April 2022

A Functional Optimization Approach to Stochastic Process Sampling

Ryan Matthew Thurman
University of South Florida

A Functional Optimization Approach to Stochastic Process Sampling

by

Ryan Matthew Thurman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Mathematics
with a concentration in Statistics
Department of Mathematics and Statistics
College of Arts and Sciences
University of South Florida

Co-Major Professor: Razvan Teodorescu, Ph.D.
Co-Major Professor: Iuliana Teodorescu, Ph.D.
Andrei Barbos, Ph.D.
Dmytro Savchuk, Ph.D.
Sherwin Kouchekian, Ph.D.

Date of Approval:
April 19, 2022

Keywords: Probability, Optimal Sampling, Large Deviations Theory, Bayesian Estimation, Non-Stationarity, Dynamic Linear Models

DEDICATION

This project is dedicated to a stranger of about 20 years by the name of Terry Gene Thurman, for doing such a poor job raising me that I had to raise myself, thus making me the man I am today, which is to say a strange man with too much passion and even more stress. A lifetime of silence will never be enough.

ACKNOWLEDGMENTS

Beginning from the start of my academic career, I would like to thank Ashley Levine and his parents for giving me a floor to sleep on when I chose to leave home at about 18 years old. I'd like to thank Stephen Warrior and his family, who allowed me to live with them briefly while I began going to community college. It is Christina Warrior that suggested I go to college in the first place, so a special thank you to her. Next, I'd like to thank Will and Kris McCullough for housing me for so many years without asking for much. My days with Will and his mom represent some of my best times, doing tricks on BMX bikes and studying all of the time.

I'd like to thank my band mates from Deathbed December, whom I toured and played drums with in the dirtiest of fashions. This includes Chris Day, Ashley Levine, Shane Davis, and David Anderson. I'd like also to thank the members of the other bands I played drums in, including Particle Motion, Greg and the Gregs, and Dirt Circus. Two of those bands included Austin Peterson, whom I am still friends with to this day. Thank you, Austin, for always taking my ideas and making them better; that, and introducing me to camping, which is an everyday element of my life these days.

I'd like to thank my many girlfriends for giving me something like love as well as a place to sleep for a while. More important members of this group, if only for the time spent with them, include Kelly Britt, Jamie Diehl, and Lauren Keroack. Thank you, girlfriends, for supporting me emotionally while I tried to do the same for you.

I'd like to thank my academic mentors, Drs. Jamie Goldenberg (undergraduate thesis advisor), Gan Ladde (graduate thesis advisor), Stephen Suen (associate chair; RIP), Iuliana and Razvan Teodorescu (dissertation co-major professors), Dmytro "Dima" Savchuk (dissertation committee member and all around nice guy), Andrei Barbos (dissertation chair), and Sherwin Kouchekian (dissertation committee member).

I'd like to also thank the Doctorate of Business Administration folks, who took me from a statistician to a statistical consultant and business owner. Those lunches somehow taught me business. Thank you, Dr. Matt Mullarkey, Dr. Grandon Gill, Michelle Walpole and Lauren Baumgartner. Thank you, Paula Chapman, for giving me a chance to work in a serious research role among combat medics and other soldiers with only a Bachelor's degree in Psychology.

I'd like to thank my grandmother, Ruby "Jean" Thurman, who doesn't like the name Ruby in the first place. She's one of the primary reasons I'm here on Earth right now. I'd also like to thank all of my party friends, which is a group big enough to need its own book for listing. Thanks to a group of statistical women I partied with for years and years, referred to blandly as "The Group." This group included Lizzy Miller, Tiffany Forest, Jodi Gubernat, and Malena Allison. Thanks to my fellow graduate students for giving me laughter in the form of awesomely broken phrases. Thank you to the cooler arms of the U.S. Government for giving me free places to camp while I finish this project. Thank you to my Instagram followers and mates for making me feel somewhat famous for several years. I'm sure this had an overall positive effect on me, give or take. Thank you to a kindred spirit by the name of Topher at Ozo West Boulder and his employees for giving me a place to work from while I was living in a car outside of Boulder, Colorado. And with flailing hands and a rapid mouth, I thank my improvisation friends. This group included Mark Zimmer, Justin Severn, Adam Bakst, and a whole host of other funny people. Thank you, dear kitty cats of the world, for giving me something non-human to love with all of my heart too. This group of fine cats includes Dorian, Basil, The Keroack Cats, The Chicago Cats of Lutz, and The McCullough Cats. Lastly, thank you to all of my students over the years, both good and bad. Some of you have been awesome enough to make up for the lack of awesomeness in most of the others, who, like the Scarecrow in Victor Fleming's 1939 *The Wizard of Oz*, think that pieces of nice paper make knowledge.

TABLE OF CONTENTS

ABSTRACT

The goal of the current research project is the formulation of a method for the estimation and modeling of additive stochastic processes with both linear- and cycle-type trend components as well as a relatively robust noise component in the form of Levy processes. Most of the research in stochastic processes tends to focus on cases where the process is stationary, a condition that cannot be assumed for the model above due to the presence of the cyclical sub-component in the overall additive process. As such, we outline a number of relevant theoretical and applied topics, such as stochastic processes and their decomposition into sub-components, linear modeling techniques, optimal sampling, harmonizable processes, dynamic linear models, Bayesian estimation and modeling, as well as non-parametric inference, all en route to the final chapter where we formulate a protocol for the estimation of this model among the theories of large deviation functionals, optimization, and Bayesian inference.

