

UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

An Empirical Bayes Approach to Nonparametric Regression with Correlated Errors

Permalink

<https://escholarship.org/uc/item/50c923qg>

ISBN

9798186386698

Author

Marzban, Ethan P

Publication Date

2026-06-14

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

An Empirical Bayes Approach to Nonparametric Regression with Correlated Errors

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Statistics and Applied Probability

by

Ethan P. Marzban

Committee in charge:

Professor Andrew V. Carter, Chair
Professor Mengyang Gu
Professor Yuedong Wang

June 2026

The Dissertation of Ethan P. Marzban is approved.

Professor Mengyang Gu

Professor Yuedong Wang

Professor Andrew V. Carter, Committee Chair

June 2026

An Empirical Bayes Approach to Nonparametric Regression with Correlated Errors

Copyright © 2026

by

Ethan P. Marzban

Dedicated to my friends and family.

Acknowledgements

I am deeply indebted to the tutelage and support of a multitude of people; far too many to list in their entirety. To anyone whose name is not listed below: I apologize, and assure no slight was intended.

First and foremost, I thank my thesis advisor Dr. Carter - without his guidance, and immense patience, I am certain I would not have been able to complete this journey. Dr. Wang and Dr. Gu were also instrumental in my success, having taught me practically everything I know about Gaussian Processes and Applied Statistics. Dr. Ravat and Dr. Miller shaped me into the educator I am today, and I am forever grateful for their guidance as mentors.

I would also like to thank my friends and colleagues, for all the sleepless nights spent studying for exams, and all their emotional support. Erika (and her cats, Bean and Shiro), Karina, and Blythe (and Arlo) have been by my side since the very beginning, when our paths crossed during a very “normal” start to our graduate careers in 2020. Caroline and Megan provided advice and support that is truly immeasurable, especially as I navigated the uncertainty of my early stages in the department. To the most recent iteration of the Grad Student Committee (Mallory, Denisse, Jen, and Summer), thank you for your dedication and friendship. The incredible faculty and staff of the PSTAT department have also been pillars of my success (with a very special thanks to Myranda). My cat, Kitty, has dedicated numerous hours of emotional labor to my well-being, and I have no doubt my mental state would have suffered significantly if it weren’t for her.

Finally, but most importantly, I would be nothing without my family - including those who are no longer with us - as they mean the world to me. Thank you for your unconditional love and support.

Curriculum Vitæ

Ethan P. Marzban

Education

2026	Ph.D. in Statistics (Expected), University of California, Santa Barbara.
2022	M.A. in Statistics, University of California, Santa Barbara.
2020	B.A in Statistics, University of California, Berkeley

Past Employment

- **Teaching Associate:** Department of Statistics and Applied Probability at the University of California, Santa Barbara
- **Teaching Assistant:** Department of Statistics and Applied Probability at the University of California, Santa Barbara
- **Lead Teaching Assistant:** Department of Statistics and Applied Probability at the University of California, Santa Barbara

Future Employment

- **Assistant Professor:** Department of Statistics, California Polytechnic State University (Fall 2026 start date)

Service

- **Officer and Founding Member:** Graduate Student Committee; Department of Statistics and Applied Probability at the University of California, Santa Barbara
- **President and Founding Member:** Santa Barbara Statistics Initiative (a registered chapter of the American Statistical Association); University of California, Santa Barbara

Abstract

An Empirical Bayes Approach to Nonparametric Regression with Correlated Errors

by

Ethan P. Marzban

Nonparametric regression has been well established as a flexible and applicable approach to regression. Many nonparametric regression techniques, however, break down in the presence of correlated noise, leading to fits that are far too noisy. In this thesis we develop an estimation technique for nonparametric regression that can be extended to cases where the noise is correlated. In the first chapter, we review relevant prerequisites from nonparametric regression, Gaussian Processes, and the spectral domain. We also include an overview of relevant literature. In the second chapter, we propose our estimation procedure, which is rooted in an Empirical Bayes framework. Hyperparameters are estimated in the spectral domain using a Generalized Linear Model, which allows our technique to be extended to problems where noises are correlated. We also demonstrate the effectiveness of the estimation procedure on several non-simulated datasets, also demonstrating that our estimation procedure works well even when the signal function possesses a noticeable trend. Simulated data also reveals that our estimation procedure performs comparably to already-existing corrections to nonparametric regression for correlated errors. In the third chapter, we construct approximate credible intervals for the approval ratings dataset, along with a dataset tracking cases of COVID-19

recorded in Santa Barbara county. Finally, in the fourth chapter, we calculate the rate at which our estimates concentrate to the truth, demonstrating that our estimators possess a certain statistical consistency.

Contents

Curriculum Vitae	vii
Abstract	viii
1 Preliminaries	1
1.1 Introduction	1
1.2 Gaussian Processes	2
1.3 Spectral Domain	5
1.4 Prior Work	7
2 Description of the Estimation Procedure	11
2.1 Description of the Estimation Procedure	11
2.2 Simulations	19
2.3 Applications to Data	22
3 Credible Intervals	30
3.1 Derivation for MA(2) Noise	31
3.2 Extensions to Higher-Order Noise	40
4 Asymptotics	42
4.1 Asymptotic Rates	42
4.2 White Noise	43
5 Conclusion	48
5.1 Future Work	49
5.2 General Disclosures	50
A Derivative Calculations for Credible Intervals	51

B Bias-Variance Decomposition for MSE	55
Bibliography	57

CHAPTER 1

Preliminaries

1.1 Introduction

Given equally-spaced univariate data $\mathcal{D} := \{(t, y_t)\}_{t=1}^n$ on a response variable y , a typical univariate nonparametric regression model posits

$$y_t = f(t/n) + \omega_t \tag{1.1}$$

for some unknown signal function $f : \mathbb{R} \rightarrow \mathbb{R}$ and noise process $\{\omega_t\}_{t=1}^n$. It is common to assert that f maps from the unit interval $[0, 1]$, which is why the argument to f has been specified as (t/n) in Equation (1.1). A typical nonparametric regression problem seeks to estimate the true signal function that gave rise to a particular set of data $\mathcal{D}$, under as few assumptions as possible. One common assumption is that of independent and identically distributed (IID) Gaussian Noise:

$$\omega_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

where $\sigma^2 > 0$ is a deterministic (but unknown) “**noise variance**” parameter. In many practical situations, however, this assumption is overly restrictive. Furthermore, when incorrectly applied, the assumption of independent (or even uncorrelated) errors may lead to poor estimates. In this dissertation, we adopt a more general assumption for the noise term, modeling it as an MA(q) (Moving Average of Order q) process:

$$\omega_t = \varepsilon_t + \sum_{j=1}^q \psi_j \varepsilon_{t-j} \quad \text{where} \quad \varepsilon_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1.2)$$

for deterministic moving average coefficients $\{\psi_j\}_{j=1}^q$.

In this chapter of the dissertation, we review relevant prerequisite material, particularly on Gaussian Processes (Section 1.2) and the spectral domain (Section 1.3). Section (1.4) provides an overview of some relevant prior work.

1.2 Gaussian Processes

Definition 1.2.1 (Gaussian Process). A stochastic process $\{X_t\}_{t \in \mathcal{T}}$ is defined to be a **Gaussian Process** if the finite dimensional distributions of $\{X_t\}$ are multivariate Gaussian [1]. In other words, for any $k \in \mathbb{N}$ and any subcollection of indices $(n_1, \dots, n_k) \subseteq \mathcal{T}$, the random vector $(X_{n_1}, \dots, X_{n_k})$ follows a multivariate Gaussian distribution.

Note that a Gaussian Process, in essence, induces a distribution from which draws are *functions*. That is, if a random function $f(\cdot)$ is such that for any collection of inputs $(t_1, \dots, t_n)$ the random vector $(f(t_1), \dots, f(t_n))$ follows a multivariate

normal distribution, then we may say that f itself follows a Gaussian Process.

As the multivariate normal distribution is parametrized by a mean vector and covariance matrix, so too is a Gaussian Process parametrized by a **mean function** $\mu(\cdot)$ and **covariance kernel** $K(\cdot, \cdot)$. We shall use the notation

$$f \sim \mathcal{GP}(\mu(t), K(s, t))$$

to mean that the function $f(\cdot)$ follows a Gaussian Process with mean function $\mu(\cdot)$ (often assumed to be identically zero) and covariance kernel $K(\cdot, \cdot)$.

The mean function controls the “trend” of resulting draws, and the covariance kernel controls the “smoothness” of resulting draws. Specifically, the covariance kernel determines the strength of correlation between outputs at any two inputs. A popular assumption for the covariance kernel is that of **stationarity**, which asserts that there exists a function C such that

$$K(s, t) = C(s - t)$$

that is; the dependence between $f(t)$ and $f(s)$ for two inputs s and t can be expressed as a function of the difference $(s - t)$ in inputs. One specific class of stationary covariance kernel is that of the **Squared-Exponential**, which admits a single hyperparameter h (often called the **lengthscale**, but what we will call a **bandwidth**):

$$C(r) = \exp \left\{ -\frac{r^2}{2h} \right\} \quad (1.3)$$

Another popular choice of stationary covariance kernels is that of the **Matérn** ker-

nel [2]

$$C_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \cdot \left(\frac{r\sqrt{2\nu}}{h} \right)^\nu \cdot K_\nu \left(\frac{r\sqrt{2\nu}}{h} \right) \quad (1.4)$$

where $K_\nu(\cdot)$ denotes a modified Bessel K function. Taking $\nu = 5/2$ allows for a simpler closed-form expression for the kernel [2]:

$$C_{\nu=5/2}(r) = \left(1 + \frac{r\sqrt{5}}{h} + \frac{5r^2}{3h^2} \right) \cdot \exp \left\{ -\frac{r\sqrt{5}}{h} \right\} \quad (1.5)$$

Figure (1.1) illustrates several sets of draws from Gaussian Processes with zero mean (i.e. $\mu \equiv 0$), using both the Squared-Exponential and Matérn-5/2 kernels. This figure, in addition to illustrating what sample draws from a Gaussian Process look like, also demonstrates how different choices for the **hyperparameter** values in each covariance kernel affect the overall shape of the resulting draws.

Indeed, as h increases in value the resulting draws become smoother - hence our rationale for calling it a bandwidth parameter. It is perhaps pertinent to note, however, that draws using the Matérn-5/2 kernel are in general rougher than their counterparts using the Squared-Exponential; for example, draws from a GP with Squared-Exponential kernel with $h = 0.1$ are smoother than draws from a GP with Matérn-5/2 kernel with $h = 0.1$. In fact, as $\nu \rightarrow \infty$, the Matérn covariance kernel converges to the Squared-Exponential kernel [2].

There are, of course, many other classes of covariance kernels, each used in different settings. Examples of such kernels can be found, for instance, in Chapter 4 of [2] and Chapter 1 of [3].

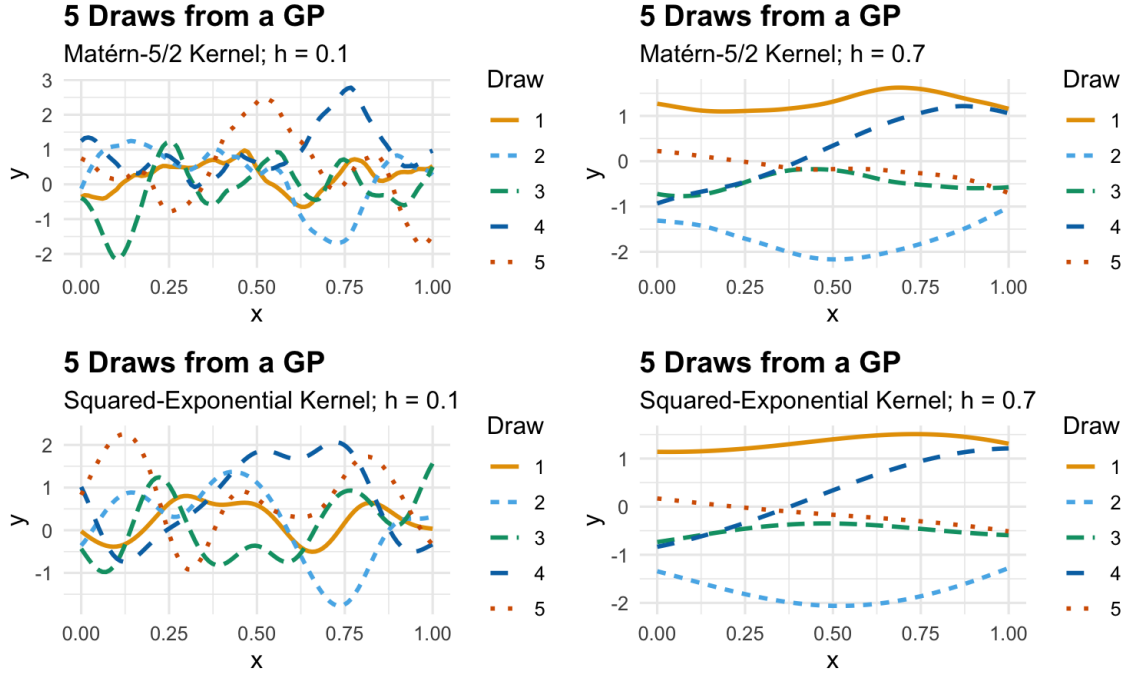


Figure 1.1: **Top:** Two sets of five draws from a Squared-Exponential Kernel. Left: $h = 0.1$. Right: $h = 0.7$. **Bottom:** Two sets of five draws from a Matérn-5/2 kernel. Left: $h = 0.1$. Right: $h = 0.7$.

1.3 Spectral Domain

Utilizing the spectral domain is appealing for a number of reasons. For one, if we assume appropriate stationarity then the Fourier coefficients of the signal and noise processes will be independent. Additionally, as Fourier transforms are linear, we maintain a preservation of separation between signal- and noise-related terms, allowing for easier estimation of these components. To that end, it may be pertinent to review some spectral results relating to covariance kernels. First, we review **Bochner's Theorem**:

Theorem 1.3.1 (Bochner's Theorem). A function $K : \mathbb{R} \rightarrow \mathbb{C}$ is the covariance

function of a stationary mean-square continuous process if and only if it can be represented as

$$K(r) := \int_{\mathbb{R}} e^{2\pi i sr} d\lambda(s) \quad (1.6)$$

where λ is a positive finite measure [2].

