &\sim \mathcal{N}(0, 1) \\
\sigma_\mu &\sim \text{HalfNormal}(0, 0.5) \\
z_{\mu,d} &\sim \mathcal{N}(0, 1) \quad \text{for } d = 1, \ldots, D \\
\mu_d &= \mu_0 + \sigma_\mu z_{\mu,d}
\end{aligned}$$

We also introduce compound-specific adjustments β_c (represented as `beta_comp` in Stan, derived from `sigma_comp` and `z_beta`, with one compound fixed to 0).

Their definitions are as follows.

$$\begin{aligned}
\sigma_{\text{comp}} &\sim \text{HalfNormal}(0, 0.5) \\
z_{\beta,c} &\sim \mathcal{N}(0, 1) \quad \text{for } c = 1, \ldots, C-1 \\
\beta_1 &= 0 \\
\beta_c &= \sigma_{\text{comp}} z_{\beta,c} \quad \text{for } c = 2, \ldots, C
\end{aligned}$$

Central to our model is the latent degradation process, θ_n^{std}, which evolves over laps. It resets to zero at the commencement of each new stint (i.e., if $\text{reset}_n = 1$ in the data, where n is the lap index) and also for the very first lap overall ($n = 1$). For subsequent laps within a stint ($n > 1$ and $\text{reset}_n = 0$), its evolution is defined as:

$$\theta_n^{\text{std}} = \theta_{n-1}^{\text{std}} + \text{drift}_n + \eta_n, \quad \text{where } \eta_n \sim \mathcal{N}(0, \tau^2) \tag{3.2}$$

The term τ (Stan: `tau`) captures the scale of the random walk noise and is given a $\text{HalfNormal}(0, 0.2)$ prior. The systematic degradation per lap, drift_n, is given by:

$$\text{drift}_n = \delta_{\text{driver}[n]} + \delta_{\text{compound}[n]} \tag{3.3}$$

Here, $\delta_{\text{driver}[n]}$ represents the driver-specific component of the drift, and $\delta_{\text{compound}[n]}$ represents the compound-specific component. The driver-specific drift includes a global mean

drift μ_δ, whereas the compound-specific drift is modeled as deviations around zero. This formulation assumes that the primary, consistent positive drift (degradation) is captured by μ_δ, while compound effects modulate this around a net-zero average, after accounting for the main driver effect. The specific definitions are:

$$\begin{aligned}
\mu_\delta &\sim \mathcal{N}(0, 0.2) \\
\sigma_{\delta,\text{driver}} &\sim \text{HalfNormal}(0, 0.2) \\
z_{\delta,\text{driver}[d]} &\sim \mathcal{N}(0, 1) \quad \text{for } d = 1, \ldots, D \\
\delta_{\text{driver}[d]} &= \mu_\delta + \sigma_{\delta,\text{driver}} \cdot z_{\delta,\text{driver}[d]}
\end{aligned}$$

The compound-specific degradation components δ_c (Stan: `delta_compound`) are modeled similarly but without a shared mean μ_δ:

$$\begin{aligned}
\sigma_{\delta,\text{compound}} &\sim \text{HalfNormal}(0, 0.2) \\
z_{\delta,\text{compound}[c]} &\sim \mathcal{N}(0, 1) \quad \text{for } c = 1, \ldots, C \\
\delta_{\text{compound}[c]} &= \sigma_{\delta,\text{compound}} \cdot z_{\delta,\text{compound}[c]}
\end{aligned}$$

It's important to note that priors for positively constrained scale parameters (like σ_μ, σ_{comp}, $\sigma_{\delta,\text{driver}}$, $\sigma_{\delta,\text{compound}}$, τ) are specified conceptually as HalfNormal. In the Stan implementation, these are achieved by declaring the parameter with a lower bound of 0 (e.g., `real<lower=0> sigma_param`) and assigning a `normal(0, scale)` prior, which effectively uses the positive half of the normal distribution.

The priors for the remaining scale parameters are specified as:

$$\begin{aligned}
\sigma &\sim \text{HalfNormal}(0, 0.5) \\
\tau &\sim \text{HalfNormal}(0, 0.2)
\end{aligned}$$

The standardized error terms $z_{\eta,n-1}$ for the process noise are $z_{\eta,n-1} \sim \mathcal{N}(0, 1)$, so that $\eta_{n-1} = \tau z_{\eta,n-1}$.

This formulation replaces the previous description of δ_s and $\bar{\delta}_c$. The degradation increment is now a sum of driver and compound effects, each with its own hierarchical structure.

3.3 Inference and Diagnostics

The model parameters were estimated using Hamiltonian Monte Carlo (HMC) as implemented in `Stan`, the details of which are described in Chapter 2. We employed four chains, each with a substantial number of warmup and sampling iterations (4000 warmup and 8000 sampling iterations per chain). Convergence was assessed by ensuring $\hat{R}$ statistics were below 1.05 and effective sample sizes (ESS) were adequate (e.g., > 400) for key parameters, as discussed in Chapter 2.