CHAPTER 1:
THEORETICAL BACKGROUND

1.1 Introduction

The purpose of the current document is to provide a physical space for the author to write and elaborate their ideas, notions, and findings related to the pursuit of an algorithm for trend estimation using an optimal subset of a larger stochastic process. While the author has taken the time and effort to make the contents of this document digestible by as wide an audience as possible, it is believed that the reader will be best served if they have at least a working knowledge of measure-theoretic probability, stochastic processes, Bayesian estimation, and linear modeling; as well as a solid understanding of calculus and algebra. There are a number of proofs in this document, meaning that the reader should also have a relatively good understanding of logic itself, which is, in some ways, the foundation of mathematics.

Chapter 1 is devoted to the theoretical definitions, findings, and results necessary to (somewhat) fully understand the current research project in general. There, the reader will find many common results, as well as some of the more obscure concepts that prove useful in conjunction with the aims of this document.

Chapter 2 is devoted to the applications of the theories and results covered in Chapter 1, including some useful examples and more detailed information not directly related to theory, such as specific examples of the more general concepts outlined in Chapter 1. At the culmination of Chapter 2, the reader should be in a position to understand the aims and goals of the current research project.

Chapter 3 is devoted to the formulation, understanding, and use of an algorithm forming the primary purpose of this document in the first place. In other words, Chapters 1 and 2 exist to bolster our understanding of the concepts at play in Chapter 3. In this

chapter, the reader will find the treatment of a few novel results for trend estimation in non-stationary stochastic processes with cycle- and linear-type trends. Future research is expected to fill in the gap between the purely theoretical results provided here and what industry expects for general use in the form of numeric input-output protocols programmed into a computer for the lay-analyst.

Let us begin our theoretical exploration of required concepts with *stochastic processes*, which largely form the concept basis for the current project. To the uninformed reader perhaps worried about the difficulty of understanding such a concept, begin by noting that stochastic processes are, in many ways, just an extension to the idea of random variables, which most undergraduates learn about to some degree during their studies.

1.2 Stochastic Processes

In the simplest sense, a *stochastic process* is a collection of random variables, possibly infinite in number, having some important index, such as time or space, for example. For our current purpose, let us suppose that our stochastic process is indexed by time, taking values in a subset of non-negative real numbers, $\tau \subseteq \mathbb{R}_+ = [0, \infty)$. If our subset, τ, is countable (i.e., consists of some subset of the natural numbers, possibly even the natural numbers themselves), then our process is said to be a *discrete-time process* or in the context of the current study, a *time series*. In the event that the subset, τ, is not a countable set, then we refer to the process as a *continuous-time process*. Of course, one may assume that it is usually measurement accuracy that creates a time series (as opposed to the process inherently being discrete), with most natural processes being continuous in nature. Notwithstanding this point, there is much to gain from treatment of both discrete- and continuous-time processes as separate concepts.

Since the reader may not be fully familiar with the concept of a random variable in the strictest of senses, some definitions are in line here. Before diving into the definition of a random variable, let us discuss some necessary and preliminary concepts, such as measurability and probability spaces. Suppose that we have both a set, say X, and a sigma-algebra defined on the set, a sigma-algebra (often presented as "σ-algebra") merely

being a collection of subsets (of X in this case) which includes the set itself, is closed under both complements and countable unions. Under the aforementioned conditions, the pair (X, Σ_X) is called a *measurable space* (with Σ_X representing the previously-discussed collection of subsets of X). Now that we have some understanding of what a measurable space is, we can move on to define the concept of a *measurable function*.

Suppose that we are given two measurable spaces, (X, Σ_X) and (Y, Σ_Y), with each set, X and Y, having their own respective sigma-algebras, Σ_X and Σ_Y. A *measurable function*, f, is a mapping from X to Y, such that for every subset of Σ_Y, say $E \subseteq \Sigma_Y$, the inverse image of f (over E and denoted $f^{-1}(E)$) is an element of Σ_X. Stated slightly more mathematically, we have that a function, f, is measurable if

$$f^{-1}(E) \coloneqq \{x \in X | f(x) \in E\} \in \Sigma_X \tag{1}$$

Moving forward from the concept of measurable spaces and measurable functions, we now address the concept of a probability space in route to understanding, in a decently strict sense, random variables. Suppose that we endow a measurable space, for example (X, Σ_X), with a measure, μ, such that the following conditions on μ are granted as true: the measure is non-negative, meaning that for every $E \in \Sigma_X$, we have that $\mu(E) \geq 0$; the measure is equal to 0 over the empty set, a condition usually written as $\mu(\emptyset) = 0$; and the measure is countably additive, meaning that for all countable collections of disjoint subsets of Σ_X, we have that

$$\mu\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i) \tag{2}$$

Now that we have given definition to measures, we are prepared to understand the concept of a probability space. Suppose that we are given a *sample space*, which is merely a collection of all possible events for a given experiment (i.e., a collection of all possible outcomes of an experiment, the "things" that can happen). Suppose that we denote this sample space in the usual way: Ω. Once a sample space is given, we may form, choose, or derive a σ-algebra for Ω, denoting it too in the usual way as $\mathcal{F}$. Suppose, additionally,

that our currently defined space, $(\Omega, \mathcal{F})$, is measurable in the sense previously described. Lastly, suppose that we endow this space with a measure, P, having certain conditions for probabilistic use granted. Namely, we assume that the measure, P, maps $\mathcal{F}$ to the set $[0, 1]$. We refer to such a measure as a probability measure. The triple, $(\Omega, \mathcal{F}, P)$, is referred to as a *probability space*.

We are finally in a position to understand the concept of a *random variable*. To wit, suppose that we are given a probability space, $(\Omega, \mathcal{F}, P)$, and a measurable space, (E, Σ_E), the latter representing the possible values and the collections of possible values of the random variable to be defined. An (E, Σ_E)-valued random variable is a measurable function from Ω to E. Essentially, Ω represents the abstract events that can result from a given experiment, be they numeric (e.g., the boxing match is concluded in less than 11 minutes) or character in nature (e.g., Boxer A beats Boxer B to win the match). The random variable takes such information, which is not usually numeric inherently, and provides us with a numeric measure of the event or events in consideration. In this way, random variables provide us with a means to mathematically describe the world around us, especially those parts of the world dealing in events that cannot be fully described before they occur, namely probabilistic events.