For the purposes of this dissertation, we will use the terms “kernel” and “covariance kernel” interchangeably. If λ has a density $S(s)$, then S is referred to as the **spectral density** of $K(\cdot)$ and the Wiener-Khintchine theorem asserts that the kernel k and the spectral density S are Fourier duals of one another [see, for instance, pg. 82 of [2]].

As an example, the spectral density of the Squared-Exponential kernel with lengthscale parameter h is given by

$$S(s) = \sqrt{2\pi h^2} \exp \left\{ -2\pi^2 h^2 s^2 \right\} \quad (1.7)$$

which is proportional to the density of a Normal distribution. The spectral density of the Matérn kernel with roughness parameter ν and lengthscale parameter h is given by

$$S(s) = \frac{2\sqrt{\pi} \cdot \Gamma\left(\nu + \frac{1}{2}\right) (2\nu)^\nu}{\Gamma(\nu) h^{2\nu}} \cdot \left(\frac{2\nu}{h^2} + 4\pi^2 s^2 \right)^{-(\nu + \frac{1}{2})} \quad (1.8)$$

(see page 84 of [2]) which is proportional to the density of a Student’s t –distribution.

Since there is a direct correspondence between a stationary covariance kernel and its spectral density, we can use the spectral density to simulate draws from a Gaussian Process with a desired covariance structure. For example, one can first generate the Fourier coefficients as complex-normal random variables with vari-

ances proportional to the spectral density evaluated at relevant Fourier frequencies, and then apply the inverse-Fourier Transform.

1.4 Prior Work

The use of Gaussian Processes in spatial and spatio-temporal statistics is well documented; see, for instance, [4]. In [5], Gaussian Processes are used to impute previously unobserved DNA methylation levels at various sites across the human genome. The authors leverage the fact that a Gaussian Process with Matérn covariance kernel is the solution to a particular stochastic differential equation, one that is expressible as a state space model that can be solved using Kálmán filtering. The computational cost of estimating hyperparameters is alleviated by utilizing a clever choice of “robust” prior. The setting of [5], however, is highly spatial as much of the algorithm relies on leveraging the correlations between inputs. We are, conversely, interested in a much simpler setting in which we have uncorrelated inputs. Furthermore, the authors adopt a fully Bayesian paradigm by imposing a specific robust prior on the hyperparameters. In this dissertation, we adopt an Empirical Bayes approach, opting to estimate hyperparameters using maximum likelihood estimation.

Of particular interest to us are considerations in the spectral domain. One early reference to this can be found in [6], in which the author considers a more general semiparametric model that incorporates random-effect terms with an unknown signal function. The signal function is estimated by considering its Fourier series expansion, with normally-distributed Fourier coefficients. In a marked departure

from our work, however, the author performs the estimation in a fully Bayesian paradigm and assigns hyperpriors to the hyperparameters. As has been mentioned above, our approach, in contrast, adopts an Empirical Bayesian paradigm and bypasses the assignment of hyperpriors by instead estimating hyperparameters by maximizing the marginal likelihood.

A slightly more modern approach involving the spectral decomposition of a GP can be found in [7], in which the authors consider an expansion of the unknown signal function in terms of sines and cosines with normally-distributed coefficients. To circumvent the costly matrix inversion, they propose an algorithm (called the Sparse Spectrum Gaussian Process [SSGP] algorithm) which selects optimal Fourier frequency so as to impose sparsity in the spectral domain, thereby significantly decreasing the computational cost of obtaining predictive values. Selection of the optimal frequencies is done by sampling the spectral density of the kernel function using Monte Carlo simulation.

Similar considerations have been made with regards to reduced-rank approximations to Gaussian Processes, as in [8]. The authors argue that, under appropriate conditions, the covariance kernel of a Gaussian Process can be well-approximated using a series expansion with respect to the eigenfunctions of the kernel, which in turn yields an expansion for the entire signal function. The authors are also able to provide convergence rates for the approximation. Though our method also considers a series expansion of the signal function, the authors of [8] consider a specific set of eigenfunctions based on the Laplace operator in a compact subset of $\mathbb{R}^d$ whereas, in contrast, our estimation procedure utilizes a Fourier set of basis functions (i.e. complex sines and cosines). Additionally, the reduced-rank approxima-

tion that the authors produce ends up being independent of the hyperparameters, whereas a large portion of our estimation procedure is dedicated to the estimation of the hyperparameters in the model (e.g. the signal and noise variances).

The authors of [8] also make reference to several “inducing-point” methods, such as the Nyström method, in which a set of “inducing” inputs are chosen and the eigendecomposition of their corresponding covariance matrix is transformed to match that of the actual covariance. This differs significantly from the setting of our procedure, in which we only utilize inputs for which there exists an observed response value (except, perhaps, in cases of missing data; Chapter 3 briefly discusses the handling of missing values in our estimation procedure).

In many situations the structure of the kernel is assumed to be known (e.g. Squared-Exponential, Matérn-5/2, etc.) Some work has been devoted to searching for the optimal form of the kernel; for instance, [9] outlines an algorithm that searches through a space of kernels (constructed as compositions of a handful of basis kernels) to identify the kernel that best fits the data along with any relevant hyperparameters. In our considerations, however, we will assume that the form of the kernel is known *a priori*, though we may select a kernel from the space of kernels the authors in [9] constructed.

Much of our work is motivated by the need for a nonparametric regression technique that works in the face of correlated errors; indeed, [10] remains a seminal work in this topic. In it, the authors consider smoothing splines as a nonparametric estimation technique. Smoothing splines, like many nonparametric regression techniques, require the selection of a smoothing (i.e. bandwidth) parameter, one that is often selected using cross-validation (CV) or generalized cross-validation

(GCV). In [10], the authors propose a modification to the GCV-optimal bandwidth selected under independent errors that achieves the same optimal rate ($n^{-1/5}$) as that of the case of independent errors. This modification does depend on the (unknown) correlation matrix; the authors propose a plug-in estimator for this correlation matrix term which can be used in practice. Though the initial setup of the situation discussed in this paper is similar to the setup we consider in this dissertation, the authors' approach differs somewhat significantly from ours; firstly, in that we bypass the use of smoothing splines (opting instead to use Gaussian Process regression), and secondly in that we leverage an explicitly Empirical Bayes formulation that, we believe, is easier to implement in practice and lends itself more readily to posterior uncertainty quantification.

CHAPTER 2

Description of the Estimation Procedure

In this chapter, we outline the details of our estimation procedure (Section 2.1) and demonstrate its applications to simulated (Section 2.2) and non-simulated (2.3) data.

2.1 Description of the Estimation Procedure

2.1.1 Specification of the Prior

Our first task is to select an appropriate prior for the signal function in Equation (1.1). We assume that f admits a Fourier decomposition of the form

$$f(t/n) = \sum_{k=0}^{n-1} \theta_k e^{-2\pi i k t/n} \quad (2.1)$$

Crucially, the θ_k are complex-valued; the fact that f is real-valued implies $\theta_k = \theta_{n-k}^*$. Adopting a Gaussian Process prior with stationary covariance kernel further implies that the θ_k will be independent zero-mean complex-normally distributed

random variables with variances v_k where $v_k \rightarrow 0$ as $k \rightarrow \infty$.

We assume the v_k have the form

$$v_k = \tau^2 h \left(1 + \frac{2\pi^2}{\nu} k^2 h^2 \right)^{-(\nu+1/2)} \quad (2.2)$$

so we have terms that match the spectrum of the Matérn kernel with roughness parameter ν and bandwidth h . Following [2], we assume ν to be known; for instance, $\nu = 5/2$. We therefore arrive at the following prior specification for the Fourier coefficients of the signal function:

$$\theta_k \stackrel{\perp}{\sim} \mathcal{CN}(\vec{0}, v_k \mathbf{I}_2) \quad (2.3)$$

where $\mathbf{I}_2$ denotes the 2-dimensional identity matrix and v_k is as in Equation (2.2). A notational note: $\mathcal{CN}$ denotes the **complex-normal** distribution, and we write (2.3) and subsequently similar notation to mean $\theta_k = \theta_{k,\text{Re}} + i\theta_{k,\text{Im}}$ with $\theta_{k,\text{Re}}$ and $\theta_{k,\text{Im}}$ independent, and $\theta_{k,\text{Re}} \sim \mathcal{N}(0, v_k)$ and $\theta_{k,\text{Im}} \sim \mathcal{N}(0, v_k)$.

Because the Gaussian Process with Matérn kernel places all of its probability on elements of a Sobolev space, we impose the assumption that f belongs to a Sobolev space of order $\alpha < \nu$ (where ν is the roughness parameter of the underlying Matérn covariance kernel). In addition to imposing smoothness constraints on the signal function $f()$, such an imposition also provides constraints on the behavior of v_k that are needed to establish the asymptotic properties of our estimator (considered in Chapter 4). The specification of the order of the Sobolev space is as follows: by

[11] and [12], asserting that f belongs to a Sobolev Space of order α ensures

$$\sum_k (1 + k^2)^\alpha \|\check{f}_k\|^2 < \infty \quad (2.4)$$

where $\|\check{f}_k\|^2$ denotes the squared-modulus of the k^{th} Discrete Fourier coefficient of f . Since $\mathbb{P}\{\sum_k (1 + k^2)^\alpha \check{f}_k < \infty\} = 1$, for sufficiently large k we have that $\mathbb{E}[\check{f}_k] = 0$ and $\text{Var}(\check{f}_k) = v_k$, so that

$$\mathbb{E} \left[\sum_k (1 + k^2)^\alpha \|\check{f}_k\|^2 \right] = \sum_k (1 + k^2)^\alpha (1 + k^2)^{-(\nu + \frac{1}{2})} = \sum_k (1 + k^2)^{-\alpha - (\nu + \frac{1}{2})}$$

In order for this sum to converge (and, consequently, to have a finite variance), we must therefore select

$$\alpha - \left(\nu + \frac{1}{2} \right) < -\frac{1}{2}$$

which is equivalent to taking $\alpha < \nu$. Therefore, we require the smoothness of the Sobolev Space to be less than the roughness parameter of the Matérn-5/2 kernel.

2.1.2 Specification of the Model

Assuming the noise process ω_t in (1.1) is stationary, it too will admit a Fourier decomposition:

$$\omega_t = \sum_{k=0}^{n-1} \xi_k e^{-2\pi i k t / n} \quad (2.5)$$

where the ξ_k are complex-valued, independent across k with variance $u_k := \text{Var}(\xi_k)$. Specifically, we assert

$$u_k = \frac{\sigma^2}{n} \left[b_0 + \sum_{j=1}^q b_j \cos \left(j \cdot \frac{2\pi k}{n} \right) \right] \quad (2.6)$$

which matches the spectrum of a MA(q) process (i.e. a moving average process of order q); see, for instance, [13]. Note that the b_j coefficients are nonlinear combinations of the original MA coefficients ψ_j in Equation (1.2). The exact correspondence between the b_j and ψ_j can be found, again, in [13]. A traditional white noise setting can be recovered from this choice of variance as well, by taking $b_j = 0$ for any $j \geq 1$.

Assuming normality, we then have

$$\xi_k \stackrel{\perp}{\sim} \mathcal{CN}(\vec{0}, u_k \mathbf{I}_2) \quad (2.7)$$

Letting $\eta_k := (\theta_k + \xi_k)$, this, combined with equations (1.1) and (2.3), yield the following Bayesian model for the Fourier coefficients:

$$\eta_k \mid \theta_k \stackrel{\perp}{\sim} \mathcal{CN}(\theta_k, u_k \mathbf{I}_2) \quad (2.8)$$

$$\theta_k \stackrel{\perp}{\sim} \mathcal{CN}(0, v_k \mathbf{I}_2) \quad (2.9)$$

which the reader may recognize as a variant of a traditional “Normal-normal” problem from Bayesian literature (e.g. [14]). A natural estimator $\hat{\theta}_k$ for θ_k is the posterior mean; indeed, if all hyperparameters are specified correctly, this estimate will minimize the Bayes Risk. Through a standard computation (see, for instance,

page 40 of [14]), we have

$$\hat{\theta}'_k := \mathbb{E}[\theta_k \mid \eta_k] = \left(\frac{v_k}{v_k + u_k} \right) \eta_k \quad (2.10)$$

meaning an estimator $\hat{f}(\cdot)$ for $f(\cdot)$ can be obtained by taking the inverse Fourier transform of the above $\hat{\theta}'_k$ values. Since Fourier transforms are linear, $\eta_k := (\theta_k + \xi_k)$ denotes the true Fourier coefficients of the full (signal-plus-noise) process. As such, we have the following interpretation of (2.10): the Fourier coefficients of our estimator $\hat{f}(\cdot)$ are constructed by weighting (shrinking) the Fourier coefficients of the full (signal-plus-noise) process.

2.1.3 Estimation of Hyperparameters

Unfortunately, Equation (2.10) cannot be used as an estimator in practice as it depends on several unobserved quantities and hyperparameters.

First there is the matter that θ_k and ξ_k are themselves unknown, as they represent the *true* (theoretical) Fourier coefficients of the signal and noise processes, respectively. However, note that in Equation (2.10) these coefficients only appear as their sum, η_k . Since η_k denotes the “true” (theoretical) Fourier coefficient of $f(t/n) + \omega_t$, and since $y_t = f(t/n) + \omega_t$, a realization of η_k is provided by with $\check{y}_k$, the Discrete Fourier Transform values of the observed response values. Hence, in Equation (2.10) we may replace η_k with $\check{y}_k$. If the input values are regularly-spaced, $\check{y}_k$ can be computed using the Fast Fourier Transform (FFT).

This leaves the following hyperparameters that must be estimated: the bandwidth h (present in v_k), the signal variance τ^2 (also present in v_k), and the noise-

related coefficients (e.g. the noise variance σ^2 and any potential spectral coefficients related to the MA noise process, all present in u_k). A purely Bayesian estimation procedure would assign hyperpriors to these quantities (or perhaps a single joint hyperprior, as in [15]); instead, we adopt an Empirical Bayes approach and seek to estimate these hyperparameters using maximum likelihood estimation. More specifically, we maximize the marginal likelihood. Admittedly, our estimation procedure is perhaps unusual even from an Empirical Bayes viewpoint in that we estimate both prior components (the bandwidth and noise variance) and *model-based* components (the signal variance). Nevertheless, our utilization of data to estimate these parameters (as opposed to the specification of hyperpriors), we believe, casts our estimation in an Empirical Bayes framework.