Posterior predictive checks (PPCs), also introduced in Chapter 2, are crucial for assessing model fit. As shown in Figure 3.5, we compare the distribution of observed lap times against the distributions of replicated lap times generated from the fitted model. Good correspondence suggests the model captures the essential features of the data.

Further diagnostics include examining the residuals ($y_n - E[y_n|\text{data}]$) for any systematic patterns.

3.4 Results and Interpretation

This section presents the key findings from the Bayesian linear model. We discuss the estimated parameters, the inferred latent tire degradation, and the model's diagnostic checks.

3.4.1 Parameter Estimates

The posterior distributions of the key model parameters provide insights into driver baselines, compound effects, and degradation dynamics. Figure 3.1 displays a forest plot summarizing the 95% credible intervals for a selection of these parameters, including the overall intercept (μ_0), standard deviations of hierarchical effects (e.g., σ_μ, σ_{comp}), and key degradation-related terms.

This allows for a quantitative assessment of the magnitude and uncertainty associated with each component of the model. The hierarchical structure enables robust estimation by pooling information across drivers and compounds, as discussed in Chapter 2.

3.4.2 Latent Tire Degradation

A core output of the model is the estimated latent cumulative tire degradation, θ_n^{std}, for each lap. Figure 3.2 illustrates the posterior mean and 95% credible intervals for the cumulative degradation trajectories, potentially averaged or shown for representative stints.

This visualization is crucial for understanding the model's ability to capture the unobserved wear process and its impact on lap times.

3.4.3 Model Diagnostics

Thorough diagnostic checks were performed to ensure the reliability of the MCMC simulations and the adequacy of the model fit.

MCMC Convergence

As mentioned in Section 3.3, convergence was primarily assessed using $\hat{R}$ statistics and Effective Sample Sizes (ESS). Visual diagnostics further support these numerical summaries. Figure 3.3 presents trace plots for key parameters, showing well-mixed chains that have converged to a stable posterior distribution. Correspondingly, Figure 3.4 displays the posterior density estimates from different chains, which should largely overlap.

Posterior Predictive Checks

Posterior predictive checks (PPCs) were used to evaluate how well the model captures the observed data distribution, as introduced in Chapter 2. Figure 3.5 compares the density of the observed lap times (y) with the densities of multiple replicated datasets (y^{rep}) generated from the posterior predictive distribution.

Additionally, examination of residuals ($y_n - E[y_n|\text{data}]$) did not reveal strong systematic patterns, further supporting the model's fit.

3.5 Conclusion

In this chapter, we have detailed a Bayesian linear model incorporating hierarchical and temporal structures to capture cumulative tire degradation in Formula 1. The latent state

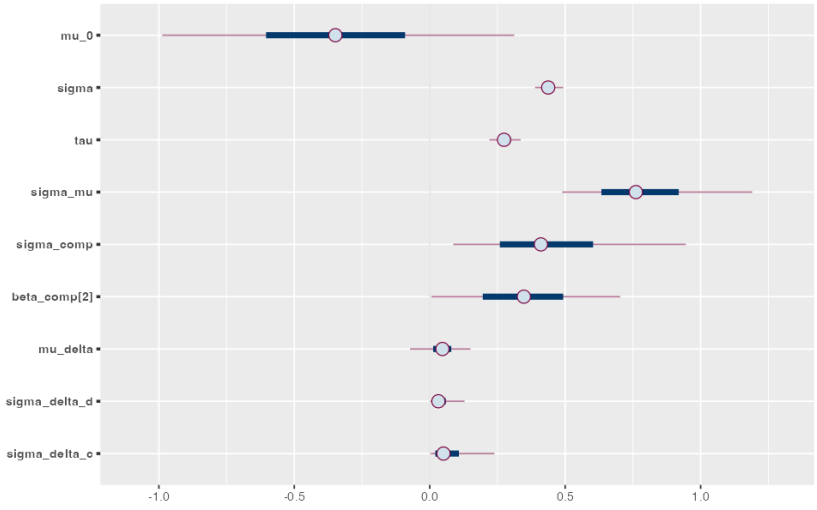

Figure 3.1: Forest plot of key parameter estimates from the linear model. The plot shows the posterior means and 95% credible intervals for the main effects, highlighting which parameters are most influential and how much uncertainty is present.