Now that we have sufficient context to understand stochastic processes beyond a definition only, let us venture to better understand certain aspects of stochastic processes that are important for the current project. In other words, let us look at some properties and results for stochastic processes relevant to the goals and aims of the current project. Before venturing too heavily into the theory of stochastic processes without some concrete examples of stochastic processes, let us look at some relevant examples.

1.2.1 Example 1. Wiener Process

Let Z_t be a stochastic process, such that $\tau = [0, \infty)$ and $Z_t \in \mathbb{R}$ for each $t \in \tau$. Suppose, further, that we endow Z_t with the following properties:

1. Initial Value Condition:

   $Z_0 = 0$

2. Independent Increments:

   $Z_u - Z_s$ is independent of $Z_r - Z_q$ whenever $q < r < s < u$

3. Gaussian Increments:

   $Z_{t+u} - Z_t$ is distributed as $N(0, u)$ when $u \geq 0$

4. Continuity of Paths:

   Z_t is continuous in t

From Property 3, we have that $Z_t \sim N(0, t)$, which is a useful property for our current purpose. Since it will be fruitful to determine some information about the moments of this process, let us do so now. From Property 3, we have that

$$E[Z_t] = 0 \tag{3}$$

and

$$V[Z_t] = t \tag{4}$$

Suppose that we are given two arbitrary times, s and t, and would like to determine the covariance between Z_s and Z_t, whenever $s \leq t$.

$$\begin{aligned} COV[Z_s, Z_t] &= E[(Z_s - E[Z_s])(Z_t - E[Z_t])] && (5)\\ &= E[Z_s \cdot Z_t] && (6)\\ &= E[Z_s \cdot (Z_t - Z_s + Z_s)] && (7)\\ &= E[Z_s \cdot (Z_t - Z_s)] + E[Z_s^2] && (8)\\ &= 0 + s && (9)\\ &= s && (10) \end{aligned}$$

where the first equation is true by the definition of covariance, the second is true by the expectation equation above, the third is true by substitution, the forth by factoring, the fifth by the independent increment condition and the variance equation above. As such,

the covariance between two (possibly different) points of a Wiener Process is equal to the value of the minimum of the two times, which is, in this case, s, by the condition that $s \leq t$. In other words, the covariance between two points of a Wiener Process is equal to the variance of the process at the lesser of the two points.

Now that we have some elementary statements about the Wiener Process, let us explore a relatively vast extension of the concept, namely the *Levy Process*.

1.2.2 Example 2. Levy Process

Let Z_t be a stochastic process, such that $\tau = [0, \infty)$ and $Z_t \in \mathbb{R}$ for each $t \in \tau$. Suppose, further, that we endow Z_t with the following properties:

1. Initial Value Condition:

   $Z_0 = 0$ (Almost Surely)

2. Independent Increments:

   $Z_u - Z_s$ is independent of $Z_r - Z_q$ whenever $q < r < s < u$

3. Stationary Increments:

   $Z_{t+u} - Z_t$ follows the same distribution as Z_u when $u \geq 0$

4. Continuity of Probability:

   For any $\epsilon > 0$ and $t \in \mathbb{R}_+$, $\lim_{h \to 0} P(|X_{t+h} - X_t| > \epsilon) = 0$

where the phrase "Almost Surely" (henceforth labeled as "a.s.") simply means that $P(Z_0 = 0) = 1$. In the parlance of probability theory, this statement means that the probabilistic condition is true outside of a set of zero measure. In the event that Property 2 is applied to a sequence of non-overlapping differences, we assume that the increments are mutually independent (i.e., independent across all subsets of the included increments). Once again, the continuity of probability condition ensures that the process is only discontinuous on a set of measure equal to 0, which amounts to saying that we do not *expect* the process to be discontinuous.

To tie the current example to the previous, note that whenever $Z_t - Z_s \sim N(0, t-s)$ and $s \leq t$, the Levy Process reduces to the Wiener Process, thus establishing that the set of all Levy processes contains the set of all Wiener processes.

While closed-form moment equations are not always possible for the Levy Process, we do have the following useful property whenever the Levy Process *does* have finite moments:

$$E[X_{t+s}^n] = \sum_{k=0}^{n} \binom{n}{k} E[X_t^k] E[X_s^{n-k}] \tag{11}$$

which establishes a binomial identity between the moments of the process at $t+s$ and the moments of the process at the points, s and t, separately. This property will prove useful in the pursuit of our current goals, especially in the longer-term study of the concepts.

1.2.3 Infinite Divisibility

Another property of Levy Processes that will prove useful for the current research is the concept of *infinite divisibility*. A probability distribution, say F, is infinitely divisible if, for every $n \in \mathbb{N}_1^\infty = \{1, 2, \cdots\}$, there exists n *Independent and Identically Distributed* (iid) random variables (each represented here by X_{ni}) such that

$$Z_n = \sum_{i=1}^{n} X_{ni} \tag{12}$$

where this sum, Z_n, has the same distribution, F.

If we suppose that Z_t is a Levy Process, such that $t \in [0, \infty)$, and write Z_t as the sum of n increments, we have

$$Z_t = \sum_{i=1}^{n} (Z_{it/n} - Z_{(i-1)t/n}) \tag{13}$$

which indicates that every Levy Process is infinitely divisible, since each of the increments in the sum above is iid by Properties 2 and 3 above. This is a very useful property for our current aims and goals, as it is used throughout the remainder of this document.