As mentioned in Chapter (1) above, we consider estimating the parameters of interest in the spectral domain. Equations (2.3) and (2.7) together imply that

$$\eta_k := (\theta_k + \zeta_k) \stackrel{\perp}{\sim} \mathcal{CN}(0, (v_k + u_k)\mathbf{I}_2)$$

which in turn implies

$$\frac{1}{2}\|\eta_k\|^2 \sim \text{Exponential}(v_k + u_k)$$

or, plugging in the expressions for u_k and v_k from (2.6) and (2.2), respectively, and setting $\nu = 5/2$,

$$\frac{1}{2}\|\eta_k\|^2 \sim \text{Exponential} \left[\tau^2 h \left(1 + k^2 h^2 \right)^{-3} + \frac{\sigma^2}{n} b_0 + \sum_{j=1}^q \frac{\sigma^2}{n} b_j \cos \left(j \cdot \frac{2\pi k}{n} \right) \right] \quad (2.11)$$

where the b_j are nonlinear combinations of the original MA coefficients. This motivates us to estimate v_k and u_k by fitting a Gamma Generalized Linear Model (GLM) of the form

$$\frac{1}{2} \|\check{y}_k\|^2 \sim \text{Exponential} \left[\beta_1 \left(1 + k^2 h^2\right)^{-3} + \beta_0 + \sum_{j=1}^q \beta_{j+2} \cos \left(j \cdot \frac{2\pi k}{n} \right) \right] \quad (2.12)$$

In this way, an estimator $\hat{\beta}_0$ provides an estimator for σ^2 , an estimator $\hat{\beta}_1$ provides an estimator for the signal variance τ^2 , and the estimators $\hat{\beta}_j$ provide estimators for the coefficients of the MA noise process. Together, these can be used to construct estimators $\hat{v}_k$ and $\hat{u}_k$, which can be substituted into (2.10).

Finally, note that the GLM specified in (2.12) depends on the bandwidth parameter h , which itself is unknown. It is natural to estimate h by way of maximum likelihood estimation; in practice, due to the highly nonlinear relationship between h and the likelihood, this estimation does not admit a closed-form solution. As such, an iterative solver is needed to carry out the optimization. One option is to utilize a grid-search, searching over a set of possible candidate bandwidth values and identifying the one that results in a model with minimum deviance. In practice, such a grid-search works well, as demonstrated in the next section. More advanced optimization techniques could be utilized instead; for example, a Newton-Raphson algorithm could be used to simultaneously estimate all parameters in the GLM (including the bandwidth). Admittedly, both approaches can suffer from numerical instability; future work could certainly be dedicated to developing a more efficient and computationally robust optimization procedure.

2.1.4 The Full Fitting Algorithm

With the results of the previous two sections in mind, we propose our estimation scheme in Algorithm (1).

Algorithm 1: Fitting a GP to Data with $MA(q)$ Noise using Fourier Coefficients

1. Obtain the coefficients $\check{y}_k$ of the Discrete Fourier Transform of the y_t values, and use this to compute $(1/2)\|\check{y}_k\|^2$ for $k = 0, \dots, B$ where $B = (n-1)/2$ if n is odd and $B = n/2$ if n is even.
2. Identify the bandwidth $\hat{h}$ that minimizes the deviance of the following Gamma Generalized Linear Model:

$$\frac{1}{2}\|\check{y}_k\|^2 \sim \text{Exponential} \left[\beta_1 \left(1 + k^2 h^2 \right)^{-(\nu + \frac{1}{2})} + \beta_0 + \sum_{j=1}^q \beta_{j+2} \cos \left(j \cdot \frac{2\pi k}{n} \right) \right]$$

3. Utilizing the optimal bandwidth from step (2) above, refit the Gamma GLM and obtain fitted values $\hat{\alpha}_k, k = 1, \dots, B$ for $\alpha_k := (1/2)\|\eta_k\|^2$. Estimates for v_k are computed as

$$\hat{v}_k := \hat{\alpha}_k - \hat{\beta}_0 - \sum_{j=1}^q \hat{\beta}_{j+2} \cos(2\pi jk/n)$$

4. Obtain estimates $\hat{\theta}_k$ for the θ_k values using $\hat{\theta}_k := w_k \check{y}_k$ where $w_0 := 1$ and, for $k \geq 1$,

$$\hat{w}_k := \left(\frac{\widehat{v_k}}{u_k + v_k} \right) = \left(\frac{\hat{v}_k}{\hat{\alpha}_k} \right)$$

5. Apply the inverse Fourier transform to the $\hat{\theta}_k$ values obtained in step (4) above to obtain fitted values $\hat{y}$.
-

2.2 Simulations

To illustrate the effectiveness of our estimation procedure, we first consider simulated data. As a simple benchmark against which to compare our estimates we use a smoothing spline via the `assist` package in R. Specifically, this allows us to implement the corrections outlined in [10] to account for the fact that we have correlated errors.

We generated 2,000 simulated datasets according to model (1.1) using a Gaussian Process prior with Matérn-5/2 kernel and MA(2) noise. To each dataset, we fit four models: white noise (i.e. ignoring the noise correlation), MA(2) noise, MA(4) noise (i.e. overfitting), and correlation-adjusted smoothing splines (using a correction for an MA(2) process). The White Noise, MA(2), and MA(4) fits were generated using Algorithm (1). Shown below are a series of plots, illustrating different findings from the simulations. Simulations were conducted on a 2023 MacBook Pro with Apple M2 Pro chip, and had a reported runtime of around 42.3 minutes (2540.996 seconds).

First, in figure (2.1) we display a side-by-side boxplot of the log-MSEs from the different fits. What we see is that the white noise fit results in drastically higher (and more variable) MSEs than the other three methods. This result is to be expected; as has been mentioned previously, fully ignoring the correlation structure of a dataset can lead to disastrously overfitted results.

In many ways, however, the data in Figure (2.1) are “paired,” since each of the 2,000 simulations gave rise to an MSE from each box. To that end, it is perhaps more statistically sound to consider scatterplots that compare the various MSEs. Such

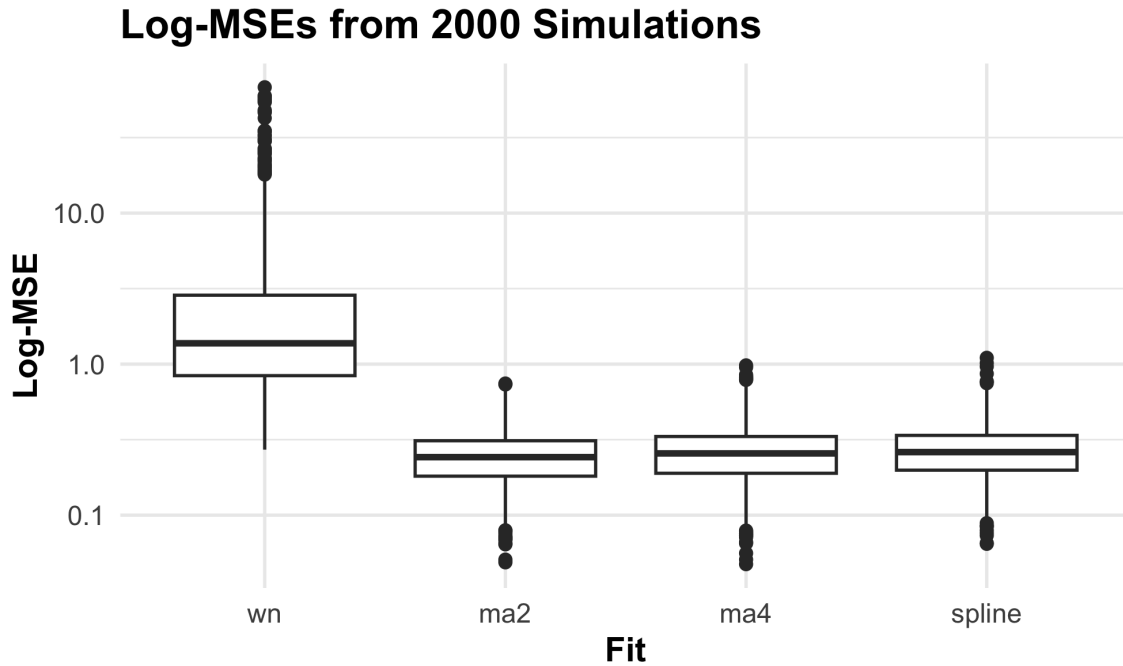


Figure 2.1: Log-MSEs obtained from White Noise, MA(2), and MA(4) fits using algorithm (wn, ma2, and ma4, respectively) along with those resulting from correlation-adjusted smoothing splines (splines). Spline fits are obtained from the `ssr()` function in the `assist` library, specifying an MA(2) noise structure. MSE is calculated as `mean((fitted - observed)^2)`.

scatterplots are shown in Figure (2.2). The top three plots of Figure (2.2) compare the white noise fits to the three other fits, once again highlighting that a white noise fit to data with correlated errors leads to MSE values that are large and variable.

The two plots in the middle row of Figure (2.2) compare the spline fit to fits resulting from our spectral algorithm. What we see is that when we correctly specify the order of the MA process our algorithm performs slightly better than the spline fit, as indicated by the fact that most of the points on the left figure lie above the diagonal. When we overfit (specifying the noise order to be higher than it actually is), our algorithm performs comparably to the spline fit, albeit with a bit more

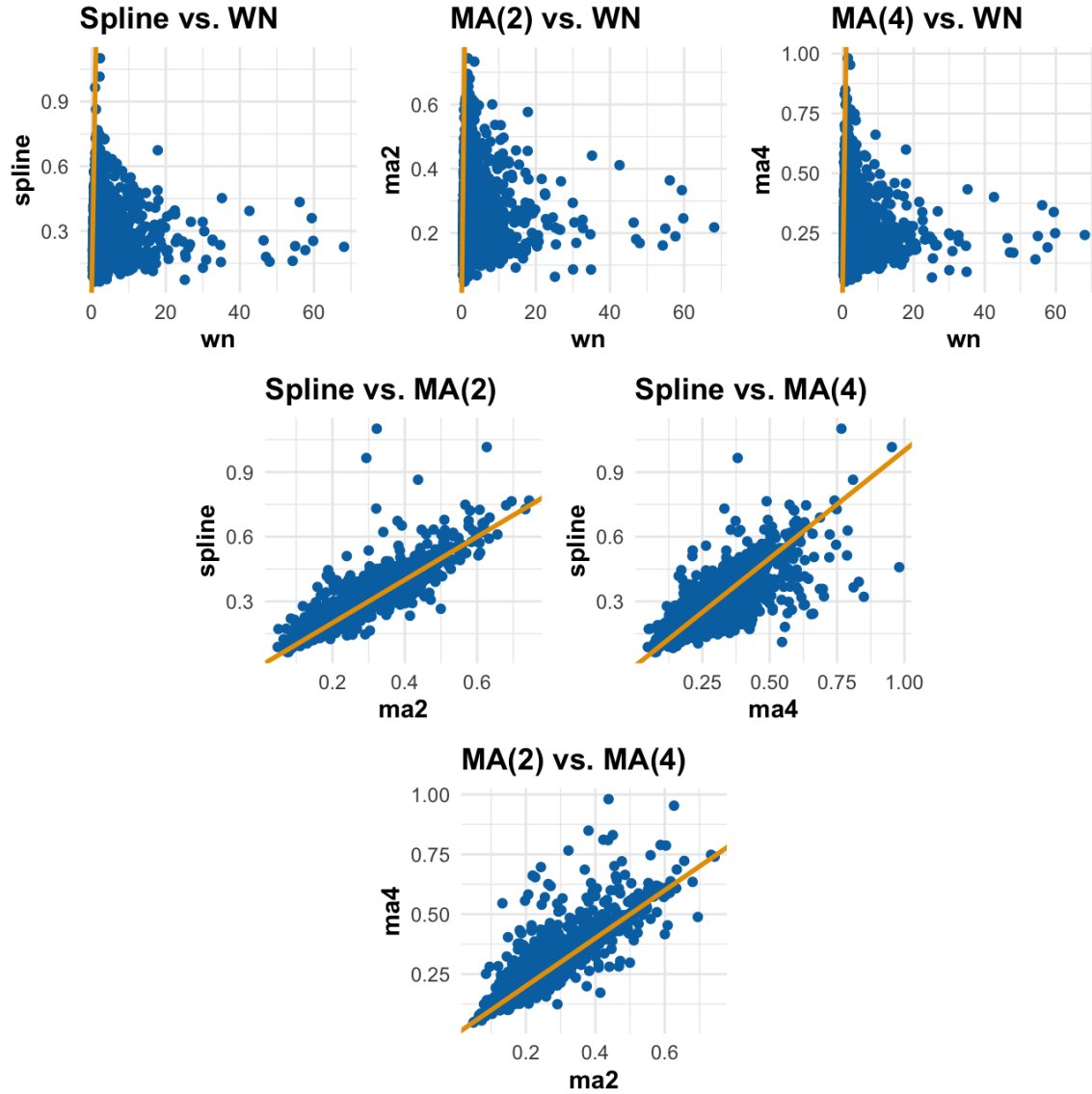


Figure 2.2: Various comparisons of MSEs resulting from the four methods (White Noise, MA(2), MA(4), and correlation-adjusted smoothing spline). White Noise MA(2), and MA(4) fits were generated using Algorithm (1), and splines fits were generated using the `ssr()` function from the `assist` package. In all plots, the diagonal is shown in orange.

spread in the MSEs. The high spread is, in many ways, expected, in light of the fact that the MA(4) fit is overfitting the data.

The effects of overfitting manifest themselves further in the single plot on the third line of Figure (2.2), in which we compare the MA(2) and MA(4) fits (both generated using our proposed estimation procedure). First we note that the MA(4) fit performs worse than the MA(2) fit (as indicated by the fact that most of the points lie above the diagonal). Additionally, we once again see a noticeable spread in the vertical direction, potentially indicating some slight numerical instabilities that are somewhat common among overparametrized GLMs.

Finally, in Figure (2.3) we display the distribution of log-MSEs resulting from each method. As expected, the distribution of MSEs from non-white-noise-fits are somewhat symmetric, whereas the distribution of MSEs from the white noise fit is noticeably right-skewed.