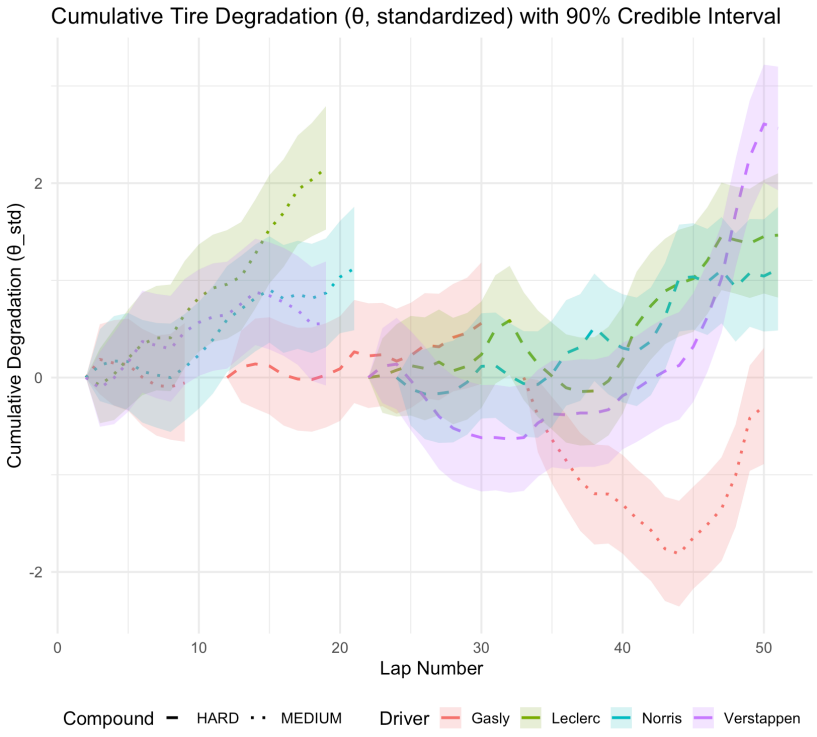

Figure 3.2: Estimated cumulative tire degradation (θ_n^{std}) with 95% credible intervals. The plot shows how the model infers the hidden degradation process over time.

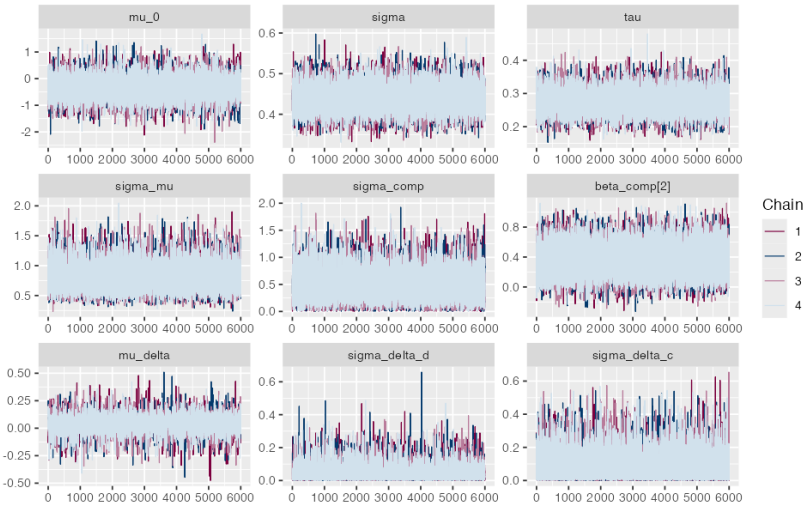

Figure 3.3: Trace plots for key parameters of the linear model. Each line represents a Markov chain for a parameter. Good mixing and lack of trends suggest convergence.

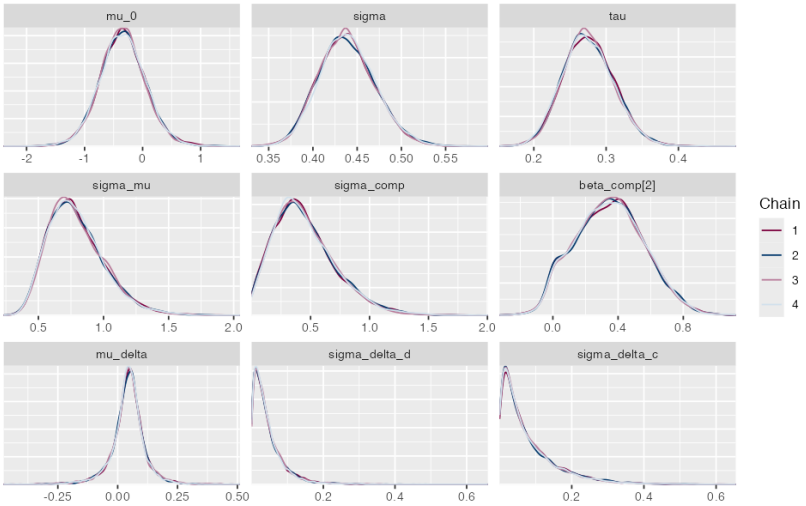

Figure 3.4: Posterior density plots for key parameters of the linear model. Overlapping densities from different chains indicate good mixing and convergence.