1.2.4 Levy-Khinchin Representation for Levy Processes

Let us turn next to a very useful result for Levy Processes, namely the *Levy-Khinchin Representation* of a Levy Process, which should greatly assist the reader in understanding Levy Processes in general. Let us begin by defining the *Characteristic Function* (CF) of a random variable.

Definition 1.1 (Characteristic Function)**.** Let Z_t be a random variable for each fixed value of $t \in [0, \infty)$ (i.e., let Z_t be a stochastic process defined on non-negative real numbers). Then, the following function, mapping the real numbers to the complex numbers, is known as the *characteristic function* of Z_t

$$\phi_{Z_t}(u) = E[e^{iuZ_t}] \tag{14}$$

which is a function that fully determines the probability distribution of the random variable upon which it is calculated, in this case Z_t for each fixed value of t. In the case where the random variable of interest can be represented as the sum of independent random variables, such as the case where the distribution of the random variable is infinitely divisible, the characteristic function of the sum of such random variables can then be written as the product of the characteristic functions across each of the independent random variables. This property of the characteristic function can easily be verified via the following route (where $Y = \sum_{i=1}^{n} X_i$)

$$\begin{aligned}
\phi_Y(u) &= E[e^{iuY}] && (15)\\
&= E[e^{iu\sum_{i=1}^{n} X_i}] && (16)\\
&= E[e^{iuX_1} e^{iuX_2} \cdots e^{iuX_n}] && (17)\\
&= \prod_{i=1}^{n} E[e^{iuX_i}] && (18)\\
&= \prod_{i=1}^{n} \phi_{X_i}(u) && (19)
\end{aligned}$$

where line 1 makes use of the definition of the characteristic function, line 2 makes use of

substitution, line 3 makes use of a very well-known property of the exponential function, line 4 utilizes the property that the expectation over products of independent random variables is equal to the product over each random variable's expectation, and the final line merely uses, once again, the definition of the characteristic function.

Now that the characteristic function has been defined, let us continue our treatment of the Levy-Khinchin Representation of a Levy Process. Let us state the theorem, first, and seek to understand its components, second.

Theorem 1.1 (Levy-Khinchin Representation)**.** If Z_t is a Levy Process defined on $t \in [0, \infty)$, then the characteristic function for Z_t has the following representation

$$\phi_{Z_t}(u) = E(e^{t\psi(u)}) \tag{20}$$

where

$$\psi(u) = ibu - \frac{1}{2}cu^2 + \int_{\mathbb{R}} (e^{iux} - 1)v(dx) - \int_{|x|<1} (iux)v(dx) \tag{21}$$

Proof. □

Before venturing to prove this theorem, let us begin by describing the nature of the form of $\psi(u)$ above. First, let's describe a *Brownian Motion process with drift*, which is merely a process defined by

$$B_t = \mu t + \sigma Z_t \tag{22}$$

where Z_t is a Wiener Process as described previously. Working from the moments of the Wiener Process, we note that

$$E[B_t] = E[\mu t + \sigma Z_t] \tag{23}$$

$$= E[\mu t] + E[\sigma Z_t] \tag{24}$$

$$= \mu t + \sigma E[Z_t] \tag{25}$$

$$= \mu t \tag{26}$$

where line 1 utilized substitution, line 2 uses the additive property of expectation, line

3 uses the property that the expectation over a constant is equal to that constant, as well as the property that $E[aX] = aE[X]$. Lastly, line 4 uses the zero-mean property of the Wiener process. The form obtained indicates that the expected value of a Brownian Motion process with drift is equal to some constant, here μ, multiplied by the given value of time, t. Further, this indicates that the process drifts in time, which is expected, given the definition of the process. Moving on, let's address the variation of this process.

The variance of the process may be determined as follows:

$$V[B_t] = V[\mu t + \sigma Z_t] \tag{27}$$

$$= V[\sigma Z_t] \tag{28}$$

$$= \sigma^2 V[Z_t] \tag{29}$$

$$= \sigma^2 t \tag{30}$$

where the first line uses substitution, the second uses the property that $V[a+X] = V[X]$ whenever a is constant, the third uses the variance property that $V[aX] = a^2V[X]$, and the final line uses the previously obtained result that the variance of a Wiener process, say Z_t is equal to t. From this result, we see that the variance of the Brownian Motion process with drift, drifts in time according to the square of σ.

Given that linear combinations of Normally distributed random variables are, themselves, Normally distributed, we have (for each fixed value of time) that

$$B_t \sim N(\mu t, \sigma^2 t) \tag{31}$$

which implies that the characteristic function for this process is given by

$$\phi_{B_t}(u) = E(e^{iuB_t}) \tag{32}$$

$$= e^{iu\mu t - \frac{1}{2}u^2\sigma^2 t} \tag{33}$$

where the first line uses the definition of the characteristic function and line 2 uses the form of the characteristic function for Normally distributed random variables. Since

the characteristic function fully determines the probabilistic structure of a given random variable, we see, by letting $b = \mu t$ and $c = \sigma^2 t$, that the Levy Process (and more generally, any process with an infinitely divisible distribution) is composed of the sum of a Brownian Motion process with drift and two other as yet not described random variables. It is to these latter two random variables that we turn our attention to next.

En route to describing the second random variable that makes up the Levy Process according to the theorem above, let us describe the *Compound Poisson Process*, which is a process with the following form

$$Z_t = \sum_{i=1}^{N_t} U_i \tag{34}$$

where N_t is a counting Poisson process (with rate λ), meaning that it takes values in the non-negative integers and is non-decreasing in t. Such a Poisson process may be used to model the number of arrivals up to and including time t, be they arrivals to a store, technical service, or any other situation in which individual units or persons arrive. U_i is the size or value associated with the ith arrival. The U_i's are assumed to be independent and identically distributed (with common distribution, say F_{U_i}), taking values in $\mathbb{R}$. Further, it is generally assumed that the U_i's are independent of the underlying Poisson process, N_t. The U_i's may represent the amount of time a customer spends in a store, the amount of money they spend, the amount of time in technical service queue, or any other value that may be associated with the arrivals of a counting Poisson process.