2.3 Applications to Data

In this section, we apply Algorithm (1) several non-simulated datasets. Our goal is to illustrate that our algorithm produces reasonable fits to real-world data, and also to further stress the point that failing to capture correlation structures (i.e. assuming independent noise where noise is *not* independent) is, in many cases, undesirable.

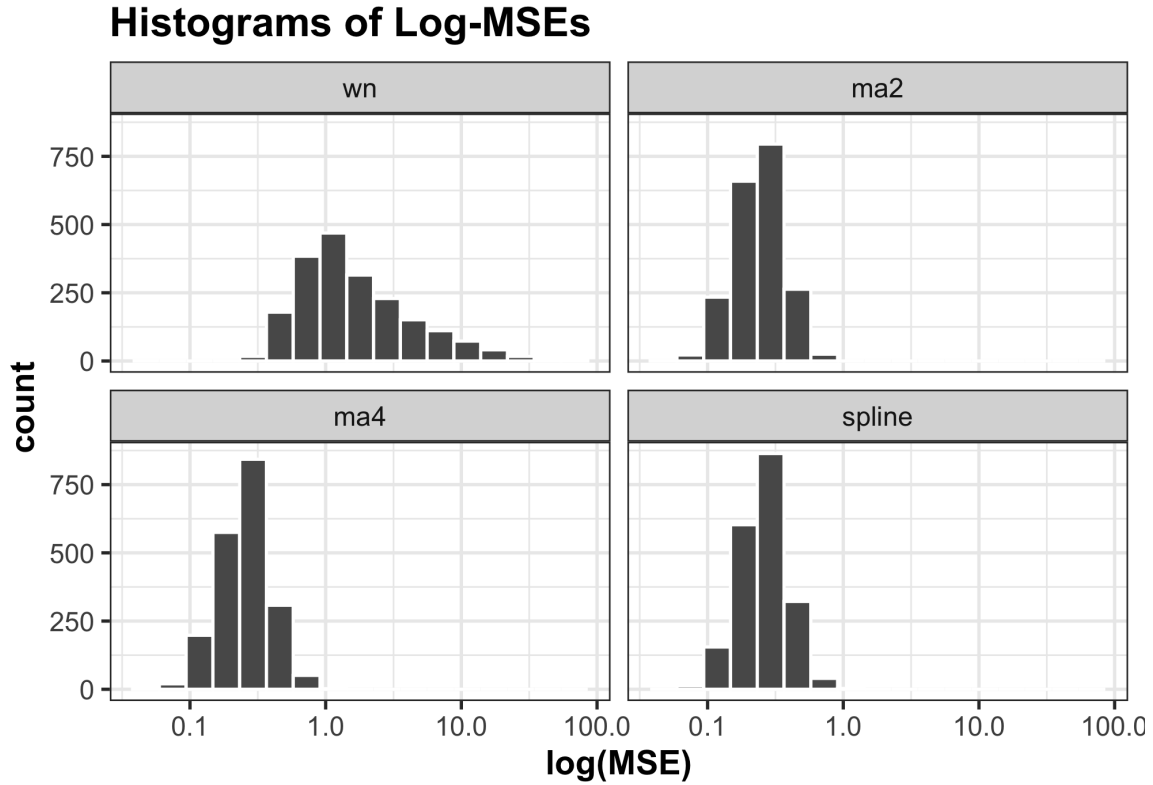


Figure 2.3: Comparison of MSEs resulting from White Noise fits, as compared to MA(2) and MA(4) fits (resulting from our proposed algorithm) and spline fits (from the `ssr()` function in the `assist` package). The diagonal is shown in orange.

2.3.1 Approval Ratings for President Obama

Our first datasets tracks the approval ratings (in percentages) for President Obama during his two terms in office. Data has been accessed from the American Presidency Project [16], and was originally collected by Gallup. One important feature of this dataset is that approval ratings are actually locally averaged, and represent the average (mean) rating across three days: the start date, the day after, and two days after (e.g. the reported average for “July 1” is the average of ratings on July 1,

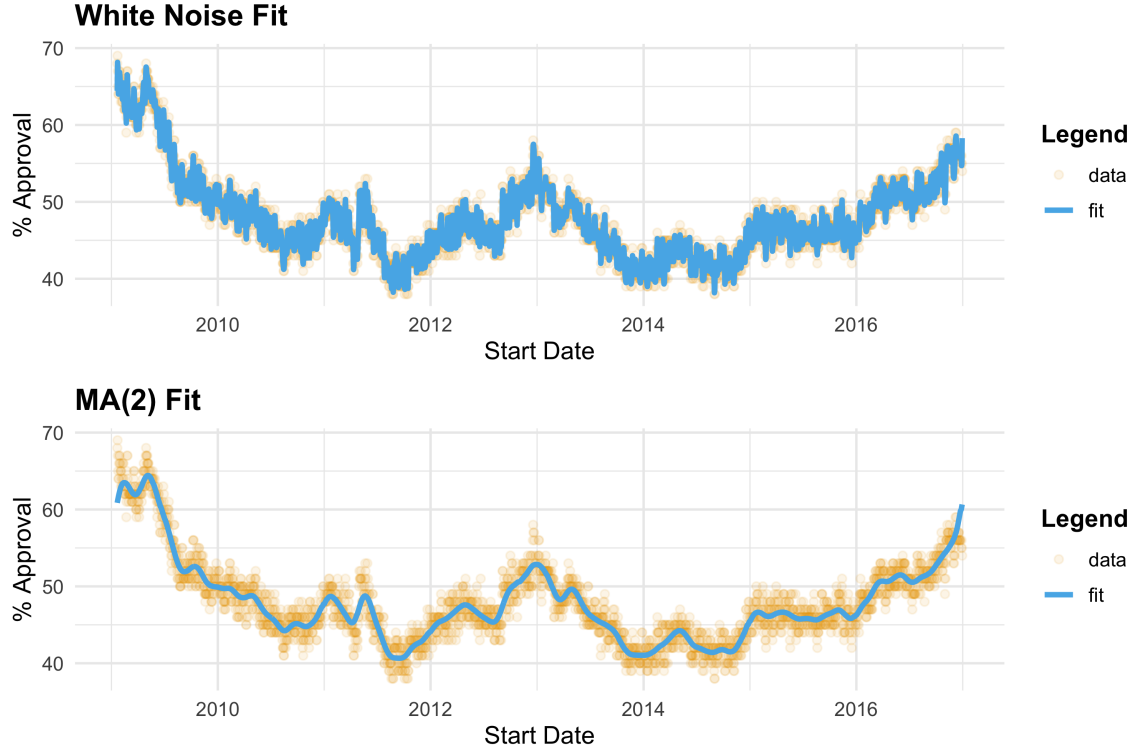


Figure 2.4: Approval percentages for President Obama during his two terms in office. Each point represents the average of approval ratings across three days; the date in question, the day before, and two days before. Data sourced from The American Presidency Project at the University of California, Santa Barbara ([16]). Top: optimal fit assuming white noise. Bottom: optimal fit using Algorithm (1), with bandwidth optimized using a grid-search.

2, and 3).

This means each point is essentially the result of a moving average process of order two. Consequently, modeling the noise as white (uncorrelated) noise is ill-advised; instead, it makes sense to adopt model (1.1) with ω_t modeled as an MA(2) process, i.e. $\omega_t = \varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2}$ for i.i.d. $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. In Figure (2.4), we demonstrate two fits to the approval rating dataset: one (incorrectly) assuming white noise, and one (correctly) assuming MA(2) noise. In both fits, the optimal

bandwidth value was found using a grid-search for the value that minimizes the deviance of the GLM. We note that the white noise fit is considerably noisier than the fit where the correlation structure was explicitly modeled, a result that is to be expected and is well-documented in the nonparametric regression literature. The intuition is as follows: when the model is specified to contain no correlation, yet the model detects autocorrelation in the data, the model assigns the detected correlation to the signal which is, by definition, overfitting (elements that should have been modeled as noise have been modeled as part of the signal). Indeed, under the assumption of white noise, our algorithm produces a fit that is comparable to a kernel smoothing fit, with bandwidth selected using Leave-One-Out Cross-Validation (LOOCV); Figure (2.5) illustrates this similarity. In this way, our algorithm can be stated as performing “as good or better” than a kernel smoother with bandwidth selected by LOOCV.

2.3.2 Datasets with Trend

As with many Fourier-based estimation procedures, our estimation procedure must be modified slightly in the presence of a trend in the data. Consider a deterministic sequence h_t , capturing a trend as a function of time; model (1.1) can then be extended to

$$y_t = f(t/n) + h_t + \omega_t \quad (2.13)$$

For example, if there is evidence to suggest that the response y_t grows linearly with time, we take $h_t = c_0 + c_1 t$ for deterministic constants c_0 and c_1 , leading to

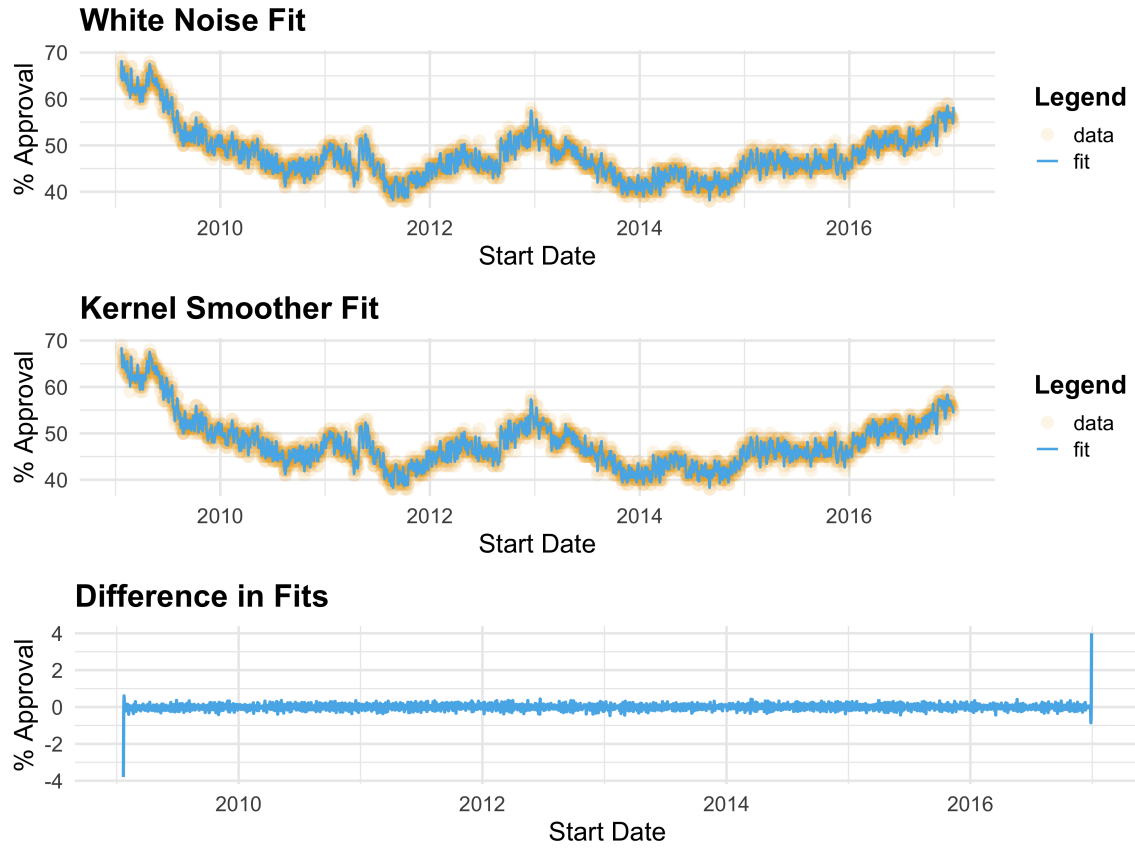


Figure 2.5: Comparison of two fits, both under the assumption of white noise. Top: fit generated by Algorithm (1) with bandwidth selected using a grid-search to minimize the GLM deviance. Middle: kernel smoothing fit with bandwidth selected by LOOCV (Leave-One-Out Cross-Validation). Bottom: fitted values from our algorithm minus fitted values from the kernel smoothing fit, which illustrates that the fitted values generated by the two methods are comparable, save for the endpoints (an artifact of many Fourier-based estimation procedures).

the model

$$y_t = f(t/n) + c_0 + c_1 t + \omega_t \quad (2.14)$$

A somewhat naïve approach to modeling the signal function is to “de-trend” the data by subtracting the deterministic trend from both sides:

$$(y_t - c_0 - c_1 t) = f(t/n) + \omega_t \quad (2.15)$$

Treating $y'_t := y_t - c_0 - c_1 t$ as a new response variable, equation (2.15) is then precisely of the form (1.1) except on the new de-trended response y'_t . In practice, this suggests:

1. Fitting a linear trend to the data using simple linear regression
2. Applying Algorithm (1) to the residuals resulting from the regression in step 1
3. Adding the fitted trend back to the resulting estimates to obtain “re-trended” fitted values

To demonstrate this, we consider a popular dataset that tracks the average monthly CO₂ emissions from the Mauna Loa volcano in Hawai'i, displayed in Figure (2.6). This dataset displays a clear nonlinear trend; nevertheless, we proceed by first removing a linear trend (using a simple linear regression) and then applying our estimation procedure to the resulting residuals. Noting the marked seasonality in our data, we modeled the noise as an MA(12) process (the fact that observations are taken monthly motivates us to try and capture annual dependencies, hence the

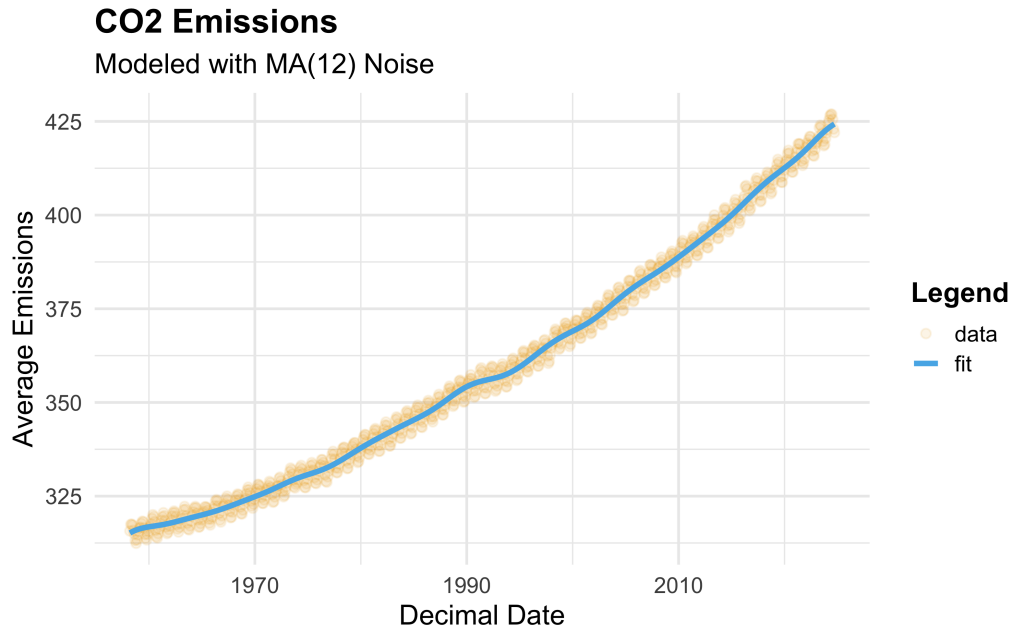


Figure 2.6: Average CO₂ emissions from the Mauna Loa Observatory. The data is first de-trended via a linear regression; Algorithm (1) is then applied to the resulting residuals, assuming MA(12) noise.

choice of order); the resulting estimates (after re-adding the linear trend) are displayed in Figure 2.6 below. Note that the resulting fit does not suffer from any of the traditional edge effects common to Fourier-based methods.