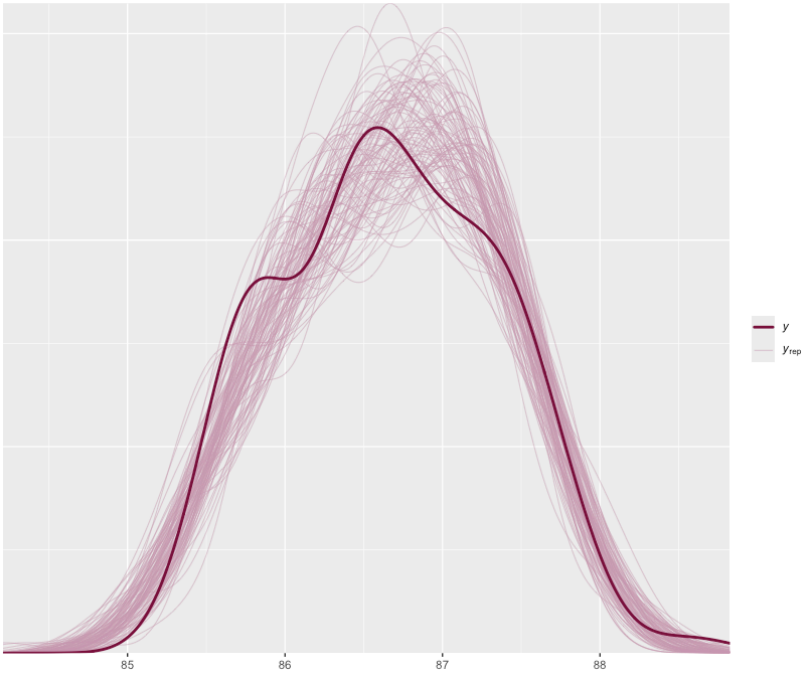

Figure 3.5: Posterior predictive check for the lap time model using the linear degradation formulation. The density of the observed lap times (y) is overlaid with 50 replications from the posterior predictive distribution (y_rep).

process, θ_n, was modeled not as a simple random walk, but as a dynamic evolution with a stint-specific mean degradation rate, driver, and compound effects.

4. Bayesian Spline-Based Modeling

This chapter introduces Bayesian Natural Splines for modeling tire degradation, building on linear models and extending to non-linear, data-driven state-space structures.

4.1 Mathematical Foundations of Spline Smoothing

Splines, specifically B-splines, enable flexible non-linear regression. Given knots $\Xi = \{\xi_1 < \cdots < \xi_{K+d+1}\}$ on $[a, b]$, the B-spline basis $\{B_{k,d}(x)\}_{k=1}^{K}$ is defined recursively:

$$B_{k,0}(x) = \mathbf{1}_{[\xi_k, \xi_{k+1})}(x), \qquad B_{k,d}(x) = \frac{x - \xi_k}{\xi_{k+d} - \xi_k} B_{k,d-1}(x) + \frac{\xi_{k+d+1} - x}{\xi_{k+d+1} - \xi_{k+1}} B_{k+1,d-1}(x)$$

with compact support and partition of unity $\sum_{k=1}^{K} B_{k,d}(x) = 1$. See [3], [6].

Natural Splines are a type of B-spline with the additional constraint that they are linear beyond the boundary knots. This often provides more stable and realistic behavior at the extremes of the data range. In our implementation, the basis for Natural Splines is generated using the 'ns()' function in R.

The general idea of P-splines [7] involves modeling $f(x) = \sum_{k=1}^{K} \beta_k B_k(x)$, where smoothness is imposed by penalizing m-th order differences of the coefficients $\boldsymbol{\beta}$:

$$S(\boldsymbol{\beta}, \lambda) = \|\mathbf{y} - \mathbf{B}\boldsymbol{\beta}\|_2^2 + \lambda \|\mathbf{D}_m \boldsymbol{\beta}\|_2^2,$$

with closed-form solution:

$$\hat{\boldsymbol{\beta}}(\lambda) = (\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{D}_m^\top \mathbf{D}_m)^{-1} \mathbf{B}^\top \mathbf{y}.$$

The smoothing parameter λ controls the effective degrees of freedom (EDF), given by trace($\mathbf{H}(\lambda)$) where $\mathbf{H}(\lambda)$ is the smoother matrix. Standard selection approaches include cross-validation or AIC/BIC.

While P-splines achieve smoothness through an explicit penalty term, in our Bayesian approach using Natural Splines, smoothness is primarily achieved through the prior distribution placed on the spline coefficients β_k.

4.2 Bayesian Formulation

In a fully Bayesian approach to spline regression, priors are placed on the spline coefficients $\boldsymbol{\beta}$ and other variance components. For P-splines, the penalty term has a Bayesian interpretation as a Gaussian prior on the differences of coefficients:

$$\mathbf{D}_m\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}), \qquad \lambda = \sigma^2/\tau^2.$$

Priors on σ^2, τ^2 (e.g., inverse-gamma, half-Cauchy) yield a fully Bayesian model [11, 8]. In our Natural Spline model, we place a prior directly on the coefficients β_k. Specifically, we use $\beta_k \sim \mathcal{N}(0, 1)$, which corresponds to setting $\sigma_\beta^2 = 1$. This choice provides regularization to help prevent overfitting while maintaining a relatively simple prior structure for the spline coefficients.

4.3 State-Space Extension with Spline Drift

We embed Natural Splines in a state-space model. The core idea is that the latent tire degradation, θ_n, evolves over time, and this evolution includes a smooth, non-linear component captured by splines, potentially alongside other systematic drift factors.