Let $s < t$. Then, the increment $Z_t - Z_s$ is equal to the sum of the U_i's from $i = N_s + 1$ to $i = N_t$ (by the definition above). Since $Z_{t+h} - Z_{s+h}$ has the same distribution as $Z_t - Z_s$ for $h \in \mathbb{N}_0^\infty$, the increments of the compound Poisson process are stationary. One can see that these (increment) distributions are the same from the fact that both increments contain the same number of iid random variables. Further, since the sums generated across non-overlapping time intervals contain non-overlapping independent random variables, namely the U_i's, we have that the compound Poisson process has both independent and stationary increments. Let us now turn to the moments of this process. Using the *Law of Total Expectation*, we may obtain the mean of the process in the following way (letting

$E[U_i] = \mu$ and $V[U_i] = \sigma^2$ for all i)

$$E[Z_t] = E[E[Z_t|N_t]] \tag{35}$$
$$= E[N_t\mu] \tag{36}$$
$$= \mu E[N_t] \tag{37}$$
$$= \mu\lambda t \tag{38}$$

where line 1 uses the aforementioned law, line 2 uses the definition of expectation over sums of independent and identically distributed random variables, line 3 uses the property that $E[aX] = aE[X]$ whenever a is constant, and line 4 uses the expectation form for a Poisson process (namely, N_t). next, we move on to the variance of the compound Poisson process. Using the *Law of Total Variance*, we have that

$$V[Z_t] = E[V[Z_t|N_t]] + V[E[Z_t|N_t]] \tag{39}$$
$$= E[\sigma^2 N_t] + V[\mu N_t] \tag{40}$$
$$= \sigma^2\lambda t + \mu^2\lambda t \tag{41}$$
$$= (\sigma^2 + \mu^2)\lambda t \tag{42}$$

where line 1 uses the mentioned law, line 2 uses the definitions of expectation and variance over sums of independent and identically distributed random variables, line 3 uses the expectation and variance forms for the Poisson process as well as the properties related to constants, and line 4 is merely simplification of the result. Now, we seek to relate the obtained moments to the characteristic function of the compound Poisson process on our way to understanding the second random variable that makes up the Levy process.

Letting Z_t represent a compound Poisson process, as described above, let us venture to compute the characteristic function for this process. Working from the definition of

the characteristic function, we have

$$\phi_{Z_t}(u) = E(e^{iuZ_t}) \tag{43}$$

$$= E[E[e^{iuZ_t}|N_t]] \tag{44}$$

$$= E[E[e^{iu\sum_{i=1}^{N_t} U_i}|N_t]] \tag{45}$$

$$= E[\phi_{U_i}(u)^{N_t}] \tag{46}$$

$$= \sum_{k=0}^{\infty} \phi_{U_i}(u)^k \left(\frac{\lambda^k e^{-\lambda}}{k!}\right) \tag{47}$$

$$= \sum_{k=0}^{\infty} \frac{(\lambda\phi_{U_i}(u))^k e^{-\lambda\phi_{U_i}(u)} e^{\lambda(\phi_{U_i}(u)-1)}}{k!} \tag{48}$$

$$= e^{\lambda(\phi_{U_i}(u)-1)} \sum_{k=0}^{\infty} \frac{(\lambda\phi_{U_i}(u))^k e^{-\lambda\phi_{U_i}(u)}}{k!} \tag{49}$$

$$= e^{\lambda(\phi_{U_i}(u)-1)} \tag{50}$$

$$= e^{\lambda\int_{\mathbb{R}}(e^{ius}-1)dF_{U_i}(s)} \tag{51}$$

where line 1 uses the definition of the characteristic function, line 2 uses the Law of Total Expectation, line 3 uses substitution, line 4 uses the property of the characteristic function over sums of independent and identically distributed random variables, line 5 uses the expectation equation for functions of a Poisson random variable, line 6 utilizes a collection of the terms as well as an equivalence of form, line 7 involves the factoring out of a constant term with respect to k, line 8 uses the fact that the sum of a probability mass function over its domain is equal to 1, while line 9 uses the definition of the characteristic function.

Equating the final result above with the second random variable in the sum comprising the decomposition of the Levy process, we have the following

$$\int_{\mathbb{R}}(e^{iux}-1)v(dx) = \lambda\int_{\mathbb{R}}(e^{ius}-1)dF_{U_i}(s) \tag{52}$$

which implies that (setting $s = x$)

$$v(dx) = \lambda dF_{U_i}(x) \tag{53}$$

$$= \lambda f_{U_i}(x)dx \tag{54}$$

where the final step is true if the distribution function for U_i is differentiable for all relevant values of x, which is to say, all $x \in \mathbb{R}$, given the nature of the integral from which the terms involved come. Since v is commonly referred to as the *Levy measure*, we see that the Levy measure has a scaled relationship with the differential of the distribution function of the value distribution associated with the previously defined compound Poisson process.

Substituting the obtained form for the Levy measure, we have that the third random variable in the sum comprising the Levy decomposition has the following form

$$\int_{|x|<1} iuxv(dx) = \int_{|x|<1} iux\lambda f_{U_i}(x)dx \tag{55}$$

$$= iu\lambda \int_{|x|<1} xf_{U_i}(x)dx \tag{56}$$

which implies that the third random variable making up the sum has the form $X = \lambda \int_{|x|<1} xf_{U_i}(x)dx$, which explicitly represents a scaled version of the expectation equation for U_i over a subset of the real numbers whose absolute value is less than or equal to 1. As such, this random variable is generally taken to represent the jumps in the Levy process with small magnitude.