It should be noted that in first removing the linear trend we induce heteroskedasticity in the residuals. As such, a more careful consideration would re-weight the fitted Fourier coefficients to overcome this induced heteroskedasticity. However, empirical results reveal that the magnitude of this heteroskedasticity is negligible, further indicated by the fact that the naïve fit (which does not account for this heteroskedasticity) demonstrated in Figure (2.6) visually captures the data very well. Future work is planned to explore a more careful approach to de-trending the data, including an idea of mapping the data to a new space in which there is no trend.

We believe that such a treatment of heteroskedasticity can lead naturally into explorations of how to extend Algorithm (1) to cases in which data is missing and/or irregularly spaced. Indeed, due to both its reliance on the Fast Fourier Transform as well as the requirement of having an associated response value y_t to use, Algorithm (1) does not currently allow for missing or irregularly-spaced data. Naturally, one workaround in the presence of missing data is to interpolate missing values and proceed. In the presence of irregularly-spaced inputs, one may be tempted to use the Non-Uniform Fast Fourier Transform (NUFFT), however in doing so many mathematical properties (independence of the coefficients, separation of signal- and noise-based terms) could be lost. Instead, the data could be mapped to a space in which observations *are* regularly spaced, Algorithm (1) could be implemented on the transformed data, and the resulting fitted values could then be mapped back to the original space. Such mappings would, of course, require careful consideration and tracking of variances, which we believe would be similar to that required to establish a more rigorous framework for handling trends. Again, future work is planned to explore some of these directions more fully.

CHAPTER 3

Credible Intervals

In this chapter, we turn our attention to constructing credible intervals for our predicted values. Indeed, our estimation procedure allows us to quantify uncertainty using relatively standard first principles. In particular, inspired by the work of [17] and [18], we have

$$\text{Var}(\theta_k | \check{y}_k) = \mathbb{E} [\text{Var}(\theta_k | \check{y}_k, \hat{\beta}) | \check{y}_k] + \text{Var}(\mathbb{E}[\theta_k | \check{y}_k, \hat{\beta}] | \check{y}_k) \quad (3.1)$$

where $\check{y}_k$ denotes the k^{th} Discrete Fourier Transform (DFT) value of y_t and $\hat{\beta}$ denotes our vector of hyperparameter estimates (i.e. estimates of τ^2 , σ^2 , h , and any moving average coefficients). The first term in (3.1) can be well-approximated by taking the posterior mean $\text{Var}(f(t) | \vec{y}, \vec{\beta})$ (which can be found in an introductory Bayesian statistics textbook; for instance [14]) and plugging in $\hat{\beta}$ for β . The second term in (3.1) represents the additional variance contributed by estimating the hyperparameters, and can be well-approximated using a delta method argument (i.e. a Taylor series expansion). The posterior distribution of $(f(t) | \vec{y})$ remains

Gaussian, meaning approximate 95% credible intervals can be obtained using

$$\hat{y}_t \pm 1.96 \cdot \sqrt{\text{Var}(f(t) \mid \vec{y})} \quad (3.2)$$

where $\hat{y}_t$ denotes the fitted values obtained from Algorithm (1). In service of obtaining credible intervals for the Approval Ratings Dataset [16], we begin by deriving credible intervals under the assumption of MA(2) noise.

3.1 Derivation for MA(2) Noise

Again, let θ_k denote the complex-normal "true" Fourier coefficients of the signal process, so that

$$f(t) = \sum_{k=0}^{n-1} \theta_k e^{-2\pi i k t / n}$$

We also have $\theta_k = \theta_{n-k}^*$, to ensure $f(t)$ is real-valued. Again, let v_k denote the spectral density of the covariance kernel and let u_k denote the spectral density of the noise process. For example, a Matérn- ν kernel and an MA(2) noise process imply

$$v_k = \beta_1 (1 + k^2 \beta_2)^{-\nu'} \quad (3.3)$$

$$u_k = \beta_0 + \beta_3 \cos(2\pi k / n) + \beta_4 \cos(4\pi k / n) \quad (3.4)$$

where $\nu' := (\nu + \frac{1}{2})$. With these notations, we have the results

$$\mathbb{E}[\theta_k \mid \check{y}_k, \beta] = \left(\frac{v_k}{u_k + v_k} \right) \check{y}_k =: w_k \check{y}_k \quad (3.5)$$

$$\text{Var}(\theta_k \mid \check{y}_k, \beta) = \left(\frac{u_k v_k}{u_k + v_k} \right) =: w_k u_k \quad (3.6)$$

Beginning with Equation (2.1) and applying Equation (3.5) yields

$$\begin{aligned} \mathbb{E}[f(t) \mid \check{y}_k, \beta] &= \mathbb{E} \left[\sum_{k=0}^{n-1} \theta_k e^{-2\pi i k t / n} \mid \check{y}_k, \beta \right] = \sum_{k=0}^{n-1} \mathbb{E}[\theta_k \mid \check{y}_k, \beta] e^{-2\pi i k t / n} \\ &= \sum_{k=0}^{n-1} \left(\frac{v_k}{u_k + v_k} \right) \check{y}_k e^{-2\pi i k t / n} \end{aligned}$$

Note that we have the implicit understanding that $w_0 := 1$ (we do not weight the zeroth frequency). Hence, we can decompose our expression above as

$$\begin{aligned} \mathbb{E}[f(t) \mid \check{y}_k, \beta] &= \check{y}_0 + \sum_{k=1}^{n-1} \left(\frac{v_k}{u_k + v_k} \right) \check{y}_k e^{-2\pi i k t / n} \\ &= \check{y}_0 + \sum_{k=1}^{(n-1)/2} \left(\frac{v_k}{u_k + v_k} \right) \left[\check{y}_k e^{-2\pi i k t / n} + \check{y}_{n-k} e^{2\pi i k t / n} \right] \end{aligned}$$

Again beginning with Equation (2.1) but now applying Equation (3.6), along with the fact that the θ_k are independent (a fact that follows from stationarity assumptions) yields

$$\begin{aligned} \text{Var}(f(t) \mid \check{y}_k, \beta) &= \text{Var} \left(\sum_{k=0}^{n-1} \theta_k e^{-2\pi i k t / n} \mid \check{y}_k, \beta \right) = \sum_{k=0}^{n-1} \text{Var}(\theta_k \mid \check{y}_k, \beta) \cdot \|e^{-2\pi i k t / n}\|^2 \cdot 1 \\ &= \text{Var}(\theta_0 \mid \check{y}_0, \beta) + \sum_{k=1}^{n-1} \text{Var}(\theta_k \mid \check{y}_k, \beta) = \text{Var}(\theta_0 \mid \vec{y}, \beta) + \sum_{k=1}^{n-1} \left(\frac{u_k v_k}{u_k + v_k} \right) \end{aligned}$$

Note that, by Equation (3.6) and the fact that $w_0 := 1$,

$$\text{Var}(\theta_0 \mid \check{y}_0, \beta) = w_0 \cdot u_0 = u_0$$

For example, adopting (3.4), $\text{Var}(\theta_0 \mid \vec{y}, \beta) = \beta_0 + \beta_3 + \beta_4$. Thus, we have

$$\text{Var}(f(t) \mid \check{y}_k, \beta) = u_0 + \sum_{k=1}^{n-1} \left(\frac{v_k}{u_k + v_k} \right) u_k \quad (3.7)$$

3.1.1 Empirical Bayes Considerations

Of course, our estimation procedure does not use w_k directly since the values of β are unknown. Instead, we estimate them using maximum likelihood estimation, which requires us to think critically about any additional variance incurred by such an estimation. Following the suggestion of Carlin and Louis [17] and Morris [18], we have

$$\text{Var}(\theta_k \mid \check{y}_k) = \mathbb{E} [\text{Var}(\theta_k \mid \check{y}_k, \hat{\beta}) \mid \check{y}_k] + \text{Var}(\mathbb{E}[\theta_k \mid \check{y}_k, \hat{\beta}] \mid \check{y}_k) \quad (3.1, \text{ again})$$

In much of our work below, we will drop the final conditioning on $\check{y}_k$; e.g. writing $\mathbb{E}[\text{Var}(f(t) \mid \check{y}_k, \hat{\beta})]$ in place of $\mathbb{E}[\text{Var}(f(t) \mid \check{y}_k, \hat{\beta}) \mid \check{y}_k]$. This is purely for notational convenience and, we believe, is justified in that the final conditioning on $\check{y}_k$ is implicit.

Posterior Variance

The first term in (3.1) is approximated by plugging our parameter estimates into (3.7):

$$\mathbb{E} [\text{Var}(f(t) \mid \check{y}_k, \hat{\beta})] \approx \text{Var}(\theta_0 \mid \check{y}_0, \hat{\beta}) + \sum_{k=1}^{n-1} \left(\widehat{\frac{u_k v_k}{u_k + v_k}} \right)$$

For example, adopting the forms (3.6) and (3.4),

$$\begin{aligned} \mathbb{E} [\text{Var}(f(t) \mid \check{y}_k, \hat{\beta})] &\approx (\hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_4) \\ &+ \sum_{k=1}^{n-1} \frac{[\hat{\beta}_0 + \hat{\beta}_3 \cos(\frac{2\pi k}{n}) + \hat{\beta}_4 \cos(\frac{4\pi k}{n})] \hat{\beta}_1 (1+k^2 \hat{\beta}_2)^{-\nu'}}{\hat{\beta}_0 + \hat{\beta}_1 (1+k^2 \hat{\beta}_2)^{-\nu'} + \hat{\beta}_3 \cos(\frac{2\pi k}{n}) + \hat{\beta}_4 \cos(\frac{4\pi k}{n})} \end{aligned} \quad (3.8)$$

This is our so-called “**posterior variance**” term, which does not depend on t . (Indeed, we expect that neither term in (3.1) will depend on t due to our stationarity assumptions.)

Variance in the Empirical Estimates of the Prior

The second term in (3.1) describes the additional variance incurred by estimating the β parameters. Our general approach to calculating this term is to:

- 1) Use standard asymptotic likelihood theory to obtain the (approximate) covariance matrix of the $\hat{\beta}$ estimates
- 2) Use a Delta Method approximation to obtain the estimated variance incurred by estimating the β parameters

First, our distributional assumptions (i.e. our choice of a Gaussian Process prior and an MA noise process) imply

$$\eta_k \sim \text{Exponential}(\mu_k(\beta)) \sim \text{Exponential}(u_k + v_k) \quad (3.9)$$

where, as in Chapter 2, $\eta_k := \frac{1}{2} \|\theta_k + \tilde{\zeta}_k\|^2$. This induces the following likelihood

and log-likelihood:

$$\begin{aligned}\mathcal{L}_\theta(\mu) &= \prod_k \left[\frac{1}{\mu_k} e^{-\theta_k/\mu_k} \right] = \prod_{i=1}^n \left(\frac{1}{\mu_k} \right) \times \exp \left\{ - \sum_k \frac{\theta_k}{\mu_k} \right\} \\ \ell_\theta(\mu) &= - \sum_k \ln(\mu_k) - \sum_k \frac{\theta_k}{\mu_k} = - \sum_k \left[\ln(\mu_k) + \frac{\theta_k}{\mu_k} \right]\end{aligned}$$

This likelihood gives rise to the following gradient and Hessian:

$$\vec{\nabla} \ell(\beta) = \sum_k \frac{(\theta_k - \mu_k)}{\mu_k^2} \left[\vec{\nabla} \mu_k(\beta) \right] \quad (3.10)$$

$$\nabla^2 \ell(\beta) = \sum_k \frac{(\mu_k - 2\theta_k)}{\mu_k^3} \left[\vec{\nabla} \mu_k(\beta) \vec{\nabla} \mu_k(\beta)^\top \right] + \sum_k \frac{(\theta_k - \mu_k)}{\mu_k^2} \left[\nabla^2 \mu_k(\beta) \right] \quad (3.11)$$

Also recall that, as indicated in Equation (3.9),

$$\mu_k(\beta) := u_k + v_k = \beta_0 + \beta_1(1 + k^2\beta_2) - v' + \beta_3 \cos\left(\frac{2\pi k}{n}\right) + \beta_4 \cos\left(\frac{4\pi k}{n}\right) \quad (3.12)$$

This means we can calculate $\vec{\nabla} \mu_k(\beta)$ and $\nabla^2 \mu_k(\beta)$ explicitly:

$$\vec{\nabla} \mu_k(\beta) = \begin{pmatrix} 1 \\ (1 + \beta_2 k^2) - v' \\ -v' \beta_1 k^2 (1 + \beta_2 k^2)^{-(v'+1)} \\ \cos(2\pi k/n) \\ \cos(4\pi k/n) \end{pmatrix} \quad (3.13)$$

$$\nabla^2 \mu_k(\beta) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -v'k^2(1 + \beta_2 k^2)^{-(v'+1)} & 0 & 0 \\ 0 & -v'k^2(1 + \beta_2 k^2)^{-(v'+1)} & v'(v' + 1)\beta_1 k^4(1 + \beta_2 k^2)^{-(v'+2)} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (3.14)$$