4.3.1 Model Structure

The observed standardized lap time is modeled as:

$$y_n^{\text{std}} \sim \text{Student-t}(\nu, \mu_{\text{driver}[n]} + \beta_{\text{compound}[n]} + \theta_n^{\text{std}}, \sigma),$$

with hierarchical priors for the baseline driver effects μ_{driver} and compound adjustments β_{compound} similar to those in the linear model (Chapter 3).

The latent degradation process θ_n^{std} evolves as follows. It resets to zero at the start of each new stint or for the first lap overall. For subsequent laps within a stint, its evolution is:

$$\theta_n^{\text{std}} = \theta_{n-1}^{\text{std}} + \text{drift}_n^{\text{spline}} + \eta_n, \quad \text{where } \eta_n \sim \mathcal{N}(0, \tau^2).$$

The term $\text{drift}_n^{\text{spline}}$ represents the systematic change in degradation at lap n and is defined

as:

$$\text{drift}_n^{\text{spline}} = \delta_{\text{driver}[n]} + \delta_{\text{compound}[n]} + \sum_{k=1}^{K} \beta_{\text{spline},k} B_k(x_n).$$

Here, $\delta_{\text{driver}[n]}$ and $\delta_{\text{compound}[n]}$ are hierarchical driver and compound-specific drift components, respectively, analogous to those in the linear model (defined with a global mean drift μ_δ for drivers and deviations for compounds). The term $\sum_{k=1}^{K} \beta_{\text{spline},k} B_k(x_n)$ is the contribution from the Natural Spline basis functions $B_k(x_n)$ evaluated at tire life x_n, with coefficients $\beta_{\text{spline},k}$. The random noise η_n is scaled by τ, i.e., $\eta_n = \tau z_{\eta,n}$ with $z_{\eta,n} \sim \mathcal{N}(0, 1)$.

In this formulation, x_n represents the tire life at lap n. It is important to note that for the spline basis functions $B_k(x_n)$ to be effective and numerically stable, the input x_n (tire life) is standardized or normalized before being passed to the basis functions. This often involves scaling x_n to a range like $[0, 1]$ based on the minimum and maximum observed tire life in the training dataset, or by subtracting the mean and dividing by the standard deviation. This preprocessing step ensures that the spline knots are appropriately distributed across the effective range of the predictor and improves the conditioning of the basis matrix.

Priors for $\beta_{\text{spline},k}$, variance components, and Student's t degrees of freedom follow standard weakly informative choices.

4.4 Stan Implementation

The Stan model includes the spline basis (`X_spline`), hierarchical parameters, and a drift term incorporating the spline. The likelihood uses a Student's t-distribution for robustness. Priors and transformed parameters reflect the structure above. Posterior inference is performed via HMC.

4.5 Application to F1 Data

Key preprocessing: standardizing lap times, constructing tire life covariates, encoding categorical predictors, building the Natural Spline basis matrix `X_spline` in R using the 'ns()' function.

Model fitting uses Stan, extracting posterior samples for all parameters, including

Natural Spline coefficients and latent states. Results are summarized by plotting. The estimated drift function $\delta(n)$ (Figure 4.1) illustrates the inferred rate of degradation over tire life, complete with uncertainty quantification. Similarly, the evolution of the latent degradation state θ_n^{std} is visualized in Figure 4.2, showing the cumulative effect on performance.

Priors for the hierarchical drift components (e.g., $\mu_\delta, \sigma_{\delta,\text{driver}}, \sigma_{\delta,\text{compound}}$), variance components (σ, τ), and Student's t degrees of freedom (ν) follow standard weakly informative choices, similar to those detailed in Chapter 3 and specified in the Stan code (e.g., HalfNormal for scales, Exponential for ν).

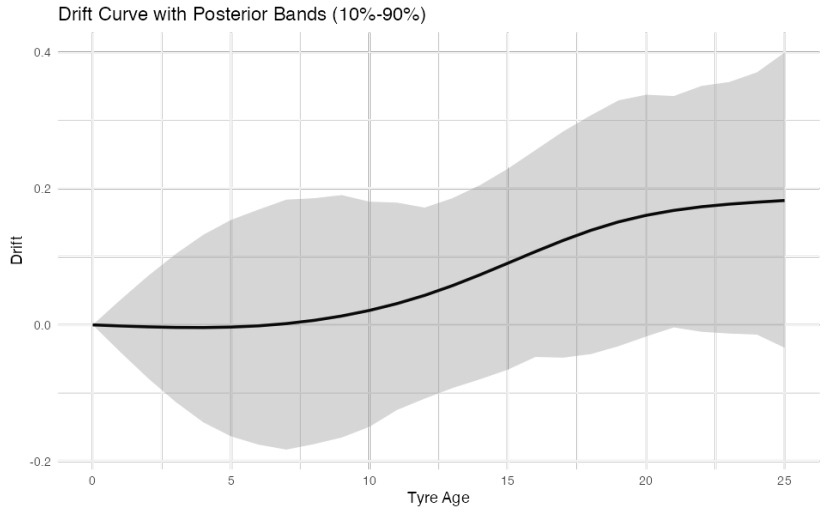

Figure 4.1: Estimated drift function $\delta(n)$ for a representative scenario, showing the mean posterior estimate and 95% credible bands. This highlights the non-linear nature of the degradation rate captured by the Natural Spline model.