1.3 Decomposition of Stochastic Processes

Often, a stochastic process (more specifically, a time series) is decomposed into sub-components, each having features that, once viewed collectively (e.g., as a sum or product of the sub-components), are easier to study than the entire process. In other words, the sub-components of an otherwise complex stochastic system, provide the researcher with a means to understanding the overall process, which is often comprised of more easily understood portions, referred to throughout this document as the *components* or *sub-*

components of the overall stochastic process. For our current aims, we will refer to three separate components which additively make up our overall stochastic process: linear trend, cyclical trend, and noise.

For the purpose of the current research, suppose that we are given a stochastic process, S_t, and would like to decompose this process into three separate components, X_t, Y_t, and Z_t, representing the trend, oscillatory/cycle, and noise sub-components of the process, respectively. The primary idea of decomposition in the context of a time series is that a collection of separate processes does a reasonable job of providing context for the original process without the need for more complex analytical procedures. It should be noted that each sub-component is, in turn, itself a time series (or more generally, a stochastic process). Decomposition is also quite useful for seasonally adjusting a time series, which provides a clearer picture of the long-term trends of the process.

1.3.1 Trend Components

Generally, trend is seen as a, possibly non-linear, upward or downward propensity in a stochastic process that is usually persistent and non-repeating (i.e., not repeating in the same way as, say, a cycle component). The process of fitting an Ordinary Least-Squares (O.L.S.) regression equation, which most elementary students of Statistics should be familiar with, is an example trend estimation, where the overall upward and downward trends within a given set of data are of primary concern, so much so that very strong conditions are often placed on the random portions (i.e., noise components) of a given process solely for the sake of obtaining an estimate of trend. Throughout the current research, we assume that the trend sub-component is both deterministic and monotonic in time, with other conditions being assumed where necessary. Let us recall next what it means for the trend sub-component to be *monotonic*.

Definition 1.2 (Monotonic Function)**.** Suppose that are given two arbitrary times, say s and t, such that $s \leq t$. Further, suppose that we are given a function, $X_t : \tau \to \mathbb{R}$. Then, our function, X_t, is said to be *monotonically increasing* whenever $s \leq t \implies X_s \leq X_t$. That is, such a function has outputs which preserves the order of its inputs. If the order

is reversed between inputs and outputs of the function, then the function is said to be *monotonically decreasing*. In other words, a monotonically decreasing function is one such that $s \leq t \implies X_s \geq X_t$. Note that the presence of equality in this definition implies that a monotonic function need not be one-to-one. In other words, the function may not have an inverse.

1.3.2 Cycle Components

Seeing as we are using the term *cycle* to mean a sub-component which is a deterministic Fourier series in time, we may consider this sub-component as actually representing the seasonal portions of a decomposition, since a cycle component is generally viewed as a process with repeating but non-periodic behavior, while a seasonal component is generally viewed as a process with both repeating and periodic behavior, the latter being well in line with the concept of a Fourier series. One cannot fully appreciate the utility of the Fourier series concept without it first being described, so let us describe such a process.

Definition 1.3 (Fourier Series)**.** Let Y_t represent the cycle sub-component of our stochastic process, as described in the previous paragraph. Let P represent the period of this sub-component, such that P is equal to the smallest value of T such that $Y_t = Y_{t+T}$ for every t. Then, by the definition of a Fourier series, we have that

$$Y_t = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi t}{P}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi t}{P}\right) \tag{57}$$

with the values a_n and b_n referred to as the Fourier coefficients of the process, Y_t, such that

$$a_n = \frac{2}{P}\int_P Y_t \cos\left(\frac{2\pi}{P} nt\right) dt \tag{58}$$

and

$$b_n = \frac{2}{P}\int_P Y_t \sin\left(\frac{2\pi}{P} nt\right) dt \tag{59}$$

It may be noted here that a Fourier series is a process that can be represented as an infinite sum of sines and cosines, which well justifies our use of this sub-component to

capture the periodic (i.e., sinusoidal) behavior of our overall process, S_t. Working with such processes usually involves the calculation of the coefficients first, through integration, as well as making use of the even or odd behavior of the process, Y_t. Generally, statistical analysts will assume, based on the nature of the data or the use of a spectral analysis procedure such as the *periodogram* (which is based on the concept of a discrete Fourier transform), that the period, P is known. However, such an assumption generally requires certain strong assumptions about the nature of the cycles in the data. For instance, it is possible that hierarchical/nested cycles exist within the data which have differing periods. It is also possible that the period itself changes over time. Throughout this document, we merely assume that there exists a common period, P, for the Fourier/cycle sub-component of the overall process. We do not assume that this quantity is known.

1.3.3 Noise Components

The final sub-component in consideration is the noise component, which, unlike the previous two components, trend and cycle, is non-deterministic, meaning that we cannot determine it's value before that value has occurred. We will assume throughout this research that the noise sub-component of S_t, denoted Z_t, is either a Levy Process, Wiener Process, or a mixture of the two. It is this sub-component that drives the randomness of the process, allowing us to even use the term *stochastic process* in the first place. The noise sub-component is responsible for the fluctuations in the process not described by either the trend or the cycle sub-components. Generally, noise components have their structure assumed (or at the very least verified approximately using a given set of data), while the other components (i.e., trend and cycle) are estimated through the use of some statistical modeling procedure.

As a final commentary on the relation between the sub-components of a stochastic process, recall that trend describes the long-term non-periodic propensities of the process, while the cycle component describes the repeating periodic behavior and the noise component captures the randomness of the process in a structured fashion (i.e., that which is left over or residual from what was otherwise expected based on trend and cycles alone).