We now turn our attention to the Delta Method argument to give us the estimated effect our β estimation has on the μ terms. Our Delta Method argument is:

$$\text{Var}(\mathbb{E}[f(t) | \check{y}_k, \hat{\beta}]) \approx \begin{pmatrix} \frac{\partial}{\partial \beta_0} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \\ \frac{\partial}{\partial \beta_1} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \\ \frac{\partial}{\partial \beta_2} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \\ \frac{\partial}{\partial \beta_3} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \\ \frac{\partial}{\partial \beta_4} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \end{pmatrix}^T \text{Cov}(\hat{\beta}) \begin{pmatrix} \frac{\partial}{\partial \beta_0} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \\ \frac{\partial}{\partial \beta_1} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \\ \frac{\partial}{\partial \beta_2} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \\ \frac{\partial}{\partial \beta_3} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \\ \frac{\partial}{\partial \beta_4} \mathbb{E}[f(t) | \check{y}_k, \hat{\beta}] \end{pmatrix} \quad (3.15)$$

where $\text{Cov}(\hat{\beta}) \approx -\nabla^2 \ell(\hat{\beta})$ and $\nabla^2 \ell(\beta)$ is given by Equation (3.11). We may rewrite Equation (3.15) more succinctly as

$$\text{Var}(\mathbb{E}[f(t) | \check{y}_k, \hat{\beta}]) = \left(\vec{\nabla}_{\hat{\beta}} \mathbb{E}[f(y) | \check{y}_k, \hat{\beta}] \right)^T (\text{Cov}(\hat{\beta})) \left(\vec{\nabla}_{\hat{\beta}} \mathbb{E}[f(y) | \check{y}_k, \hat{\beta}] \right) \quad (3.16)$$

where the notation $\vec{\nabla}_{\hat{\beta}}$ denotes that the derivatives in the gradient are taken with respect to the β coefficients. The remaining partial derivatives in (3.15) can be ob-

tained through repeated applications of the Quotient and Chain rules:

$$\frac{\partial}{\partial \beta_0} \mathbb{E}[f(t) \mid \check{y}_k, \hat{\beta}] = \sum_{k=1}^{(n-1)/2} \frac{-\hat{\beta}_1(1+k^2\hat{\beta}_2)^{-\nu'}}{\left[\hat{\beta}_0 + \hat{\beta}_1(1+k^2\hat{\beta}_2)^{-\nu'} + \hat{\beta}_3 \cos\left(\frac{2\pi k}{n}\right) + \hat{\beta}_4 \cos\left(\frac{4\pi k}{n}\right)\right]^2} \cdot z_{k,t} \quad (3.17)$$

$$\frac{\partial}{\partial \beta_1} \mathbb{E}[f(t) \mid \check{y}_k, \hat{\beta}] = \sum_{k=1}^{(n-1)/2} \frac{\left(\hat{\beta}_0 + \hat{\beta}_3 \cos\left(\frac{2\pi k}{n}\right) + \hat{\beta}_4 \cos\left(\frac{4\pi k}{n}\right)\right)(1+k^2\hat{\beta}_2)^{-\nu'}}{\left[\hat{\beta}_0 + \hat{\beta}_1(1+k^2\hat{\beta}_2)^{-\nu'} + \hat{\beta}_3 \cos\left(\frac{2\pi k}{n}\right) + \hat{\beta}_4 \cos\left(\frac{4\pi k}{n}\right)\right]^2} \cdot z_{k,t} \quad (3.18)$$

$$\frac{\partial}{\partial \beta_2} \mathbb{E}[f(t) \mid \check{y}_k, \hat{\beta}] = \sum_{k=1}^{(n-1)/2} \frac{-\nu' k^2 \hat{\beta}_1 \left(\hat{\beta}_0 + \hat{\beta}_3 \cos\left(\frac{2\pi k}{n}\right) + \hat{\beta}_4 \cos\left(\frac{4\pi k}{n}\right)\right)(1+k^2\hat{\beta}_2)^{-(\nu'+1)}}{\left[\hat{\beta}_0 + \hat{\beta}_1(1+k^2\hat{\beta}_2)^{-\nu'} + \hat{\beta}_3 \cos\left(\frac{2\pi k}{n}\right) + \hat{\beta}_4 \cos\left(\frac{4\pi k}{n}\right)\right]^2} \cdot z_{k,t} \quad (3.19)$$

$$\frac{\partial}{\partial \beta_3} \mathbb{E}[f(t) \mid \check{y}_k, \hat{\beta}] = \sum_{k=1}^{(n-1)/2} \frac{-\cos\left(\frac{2\pi k}{n}\right) \hat{\beta}_1(1+k^2\hat{\beta}_2)^{-\nu'}}{\left[\hat{\beta}_0 + \hat{\beta}_1(1+k^2\hat{\beta}_2)^{-\nu'} + \hat{\beta}_3 \cos\left(\frac{2\pi k}{n}\right) + \hat{\beta}_4 \cos\left(\frac{4\pi k}{n}\right)\right]^2} \cdot z_{k,t} \quad (3.20)$$

$$\frac{\partial}{\partial \beta_4} \mathbb{E}[f(t) \mid \check{y}_k, \hat{\beta}] = \sum_{k=1}^{(n-1)/2} \frac{-\cos\left(\frac{4\pi k}{n}\right) \hat{\beta}_1(1+k^2\hat{\beta}_2)^{-\nu'}}{\left[\hat{\beta}_0 + \hat{\beta}_1(1+k^2\hat{\beta}_2)^{-\nu'} + \hat{\beta}_3 \cos\left(\frac{2\pi k}{n}\right) + \hat{\beta}_4 \cos\left(\frac{4\pi k}{n}\right)\right]^2} \cdot z_{k,t} \quad (3.21)$$

where we have let $z_{k,t} := [\check{y}_k e^{-2\pi i k t/n} + \check{y}_{n-k} e^{2\pi i k t/n}]$ for ease of notation. (See the Appendix for a derivation of these terms.) Derivatives (3.17) - (3.21) can be substituted into Equation (3.15), which gives the second term in Equation (3.1).

3.1.2 Application to Approval Ratings Dataset

With this procedure in mind, we can construct approximate 95% credible intervals for the true approval percentage y_t at a time t , using the Approval Ratings dataset ([16]).

It may be illustrative to examine the contributions of each of the two terms in 3.1 separately.

The top plot of (3.1) illustrates the second term on the right hand side of Equation (3.1); i.e. the variance in the empirical estimates of the Prior. The bottom plot of

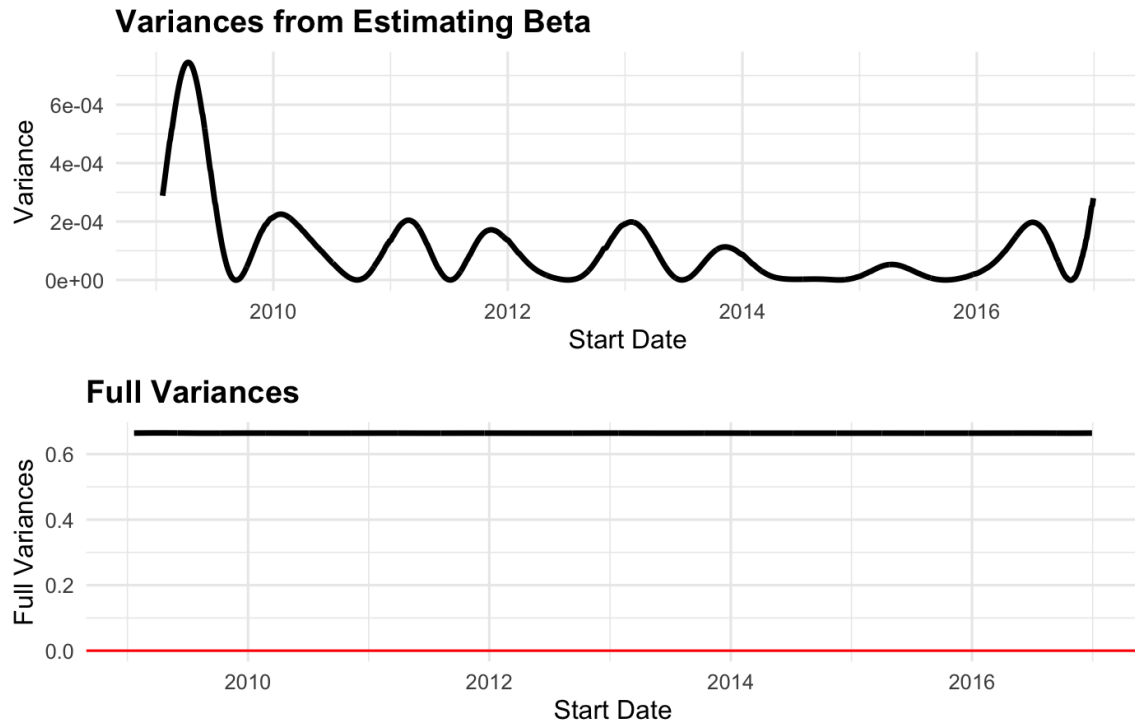


Figure 3.1: **Top:** Variance contributed by estimating the hyperparameters. **Bottom:** Full variance, including both terms in 3.6. Note that the uncertainty contributed by estimating each hyperparameter is negligible when compared to the posterior variance (the second term in 3.6), indicating that the full posterior uncertainty is not too dependent on our ability to estimate the individual hyperparameters with great accuracy.

(3.1) illustrates the full (posterior variance plus variance in the empirical estimates of the prior); i.e. the left hand side of Equation (3.1).

What we see is that the full posterior variances are dominated by the posterior variance, and that the variance incurred by estimating the model hyperparameters is negligible. This is a comforting result, indicating that small errors in estimating the model hyperparameters will not unduly influence the full posterior uncertainty. This point will be revisited in Chapter (4), where we explore the

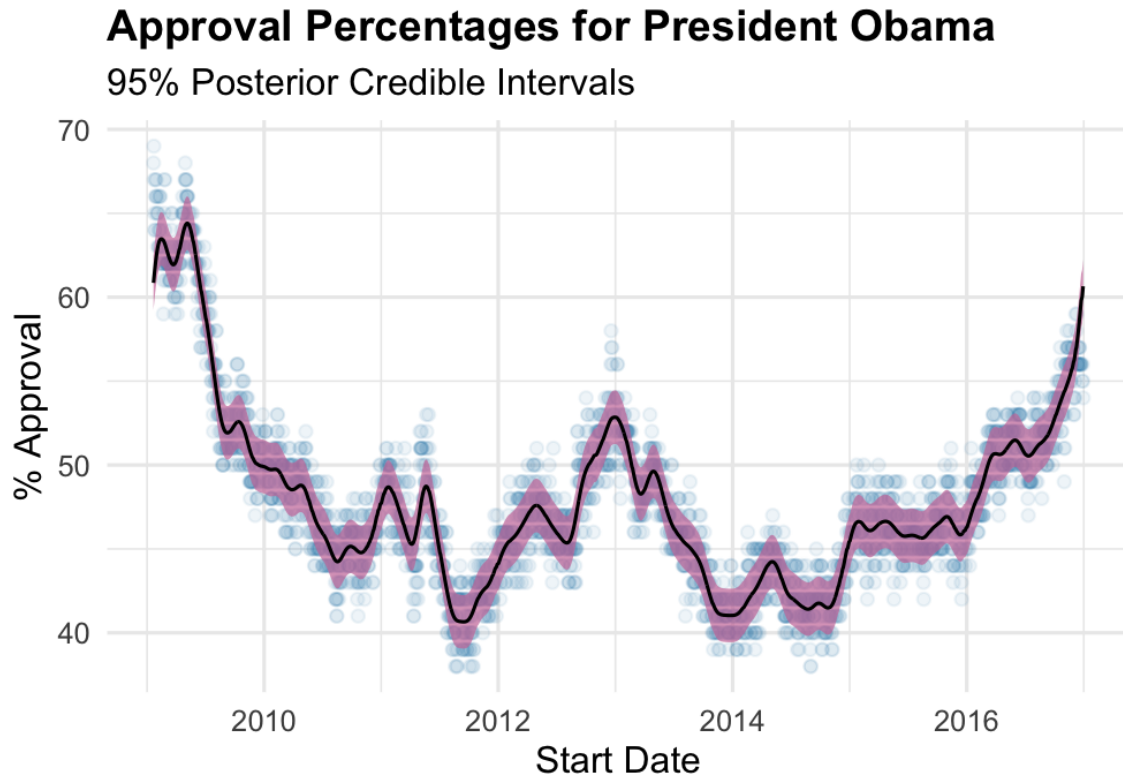


Figure 3.2: Approval Ratings Dataset, with full 95% posterior credible intervals indicated. Note that these intervals are pointwise.

large-sample properties of our estimator. For now, we draw the reader's attention to Figure (3.2), which illustrates the full pointwise 95% credible intervals for the Approval Ratings dataset introduced in Section (2.3.1), as prescribed in Equation (3.2).

We stress that these are *pointwise* intervals; i.e. they are not true confidence bands that achieve simultaneous 95% coverage.

3.2 Extensions to Higher-Order Noise

Though the derivations in the previous section assume an MA(2) process for the noise term, it is relatively straightforward to extend this process to higher-order MA terms. Specifically, v_k remains unchanged from Equation (3.3), and u_k can be modified from Equation (3.4) by simply incorporating additional cosine terms; see Appendix (A) for a general outline.