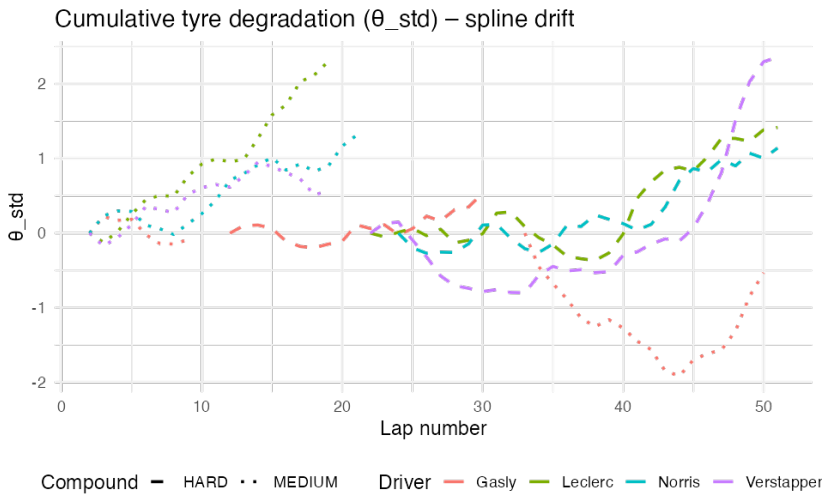

Figure 4.2: Fan chart illustrating the posterior distribution of the cumulative latent degradation θ_n^{std} over tire life. The widening bands reflect increasing uncertainty as predictions extend further into a stint.

Further visualizations include trace and density plots for spline coefficients to assess MCMC convergence and posterior distributions (e.g., Figure 4.3), and posterior predictive checks to evaluate model fit (Figure 4.4).

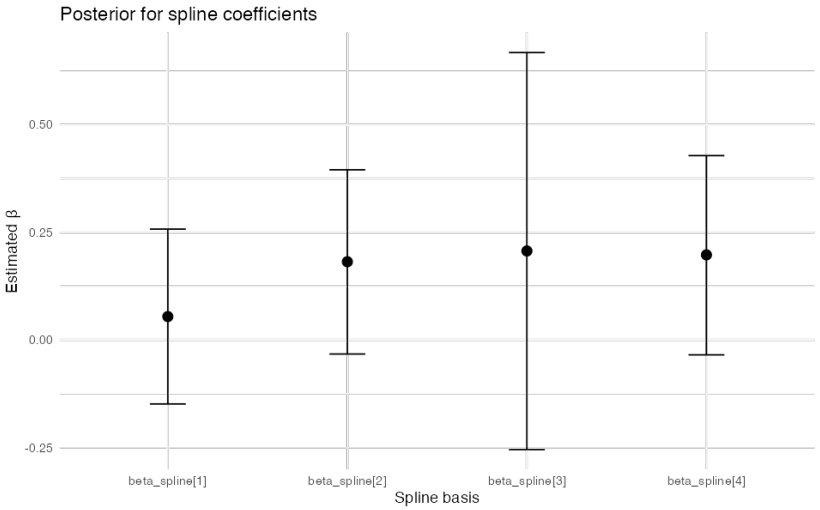

Figure 4.3: Posterior distributions of the Natural Spline basis coefficients ($\beta_{\text{spline},k}$). These coefficients determine the shape of the estimated drift function.

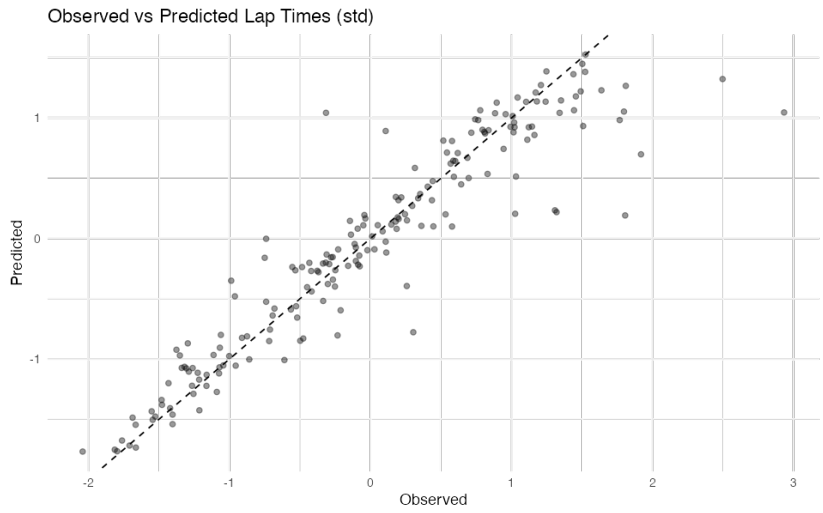

Figure 4.4: Posterior predictive check: scatter plot of observed versus predicted (replicated) lap times. Points clustering around the y=x line indicate good model calibration.