1.4 Linear Regression

One of the primary aims of the current research is to generate a procedure that isolates those points in time of a given stochastic process where trend (in our case, both linear- and cycle-type trend) may be estimated in a simple yet elegant way. In other words, we aim to use simpler procedures available to most undergraduates willing to stay awake long enough, to estimate trend, which further allow us to better understand the overall process itself. Linear regression (as an algorithm) generally seeks to determine the coefficient values (i.e., intercept of the line and slope(s) of the line) that minimizes some quantity of interest, such as the sum of squared residuals; a residual merely being a measure of the difference between a value observed in the data and the value that was expected according to the resulting model, the linear regression.

Let us explore linear regression generally before delving into the concept's specific manifestations. To this end, suppose that are given a (column) vector of outputs,

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2, \\ \vdots \\ y_n \end{pmatrix} \tag{60}$$

Let us suppose also that we are given a matrix of inputs, such that this matrix has rows that represent observations on an experimental unit (e.g., a participant, a business, a day, etc.) and columns that represent observations on a sequence of random vectors (outside of the first column, which is set to 1 for all observations, representing the constant or intercept effect). Mathematically, we write this matrix as (where X is not the same as X_t described above)

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{61}$$

Generally, the matrix $\boldsymbol{X}$, as described above, is known as the *design matrix* for a given regression problem. Moving forward, we define what are known as the *regression coefficients*, which are the values to be fitted (i.e., obtained) by the linear regression procedure. Let us represent these regression coefficients in the form of a vector that will be pre-multiplied by the design matrix to form the linear relationship between the inputs and outputs of our regression problem. Namely, we have that

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \tag{62}$$

and

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{63}$$

where $\boldsymbol{\epsilon}$ is a vector of noise components of equivalent length to our output vector, which in this case is represented by the natural number n. It is in this error term that we venture from the general to the specific, as the assumptions necessary in a given linear regression problem largely come from the nature of the noise components. Some of the assumptions necessary for linear regression in general (i.e., those not borne out of the conditions placed on the noise components) include assuming that the inputs are observed without error (a condition largely assumed more than manifested in reality), the relationship between the inputs and the outputs is linear (hence why we call the process *linear* regression), and that the design matrix be of full rank. The design matrix being of full rank means that we do not expect any of the input variables (i.e., columns 2 through p of the design matrix) to be linearly related to any of the other input variables, a condition which often ensures that our interpretation of the individual regression coefficients is not influenced by any of the other coefficients in the same vector. In other words, we would like to assume that our inputs are largely independent of one another for the sake of

descriptive ease. In the event that one (or more) of our input vectors is linearly related to any of the others, we have a condition known as *multicollinearity*, which can be a real pain for a statistical analyst, as it muddles the information usually sought after most in the linear regression analysis, especially in an applied setting, where many clients cannot easily understand such a limitation.

Other assumptions which *are* borne about by the nature of the noise component vector, include the assumption of *homoscedasticity*, which means that the variation among the errors is constant across the observed values of the inputs vectors. In other words, it is often (but not always) assumed that the variability of the noise components is *not* a function of the input variables. Additionally, it is often (but, once again, not always) assumed that each noise component (i.e., each row of the noise component vector) is independent of each other noise component. Generally, the assumptions outlined above (or, more specifically, their lack of justification in a given regression problem) can create a need for an entirely different analysis procedure, such as the need for non-linear regression, more robust errors, or the use of regularization, to name a few. As such, the assumptions of a given model should be evaluated extensively to determine if the model being used is in fact correct for a given set of data in the first place. Before setting certain conditions on the errors for the sake of a more thorough understanding and exploration of linear regression, let us explore some important results in linear regression theory, namely the Gauss Markov Theorem and its extension to correlated errors, the method known as *Generalized Least Squares*.

1.4.1 The Gauss-Markov Theorem

In preliminary summation of concept, the Gauss-Markov Theorem states that the OLS estimate, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ (to be explained below), is the *Best Linear Unbiased Estimator* (BLUE) of $\boldsymbol{\beta}$, which amounts to saying that this estimate equals the parameter (vector) it is estimating in expectation and has the lowest mean squared error among all such unbiased estimators of the same parameter. Now, let us explore this result in more detail.

Suppose, as before, that we have a vector of outputs, $\boldsymbol{y} \in \mathbb{R}^n$, a matrix of non-random inputs, $\boldsymbol{X} \in \mathbb{R}^{n \times k}$, a vector of noise components that induce randomness on the outputs, $\boldsymbol{\epsilon} \in \mathbb{R}^n$, and a non-random but unobservable parameter vector to be estimated via the linear regression procedure, $\boldsymbol{\beta} \in \mathbb{R}^k$. Suppose, also, that these separate elements of the problem are related via the following equation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{64}$$

As mentioned above, the assumptions that are made on the noise components significantly affect the nature of the regression procedure, often ruling out simpler procedures for more complicated yet more accurate procedures. For the sake of the current theorem, let us make the following assumptions on the moments of the errors of the model:

- $E[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{0}$
- $V[\boldsymbol{\epsilon}|\boldsymbol{X}] = \sigma^2 \boldsymbol{I}_n$

These assumptions amount to saying that each of the noise components has a mean of zero, that each of the noise components is uncorrelated with each other, and that the variance of each component is equal to the same value, namely σ^2.