As an example of this implementation on a dataset with higher-order noise, we consider a dataset that tracked the number of reported cases of COVID-19 in Santa Barbara County Hospitals. Due to slower reporting over weekends, such data is suspected to have a weekly autocorrelation, leading us to model the noise term ω as an MA(7) process. Due to the sharp peaks in the data, counts of cases are transformed using a square root, and analysis is performed on the square-root-cases. The data, along with pointwise 95% posterior credible intervals, is displayed in Figure 3.3

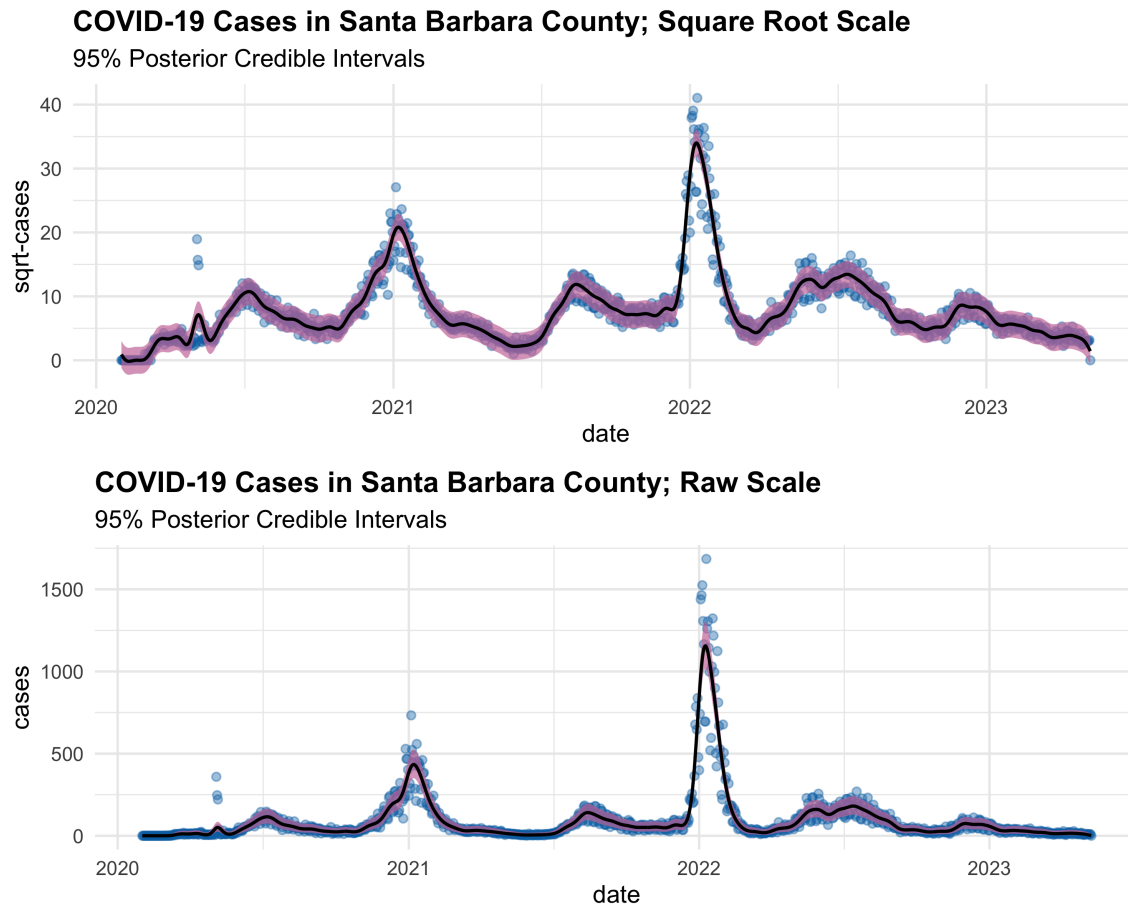


Figure 3.3: 95% credible intervals for the counts of COVID-19 cases in Santa Barbara County. **Top:** Square-root-transformed data and intervals. **Bottom:** Raw-scale data and intervals. In both plots, intervals are displayed as shaded ribbons, the fitted values are displayed as a solid black line, and the original data is displayed as points.

CHAPTER 4

Asymptotics

In this chapter, we explore the large-sample properties of our estimation procedure.

4.1 Asymptotic Rates

As noted in Section 2.1.1 we wish to consider signals in a Sobolev- α class. The Matérn kernel prior with parameter $\nu < \alpha$ ensures that the expected Sobolev norm of the signal function under the prior is finite. That is,

$$\mathbb{E} \left[\sum_k (1 + k^2)^\alpha |\check{f}_k|^2 \right] = \sum_k (1 + k^2)^\alpha (1 + k^2)^{-(\nu + \frac{1}{2})} = \sum_k (1 + k^2)^{-\alpha - (\nu + \frac{1}{2})} < \infty$$

or, as [12] and [11] note, that the Fourier terms v_k decay at a polynomial rate.

We would like a bound on $\int (\hat{f} - f)^2 dt$ which, by Parseval's Identity, is equivalent to a bound on

$$\sum_{k=0}^{\infty} \mathbb{E} \left[(\hat{\theta}_k - \theta_k)^2 \right] \tag{4.1}$$

Our prior is $\theta_k \sim \mathcal{CN}(0, v_k)$ and our model posits $(\check{y}_k \mid \theta_k) \sim \mathcal{CN}(\theta_k, u_k)$ where $\check{y}_k$ denotes the Discrete Fourier Transform values of y_t . For a Matérn- ν kernel with White Noise, we have

$$u_k = \sigma^2 / n$$

$$v_k = \beta_1 (1 + \beta_2 k^2)^{-(\nu + \frac{1}{2})}$$

Additionally, our fitted values are

$$\hat{\theta}_k = w_k \cdot \check{y}_k \quad \text{where} \quad w_k := \frac{v_k}{u_k + v_k}$$

4.2 White Noise

We first suppose the hyperparameters $\sigma^2, \beta_0, \beta_1$, and β_2 are known. A simple invocation of a bias-variance decomposition (or expanding the square and simplifying; see Appendix B) reveals that

$$\mathbb{E}[(\hat{\theta}_k - \theta_k)^2] = w_k^2 u_k + (1 - w_k)^2 v_k$$

which reveals that Equation (4.1) is equivalent to

$$\sum_{k=0}^{\infty} \mathbb{E}[(\hat{\theta}_k - \theta_k)^2] = \sum_{k=0}^{\infty} [w_k^2 u_k + (1 - w_k)^2 v_k] \tag{4.2}$$

$$= \underbrace{\sum_{k=0}^{\infty} w_k^2 u_k}_{:= (A)} + \underbrace{\sum_{k=0}^{\infty} (1 - w_k)^2 v_k}_{:= (B)} \tag{4.3}$$

It should be noted that we are considering concentration in a frequentist sense; that is, $\mathbb{E}[(\hat{\theta}_k - \theta_k)^2]$ represents the frequentist MSE as opposed to the Bayesian MSE (as we are not averaging over the prior). Rather, our argument is mainly to demonstrate that the width of the posterior credible intervals shrinks to zero as the sample size increases.

We bound terms (A) and (B) separately. In both cases, it will be useful to note that $w_k \leq 1$ and $(1 - w_k) \leq 1$, which means $w_k^2 \leq w_k$ and $(1 - w_k)^2 \leq w_k$. Furthermore, our strategy will be to utilize the interplay between u_k and v_k for small and large values of k . That is, when k is small we have that v_k dominates u_k , whereas when k is large we have that u_k dominates v_k . Our first task is to identify the cutoff point that divides “small” values of k from “large” values.

First focusing on (A), we let J denote the cutoff point so that

$$(A) := \sum_{k=0}^{J-1} w_k^2 u_k + \sum_{k=J}^{\infty} w_k^2 u_k$$

As mentioned above, v_k dominates u_k for small k meaning

$$\sum_{k < J} w_k^2 u_k \leq \sum_{k < J} w_k u_k = \sum_{k < J} \frac{u_k v_k}{u_k + v_k} \leq \sum_{k < J} u_k = \sum_{k < J} \frac{\sigma^2}{n} = \frac{\sigma^2}{n} J$$

For large k , we have

$$\begin{aligned} \sum_{k \geq J} w_k^2 u_k &\leq \sum_{k \geq J} w_k u_k = \sum_{k \geq J} \left(\frac{u_k v_k}{u_k + v_k} \right) \leq \sum_{k \geq J} v_k \\ &\leq \sum_{k \geq J} (1 + k^2)^{-(\nu + \frac{1}{2})} \leq C \cdot \sum_{k \geq J} k^{-(2\nu + 1)} \end{aligned}$$

$$\leq M \cdot \int_J^\infty k^{-(2\nu+1)} dk \leq M \cdot J^{-2\nu}$$

So, combining results reveals that the overall bound for sum (A) above is

$$\sum_{k=0}^{\infty} w_k^2 u_k \leq C \cdot \frac{\sigma^2}{n} J + M \cdot J^{-2\nu} \quad (4.4)$$

We wish to find J that minimizes this cutoff (to obtain a sort of minimax rate of convergence). Heuristically, we know that such a J value is obtained where the two terms on the right hand side of the inequality have the same order in n ; that is, we select J so that

$$\frac{J}{n} = J^{-2\nu} \implies J^{1+2\nu} = n \implies \boxed{J = n^{\frac{1}{1+2\nu}}}$$

which, when substituted back into Equation (4.4), gives

$$\sum_{k=0}^{\infty} w_k^2 u_k \leq \boxed{n^{-\frac{2\nu}{2\nu+1}}}$$

We now examine sum (B) in Equation (4.3). Note that

$$\sum_{k=0}^{\infty} (1 - w_k)^2 v_k \leq \sum_{k=0}^{\infty} (1 - w_k) v_k = \sum_{k=0}^{\infty} \left(\frac{u_k v_k}{u_k + v_k} \right)$$

which is precisely the same term that we bounded in (A), meaning we would still take $J = n^{\frac{1}{1+2\nu}}$, leading to

$$(B) := \sum_{k=0}^{\infty} (1 - w_k)^2 v_k \leq n^{-\frac{2\nu}{2\nu+1}}$$

So, our overall bound for Equation (4.1) will be

$$\sum_{k=0}^{\infty} \mathbb{E} \left[(\hat{\theta}_k - \theta_k)^2 \right] \leq 2 \cdot n^{-\frac{2\nu}{2\nu+1}} \leq n^{-\frac{2\nu}{2\nu+1}}$$

For example, if we adopt a Matérn-5/2 kernel we take $\nu = 5/2$, leading to a rate of convergence of the order $n^{-5/6}$.

4.2.1 MA Noise

We now consider the case where the noise process follows a Moving Average process of order q . Specifically, we assume

$$\omega_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots + \psi_q \varepsilon_{t-q+1} \quad \text{where} \quad \varepsilon_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

Our u_k term can now be written, referring to the parametrization in Equation (2.6) as

$$u_k = \frac{\sigma^2}{n} \left[b_0 + \sum_{j=1}^q b_j \cos \left(j \cdot \frac{2\pi k}{n} \right) \right]$$

where the b_j 's are nonlinear combinations of the original MA coefficients $\psi_1, \dots, \psi_q$ (and therefore do not depend on n). Terms (A) and (B) in equation 4.3 remain the same, except with this new expression for u_k . As such, our strategy for obtaining a bound also remains the same: first find the cutoff J , and use this to obtain optimal rates.

The bounds for large k needn't be changed, as the polynomial decay of the v_k terms ensures they will still dominate u_k when k is sufficiently large. As such, we

need only explore the bounds on (A) and (B) for small k . Note that we still have

$$\sum_{k < J} w_k^2 u_k \leq \sum_{k < J} w_k u_k = \sum_{k < J} \left(\frac{v_k}{u_k + v_k} \right) u_k \leq \sum_{k < J} u_k$$

Because u_k is no longer just σ^2/n , we cannot immediately write this as $J(\sigma^2/n)$.

However, we can now refer to Equation (2.6) to see that, since $\cos(x) \leq 1$ for any $x \in \mathbb{R}$,

$$u_k \leq \frac{\sigma^2}{n} [b_0 + b_1 + \cdots + b_q] =: C \cdot \frac{\sigma^2}{n}$$

Our bound on (A) for small k is therefore now

$$\sum_{k < J} w_k^2 u_k \leq \sum_{k < J} u_k \leq \sum_{k < J} \frac{\sigma^2}{n} [b_0 + b_1 + \cdots + b_q] \leq C \cdot \frac{\sigma^2}{n} J$$

where the constant C is the sum of the coefficients of the spectral density of the MA process. Since, as previously mentioned, our bounds for large k remain unchanged from the white noise case, this reveals that our optimal J is still that which sets

$$\frac{J}{n} = J^{-2\nu} \implies J = n^{\frac{1}{1+2\nu}}$$

which still gives us an overall bound for Equation (4.1) as

$$\sum_{k=0}^{\infty} \mathbb{E} [(\hat{\theta}_k - \theta_k)^2] \leq n^{-\frac{2\nu}{2\nu+1}}$$

CHAPTER 5

Conclusion

In this thesis, we considered an Empirical Bayes approach to nonparametric regression that leveraged spectral considerations to easily extend to cases of correlated noise. Though the topic of correlated errors in nonparametric regression is widely explored (as is outlined in Chapter 1), we believe our approach combines many of these existing techniques in a novel way. Furthermore (as shown in Chapter 3) not only does our model lend itself to easy adaptation in the case of higher-order noise terms, but, in utilizing an Empirical Bayes framework, we are able to easily perform uncertainty quantification. The effectiveness of our algorithm was demonstrated on simulated and non-simulated datasets in Chapter (2); additionally, as shown in Chapter 4, our estimator possesses the desirable property that the MSE of predicted values goes to zero at a rate bounded above by $n^{-\frac{2\nu}{2\nu+1}}$, where ν represents the roughness parameter of the underlying Matérn kernel.

5.1 Future Work

Several directions for future work present themselves. First is the adaptation of our algorithm in the presence of missing and/or non-uniformly-spaced inputs. Section (2.3.2) provides some preliminary thoughts on this, with the main idea being to map the data to a space where inputs are non-missing and uniformly-spaced, so that our algorithm may be applied and the resulting fitted values can be back-transformed to the original space. This approach, we believe, goes hand-in-hand with our plans to establish a more complete treatment of data with deterministic trends.

Another direction of future work lies in the computational implementation of our algorithm. As noted in Section (2.2), our algorithm currently relies on the base `glm()` function in R to fit the optimal hyperparameters for a fixed bandwidth, and then grid-searches over candidate values to identify the optimal bandwidth. This function, however, is notoriously finicky and is prone to false starts, warnings, and sometimes outright errors unless appropriate starting values and candidate bandwidths are selected. Some work has been dedicated to writing a Newton-Raphson algorithm to bypass these intricacies of the `glm()` function (and simultaneously fit all parameters, rather than sequentially optimizing the non-bandwidth and bandwidth parameters separately), however this algorithm still suffered from numerical instability. To that end, a more computationally robust implementation of our algorithm is certainly warranted.