4.6 Diagnostics and Discussion

Diagnostics include:

- MCMC trace/density plots and $\hat{R}$.
- Posterior predictive checks (observed vs. predicted).
- Visual inspection of residuals and spline drift function.

The estimated drift reveals how tire degradation evolves, highlighting non-linearities and differences by compound and driver. The Natural Spline-based approach captures patterns missed by linear models and can inform F1 tire strategy.

4.7 Conclusion

Bayesian Natural Spline state-space models offer a powerful and flexible framework for capturing the complex, non-linear evolution of tire degradation in Formula 1. By leveraging a data-driven spline basis and Bayesian regularization, these models can adapt to a wide variety of degradation patterns—capturing subtle features such as initial conditioning, phases of stable wear, and rapid end-of-life drop-offs. The probabilistic nature of the approach allows for credible interval estimation and principled quantification of uncertainty, which is crucial for robust inference and interpretation.

This methodology enables a detailed reconstruction of latent degradation dynamics, providing interpretable insights into how tire performance evolves over a stint. The ability to model non-linearities and heterogeneity across drivers and compounds makes the Natural Spline approach particularly well-suited for the realities of modern motorsport, where tire behavior is influenced by a multitude of interacting factors. While the model's flexibility requires careful regularization and sufficient data, its advantages in terms of expressiveness and uncertainty quantification make it a valuable tool for advanced tire analysis in racing applications.

5. Conclusion and Future Work

5.1 Comparative Analysis of Tire Degradation Models

This thesis developed and evaluated two distinct Bayesian approaches for modeling Formula 1 tire degradation: a structured linear model (Chapter 3) and a more flexible Natural Spline based model (Chapter 4). Both aim to provide insights into lap time evolution due to tire wear, but they differ significantly in their assumptions, flexibility, and interpretability. Here, we compare their key characteristics:

Modeling Approach and Flexibility: The Linear Model employs a state-space formulation with a relatively rigid structure for the degradation process, a linear drift or a random walk. This makes it suitable for capturing general trends and monotonic degradation but less so for complex, non-linear patterns. In contrast, the Natural Spline model uses a basis of natural spline functions (B-splines with linear behavior beyond the boundary knots, generated with 'ns()' in R), offering a semi-parametric and highly flexible approach. It can adapt to non-linear degradation shapes, such as initial grip improvements followed by rapid drop-offs, by learning the functional form from the data and leveraging Bayesian regularization on the spline coefficients.

Interpretability: The parameters of the linear model, such as driver-specific baselines or average degradation rates, often have direct and intuitive interpretations. In the Natural Spline model, on the other hand, individual spline coefficients are not easily interpretable on their own, but the estimated degradation curve provides a clear visual representation of the wear pattern, although the overall model complexity is higher.

Predictive Accuracy: The linear model can achieve good predictive accuracy if the underlying degradation process is indeed close to a linear dynamic or follows the assumed stochastic process. The Natural Spline model, on the other hand, can offer superior accuracy when degradation patterns are non-linear, provided it is well regularized (for example, through appropriate priors) to avoid overfitting. Its data-driven nature allows it to capture nuances that a fixed-structure model might miss.

Computational Efficiency: The linear model is generally less computationally demanding and its simpler structure often leads to faster MCMC convergence. The Natural Spline model, instead, can be more demanding due to the larger number of parameters (the spline coefficients) and the need to estimate them.

Data Requirements: Both models benefit from a sufficient amount of data. The linear model can provide reasonable estimates even with moderately sized datasets thanks to its stronger assumptions. The Natural Spline model, on the other hand, requires more data to reliably estimate complex shapes and to ensure that the priors effectively prevent overfitting, especially if many knots (and thus many coefficients) are used.

Uncertainty Quantification: A key strength of both Bayesian approaches is their intrinsic ability to quantify uncertainty. Both models provide credible intervals for parameter estimates and predictions, offering a probabilistic view of tire degradation.

In summary, the choice between these models involves a trade-off. The linear model offers simplicity, interpretability, and computational speed, making it suitable for initial analyses or when degradation is expected to be fairly regular. The Natural Spline model provides superior flexibility to capture complex realities of tire wear, potentially at the cost of increased computational resources and a greater need for careful model specification and validation.

5.2 Future Research Directions

This section discusses potential extensions to the developed tire degradation models, focusing on enhancing model flexibility, predictive power, and real-time applicability. Key directions include more expressive nonparametric models like Gaussian Processes (GPs) and advanced sequential inference techniques such as Particle Filters (PF).

5.2.1 Gaussian Processes for Enhanced Degradation Modeling

A highly promising direction is to replace or augment the spline-based or linear degradation components with a *Gaussian Process* (GP) prior [12]. GPs offer a flexible Bayesian nonparametric approach, treating the degradation trajectory itself as a random function. The prior is specified by a mean function and a covariance kernel $k(n, n')$, which encodes assumptions about the degradation trajectory's smoothness, length-scale, and

other structural properties. This nonparametric flexibility allows GPs to capture complex, highly non-linear wear patterns that might be challenging for parametric or fixed-basis approaches. The marginal likelihood in GPs naturally balances model fit and complexity, aiding in hyperparameter learning.