We are also interested in obtaining a more complete treatment of the asymptotics of our estimator. Though we are able to estimate noise-related hyperparam-

eters at parametric rates, obtaining rates for signal-related hyperparameters requires bounding the delta-method term

$$\left(\vec{\nabla}_{\vec{\beta}} \mathbb{E}[f(t) \mid \check{y}_k, \beta] \right)^T (\text{Cov}(\hat{\beta})) \left(\vec{\nabla}_{\vec{\beta}} \mathbb{E}[f(t) \mid \check{y}_k, \beta] \right)$$

where the notation $\vec{\nabla}_{\vec{\beta}}$ means derivatives are to be taken with respect to the β 's.

5.2 General Disclosures

All code was run using the R programming language [19]. Plots have been generated using the ggplot2 package [20], and multipanel plots have been generated using the gridExtra package [21]. Dataframe transformations were implemented using the tidyverse [22] and the reshape2 package [23]. The correlation-corrected smoothing splines of Section (2.2) were implemented using the assist package [24], with thanks to Dr. Yuedong Wang.

Colorblind accessible functionality has been implemented using the ggokabeito package [25]. Every effort has been made to comply with the Web Content Accessibility Guidelines (WCAG) 2.1.

APPENDIX A

Derivative Calculations for Credible Intervals

We have that

$$\mathbb{E}[f(t) \mid \vec{y}, \beta] = \sum_{k=1}^{n-1} \left(\frac{v_k}{u_k + v_k} \right) \check{y}_k e^{-2\pi i k t / n}$$

Our Delta Method implementation requires calculating $\frac{\partial}{\partial \beta_j} \mathbb{E}[f(t) \mid \vec{y}, \beta]$ for all appropriate values of j . The number of “appropriate values” is determined by the order of the noise process; for example, in an MA(2) setting we have parameters $\beta_0, \beta_1, \dots, \beta_4$ and in an MA(7) setting we have parameters $\beta_0, \beta_1, \dots, \beta_8$. In general, for an MA(q) process, we have parameters $\beta_0, \beta_1, \dots, \beta_{q+1}$ since

$$\begin{aligned} v_k &= \beta_1 (1 + k^2 \beta_2)^{-\nu'} \\ u_k &= \beta_0 + \sum_{j=1}^q \beta_{j+2} \cos(2\pi j k / n) \end{aligned}$$

By the linearity of derivatives,

$$\frac{\partial}{\partial \beta_j} \mathbb{E}[f(t) \mid \vec{y}, \beta] = \sum_{k=1}^{n-1} \frac{\partial}{\partial \beta_j} \left(\frac{v_k}{u_k + v_k} \right) \check{y}_k e^{-2\pi i k t / n}$$

meaning it suffices to calculate $\frac{\partial}{\partial \beta_j} \left(\frac{v_k}{u_k + v_k} \right)$ for the appropriate values of j .

The Quotient Rule for derivatives gives

$$\frac{\partial}{\partial \beta_j} \left(\frac{v_k}{u_k + v_k} \right) = \frac{\frac{\partial v_k}{\partial \beta_j} (u_k + v_k) - v_k \frac{\partial (u_k + v_k)}{\partial \beta_j}}{(u_k + v_k)^2} =: \frac{(1) - (2)}{(u_k + v_k)} \quad (\text{A.1})$$

where $(1) := \frac{\partial v_k}{\partial \beta_j} (u_k + v_k)$ and $(2) := v_k \frac{\partial (u_k + v_k)}{\partial \beta_j}$. For notational convenience, define $a_k := (1 + k^2 \beta_2)$ so that $v_k = \beta_1 a_k^{-v'}$.

We now outline the derivative calculation for each value of j .

- $j = 0$: Term (1) in equation (A.1) is zero and term (2) is v_k , leading to

$$\frac{\partial}{\partial \beta_0} \left(\frac{v_k}{u_k + v_k} \right) = \frac{-v_k}{(u_k + v_k)^2} = -\frac{\beta_1 (1 + k^2 \beta_2)^{-v'}}{(u_k + v_k)^2}$$

- $j = 1$

$$\begin{aligned} (1) - (2) &= a_k^{-v'} (u_k + \cancel{\beta_1 a_k^{-v'}}) - \cancel{\beta_1 a_k^{-v'}} (\cancel{a_k^{-v'}}) \\ &= a_k^{-v'} (u_k) = (1 + k^2 \beta_2)^{-v'} (u_k) \\ \Rightarrow \frac{\partial}{\partial \beta_1} \left(\frac{v_k}{u_k + v_k} \right) &= \frac{(1 + k^2 \beta_2)^{-v'} (u_k)}{(u_k + v_k)^2} \end{aligned}$$

- $j = 2$

$$\begin{aligned} (1) - (2) &= -v' k^2 \beta_1 a_k^{-(v'+1)} (u_k + \cancel{\beta_1 a_k^{-v'}}) \\ &\quad - \cancel{\beta_1 a_k^{-v'}} (\cancel{-v' k^2 \beta_1 a_k^{-(v'+1)}}) \\ &= -v' k^2 \beta_1 a_k^{-(v'+1)} (u_k) = -v' k^2 \beta_1 (1 + k^2 \beta_2)^{-(v'+1)} (u_k) \end{aligned}$$

$$\implies \frac{\partial}{\partial \beta_2} \left(\frac{v_k}{u_k + v_k} \right) = \frac{-\nu' k^2 \beta_1 (1 + k^2 \beta_2)^{-(\nu'+1)} (u_k)}{(u_k + v_k)^2}$$

- $j = 3, 4, \dots, (q+2)$: Note that v_k does not depend on $\beta_3, \beta_4, \dots, \beta_{q+2}$; only u_k does. As a result, $\frac{\partial v_k}{\partial \beta_j} = 0$ and so $(1) = 0$; furthermore,

$$\frac{\partial(u_k + v_k)}{\partial \beta_j} = \frac{\partial u_k}{\partial \beta_j} = \cos \left(\frac{2(j-2)\pi k}{n} \right)$$

Therefore, we have

$$\frac{\partial}{\partial \beta_j} \left(\frac{v_k}{u_k + v_k} \right) = \frac{-\beta_1 (1 + k^2 \beta_2)^{-\nu'} \cdot \cos \left(\frac{2(j-2)\pi k}{n} \right)}{(u_k + v_k)^2}, \quad j = 3, 4, \dots, (q+2)$$

A.0.1 MA(2) Case

It may be illustrative to write out these terms more explicitly for an MA(2) noise process. Specifically, we take

$$\begin{aligned} v_k &= \beta_1 (1 + k^2 \beta_2)^{-\nu'} \\ u_k &= \beta_0 + \beta_3 \cos(2\pi k/n) + \beta_4 \cos(4\pi k/n) \end{aligned}$$

For notational convenience, let

$$b(\vec{\beta}, k) := \frac{1}{[\beta_0 + \beta_1 (1 + k^2 \beta_2)^{-\nu'} + \beta_3 \cos(2\pi k/n) + \beta_4 \cos(4\pi k/n)]^2}$$

Then:

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} \left(\frac{v_k}{u_k + v_k} \right) &= b(\vec{\beta}, k) \left[-\beta_1 (1 + k^2 \beta_2)^{-\nu'} \right] \\
\frac{\partial}{\partial \beta_1} \left(\frac{v_k}{u_k + v_k} \right) &= b(\vec{\beta}, k) \left[(1 + k^2 \beta_2)^{-\nu'} (\beta_0 + \beta_3 \cos(2\pi k/n) + \beta_4 \cos(4\pi k/n)) \right] \\
\frac{\partial}{\partial \beta_2} \left(\frac{v_k}{u_k + v_k} \right) &= b(\vec{\beta}, k) \left[-\nu' k^2 \beta_1 (1 + k^2 \beta_2)^{-(\nu'+1)} (\beta_0 + \beta_3 \cos(2\pi k/n) \right. \\
&\quad \left. + \beta_4 \cos(4\pi k/n)) \right] \\
\frac{\partial}{\partial \beta_3} \left(\frac{v_k}{u_k + v_k} \right) &= b(\vec{\beta}, k) \left[-\beta_1 (1 + k^2 \beta_2)^{-\nu'} \cdot \cos(2\pi k/n) \right] \\
\frac{\partial}{\partial \beta_4} \left(\frac{v_k}{u_k + v_k} \right) &= b(\vec{\beta}, k) \left[-\beta_1 (1 + k^2 \beta_2)^{-\nu'} \cdot \cos(4\pi k/n) \right]
\end{aligned}$$

These are the terms that were used in Equations (3.17) through (3.21)

APPENDIX B

Bias-Variance Decomposition for MSE

Lemma B.0.1. $\mathbb{E}[(\hat{\theta}_k - \theta_k)^2] = w_k^2 u_k + (1 - w_k)^2 v_k$

Proof. We first expand the square and simplify:

$$\begin{aligned}\mathbb{E}[(\hat{\theta}_k - \theta_k)^2] &= \mathbb{E}[\hat{\theta}_k^2 + \theta_k^2 - 2\hat{\theta}_k\theta_k] \\ &= \mathbb{E}[\hat{\theta}_k^2] + \mathbb{E}[\theta_k^2] - 2\mathbb{E}[\hat{\theta}_k\theta_k] \\ &= w_k^2 \mathbb{E}[\check{y}_k^2] + v_k - 2w_k \mathbb{E}[\check{y}_k\theta_k]\end{aligned}$$

The Law of Total Expectation gives

$$\mathbb{E}[\check{y}_k] = \mathbb{E}[\mathbb{E}[\check{y}_k \mid \theta_k]] = \mathbb{E}[\theta_k] = 0$$

and the Law of Total Variance gives

$$\begin{aligned}\text{Var}(\check{y}_k) &= \mathbb{E}[\text{Var}(\check{y}_k \mid \theta_k)] + \text{Var}(\mathbb{E}[\check{y}_k \mid \theta_k]) \\ &= \mathbb{E}[u_k] + \text{Var}(\theta_k) = u_k + v_k\end{aligned}$$

$$\implies \mathbb{E}[\check{y}_k^2] = \text{Var}(\check{y}_k) + (\mathbb{E}[\check{y}_k])^2 = u_k + v_k$$

Additionally, the Law of Total Expectation also gives

$$\mathbb{E}[\check{y}_k \theta_k] = \mathbb{E}[\mathbb{E}[\check{y}_k \theta_k \mid \theta_k]] = \mathbb{E}[\theta_k \cdot \mathbb{E}[\check{y}_k \mid \theta_k]] = \mathbb{E}[\theta_k^2] = v_k$$

So, substituting these back into our initial expansion, we have

$$\begin{aligned} \mathbb{E}[(\hat{\theta}_k - \theta_k)^2] &= w_k^2 \mathbb{E}[\check{y}_k^2] + v_k - 2w_k \mathbb{E}[\check{y}_k \theta_k] \\ &= w_k^2 (u_k + v_k) + v_k - 2w_k v_k \\ &= w_k^2 u_k + w_k^2 v_k + v_k - 2w_k v_k \\ &= w_k^2 u_k + (w_k^2 - 2w_k + 1) v_k \\ &= w_k^2 u_k + (1 - w_k)^2 v_k \end{aligned}$$

□

Bibliography

- [1] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, USA, 4th ed., 2010.
- [2] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, Massachusetts, 2006.
- [3] D. Duvenaud, *Automatic model construction with Gaussian processes*. PhD thesis, Apollo - University of Cambridge Repository, 2014.
- [4] S. Banerjee, B. Carlin, and A. Gelfand, *Hierarchical Modeling and Analysis for Spatial Data (2nd ed.)*. Cambridge Series in Statistical and Probabilistic Mathematics. CRC Press, 2014.
- [5] M. Gu and Y. Xu, *Fast nonseparable gaussian stochastic process with application to methylation level interpolation*, *Journal of Computational and Graphical Statistics* **29** (2020), no. 2 250–260, [<https://doi.org/10.1080/10618600.2019.1665534>].
- [6] P. J. Lenk, *Bayesian inference for semiparametric regression using a fourier representation*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61** (1999), no. 4 863–879, [<https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00207>].
- [7] M. Lázaro-Gredilla, J. Quiñoner-Candela, C. Rasmussen, and A. Figueiras-Vidal, *Sparse spectrum gaussian process regression*, *Journal of Machine Learning Research* **11** (2010).
- [8] A. Solin and S. Särkkä, *Hilbert space methods for reduced-rank gaussian process regression*, *Statistics and Computing* **30** (2020), no. 2 419–446.
- [9] D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, and G. Zoubin, *Structure discovery in nonparametric regression through compositional kernel search*, in *Proceedings of the 30th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.), vol. 28 of *Proceedings of Machine*

Learning Research, (Atlanta, Georgia, USA), pp. 1166–1174, PMLR, 17–19 Jun, 2013.

- [10] J. Opsomer, Y. Wang, and Y. Yang, *Nonparametric regression with correlated errors*, *Statistical Science* **16** (2001), no. 2 134–153.
- [11] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, Inc., 1999.
- [12] A. van der Vaart and H. van Zanten, *Information rates of nonparametric gaussian process methods*, *Journal of Machine Learning Research* **12** (2011).
- [13] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Springer, Switzerland, 2016.
- [14] A. Gelman, J. B. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis Third edition*. CRC Press, London, England, 2021.
- [15] M. Gu, X. Wang, and J. O. Berger, *Robust Gaussian stochastic process emulation*, *The Annals of Statistics* **46** (2018), no. 6A 3038 – 3066.
- [16] J. Woolley and G. Peters, *The american presidency project*, 2025.
- [17] B. P. Carlin and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman Hall/CRC, 1998.
- [18] C. N. Morris, *Parametric empirical bayes inference: Theory and applications*, *Journal of the American Statistical Association* **78** (1983), no. 381 47–55.
- [19] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025.
- [20] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
- [21] B. Auguie, *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. R package version 2.3.
- [22] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Golemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, *Welcome to the tidyverse*, *Journal of Open Source Software* **4** (2019), no. 43 1686.

- [23] H. Wickham, *Reshaping data with the reshape package*, *Journal of Statistical Software* **21** (2007), no. 12 1–20.
- [24] Y. Wang and C. Ke, *assist: A Suite of R Functions Implementing Spline Smoothing Techniques*, 2023. R package version 3.1.9.
- [25] M. Barrett, *ggokabeito: 'Okabe-Ito' Scales for 'ggplot2' and 'ggraph'*, 2023. R package version 0.1.0.9000.