Key extensions and considerations for GP-based degradation modeling include:

- **Hierarchical GPs:** To account for multi-level data structures (e.g., different drivers, tire compounds, tracks), hierarchical GP models can be employed. These models allow individual degradation curves for each context while sharing statistical strength through common hyperpriors on covariance parameters, improving estimates in data-sparse scenarios.

- **Sparse GP Approximations:** Standard GP inference scales as $\mathcal{O}(T^3)$ with T observations. For larger datasets (e.g., high-frequency telemetry or many stints), sparse GP approximations (e.g., FITC, variational inducing points) reduce complexity, often to $\mathcal{O}(Tm^2)$ for $m \ll T$ inducing points, making GPs scalable [12]. This is crucial for handling large data volumes or for real-time updates.

Adopting GPs could significantly improve the model's ability to adapt to nuanced degradation patterns while maintaining a fully probabilistic Bayesian interpretation.

5.2.2 Particle Filters (Sequential Monte Carlo) for Online Inference

To transition from offline analysis to real-time decision support, *Sequential Monte Carlo* (SMC) methods, also known as *Particle Filters* (PF), are essential [1, 4]. SMC methods perform Bayesian inference on state-space models as new observations arrive lap-by-lap. A state-space model for tire degradation would define a latent state x_k (degradation level at lap k), a state transition model $p(x_k \mid x_{k-1})$, and an observation model $p(y_k \mid x_k)$ linking the state to data y_k (e.g., lap time). SMC approximates the evolving posterior $p(x_k \mid y_{1:k})$ using a set of weighted samples (particles). The particles are propagated (predicted), weighted by the likelihood of the new observation, and resampled. This cycle provides a continually updated estimate of the tire's degradation state. Particle filters can handle non-linear and non-Gaussian models, crucial for realistic degradation dynamics. This capability would enable live tracking of tire condition, informing strategic decisions

like pit stops. Challenges include computational cost for real-time updates and potential particle degeneracy, requiring efficient implementation and careful model design.

5.2.3 Other Promising Directions

Beyond GPs and PFs, other enhancements could include:

- **Multivariate State Modeling with Telemetry Features:** Integrating diverse telemetry data (tire temperatures, pressures, g-forces) into a multivariate state-space model or as covariates. This could involve a latent state vector capturing not just "degradation" but also related factors like thermal state, with a joint likelihood for multiple data streams (lap times, temperatures). This holistic approach could lead to more accurate and robust inference of tire health by capturing the interplay between various physical factors.
- **Dynamic Model Parameters:** Allowing key model parameters (e.g., degradation rates, noise levels, or even GP kernel parameters) to vary over time or according to changing track conditions or driving styles.

5.3 Conclusion

The methodological avenues outlined—enhancing model flexibility with GPs, enabling real-time inference via Particle Filters, and broadening scope by incorporating more data sources—each target specific aspects of the tire degradation modeling problem. In combination, these advances could substantially elevate the fidelity and practicality of the models. For example, a hierarchical sparse GP model could learn generalizable degradation patterns across races, while a particle filtering system could use those patterns to swiftly adapt predictions during an ongoing race.

The ultimate implication for in-race strategy is significant: teams would be empowered with more accurate and up-to-date probabilistic estimates of tire health. This means better timing of pit stops (avoiding both untimely tire failures and overly conservative early stops), more informed tire compound choices under evolving conditions, and the ability to anticipate performance cliff points with greater confidence.

This thesis has demonstrated the value of Bayesian statistical modeling in dissecting the complex problem of tire degradation in motorsport. The comparison between the

linear and Natural Spline models highlights the trade-offs between structural simplicity and adaptive flexibility. By pursuing the future research directions discussed, particularly Gaussian Processes and Sequential Monte Carlo methods, the gap between off-line statistical analysis and on-line, actionable decision support can be effectively bridged. This will provide Formula 1 teams with an increasingly powerful Bayesian toolkit to navigate the ever-critical challenge of tire management, ultimately sharpening their competitive edge.

Bibliography

[1] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[2] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32, 2017.

[3] Carl de Boor. *A Practical Guide to Splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, revised edition, 2001.

[4] Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656–704), 2009.

[5] James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2nd edition, 2012.

[6] Paul H. C. Eilers and Brian D. Marx. *Practical Smoothing: The Joys of P-splines*. Cambridge University Press, 2021.

[7] Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

[8] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression: Models, Methods and Applications*. Springer Science & Business Media, 2013.

[9] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2013.

[10] Rudolf E Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[11] Sergei Lang and Andreas Brezger. Bayesian p-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004.

[12] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[13] Christian P Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Science & Business Media, 2nd edition, 2007.

[14] Andrea Stella. Stella on mclaren's mastery of tyre management after f1 miami gp. Pit Debrief, 2024. Accessed: June 16, 2024.

[15] FastF1 Development Team. Fastf1: A python library for formula 1 data analysis, 2023.

[16] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, 2017.