Suppose that the coefficient vector, $\hat{\boldsymbol{\beta}}$, is a linear estimator of $\boldsymbol{\beta}$. Then, we have that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{C}(\boldsymbol{X})\boldsymbol{y} \tag{65}$$

where the matrix $\boldsymbol{C} \in \mathbb{R}^{k \times n}$ is granted dependence on the design matrix, $\boldsymbol{X}$, but not on the parameter vector being estimated, $\boldsymbol{\beta}$. Since the theorem is concerned with the best linear *unbiased* estimator, let us assume that

$$E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \tag{66}$$

where the lack of dependence of this unbiasedness on the design matrix has been explicitly stated. To establish that the estimator of $\boldsymbol{\beta}$ given at the beginning of this section is in

fact the best linear unbiased estimator, we must establish that this estimate minimizes the following quantity (for all choices of $\boldsymbol{\lambda} \in \mathbb{R}^k$)

$$V[\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}] = E[(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}} - E(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}))^2] \tag{67}$$

$$= E[(\boldsymbol{\lambda}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2] \tag{68}$$

which amounts to saying that the best linear unbiased estimator is one that not only minimizes the variance of the estimate $\hat{\boldsymbol{\beta}}$ for all choices of $\boldsymbol{\lambda}$ but also (equivalently, due to unbiasedness) minimizes the mean squared error of the estimation. In this way, such an estimator is known to be the lowest variance estimator among all similarly defined estimators, making it the *best* estimator according to this lowest variance/mean squared error criterion. Note that $\boldsymbol{\lambda}'$ represents the matrix/vector transpose of $\boldsymbol{\lambda}$.

To establish that the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is in fact the best such estimator according to the criteria above, let us define some other unbiased estimator to be compared to the given one. More specifically, let

$$\check{\boldsymbol{\beta}} = \boldsymbol{C}_1(\boldsymbol{X})\boldsymbol{y} = ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})\boldsymbol{y} \tag{69}$$

where $\boldsymbol{C}_1 \in \mathbb{R}^{k\times n}$ has been provided a subscript to differentiate it from the more general $\boldsymbol{C}(\boldsymbol{X})$ previously used. Additionally, $\boldsymbol{D} \in \mathbb{R}^{k\times n}$.

Since this estimator, different from the one stated to be optimal, must be unbiased, let us determine the nature of the (non-random) matrix $\boldsymbol{D}$.

$$E[\check{\boldsymbol{\beta}}] = E[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})\boldsymbol{y}] \tag{70}$$

$$= E[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \tag{71}$$

$$= E[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})(\boldsymbol{X}\boldsymbol{\beta})] + E[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})(\boldsymbol{\epsilon})] \tag{72}$$

$$= E[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})(\boldsymbol{X}\boldsymbol{\beta})] \tag{73}$$

$$= E[((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta})] + E[(\boldsymbol{D})(\boldsymbol{X}\boldsymbol{\beta})] \tag{74}$$

$$= (\boldsymbol{I}_k + \boldsymbol{D}\boldsymbol{X})\boldsymbol{\beta} \tag{75}$$

which implies that the matrix product $\boldsymbol{DX}$ must equal the matrix $\mathbf{0}_{k\times n}$ for the estimator $\check{\boldsymbol{\beta}}$ to be unbiased. Step 1 utilized the definition of the estimator given, step 2 utilized the relationship given for $\boldsymbol{y}$ (i.e., the linear model form), step 3 uses factoring of the terms, step 4 uses the fact that $E[\boldsymbol{\epsilon}] = \mathbf{0}_n$ and $((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})$ is constant/non-random. Step 5 uses, once again, factoring, while the final step is merely a simplification of the previous step that uses the property $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, where the identity matrix has the same dimension as the product of the inverse of $\boldsymbol{A}$ with $\boldsymbol{A}$. In the current result, the identity matrix has dimension $k \times k$, hence the subscript k.

Now, we establish the minimum variance property of the OLS estimate (against the arbitrary but different estimator $\check{\boldsymbol{\beta}}$)

$$
\begin{aligned}
V[\check{\boldsymbol{\beta}}] &= V[\boldsymbol{C}_1(\boldsymbol{X})\boldsymbol{y}] && (76)\\
&= \boldsymbol{C}_1(\boldsymbol{X})V[\boldsymbol{y}]\boldsymbol{C}_1'(\boldsymbol{X}) && (77)\\
&= \boldsymbol{C}_1(\boldsymbol{X})(\sigma^2\boldsymbol{I}_n)\boldsymbol{C}_1'(\boldsymbol{X}) && (78)\\
&= \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})' && (79)\\
&= \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{D}') && (80)\\
&= \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{DX}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{DX})' + \boldsymbol{DD}') && (81)\\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \sigma^2\boldsymbol{DD}' && (82)\\
&= V[\hat{\boldsymbol{\beta}}] + \sigma^2\boldsymbol{DD}' && (83)
\end{aligned}
$$

where step 1 uses the definition of the alternative estimator, step 2 uses the variance equation for the pre-multiplication of a random vector by a non-random matrix, step 3 uses the variance equation of the errors (since $\boldsymbol{y}$ receives all of its randomness from the errors of the model), step 4 uses the definition of $\boldsymbol{C}_1(\boldsymbol{X})$, step 5 uses a well-known property of the matrix transpose, step 6 utilizes factoring, step 7 uses the constraint imposed by the need for unbiasedness (determined in the expectation result above), and the final step involves the substitution of the variance equation for the best estimator,

which is derived as follows:

$$V[\hat{\boldsymbol{\beta}}] = V[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \tag{84}$$

$$= ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')V[\boldsymbol{y}]((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \tag{85}$$

$$= ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\sigma^2\boldsymbol{I}_n)((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \tag{86}$$

$$= \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \tag{87}$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{88}$$

Thus, it has been established that the estimator, here referred to by $\hat{\boldsymbol{\beta}}$, is in fact the best linear unbiased estimator among all similarly defined estimators. Stated differently, it has been established that the OLS estimator is BLUE. Of course, this result, while quite general in its own right, is limited by the assumptions made on the errors of the model. We now relax these assumptions in line with the needs of the current research by presenting and detailing the *Generalized Least Squares* procedure in the following section.

1.4.2 Generalized Least Squares

In order for the Gauss-Markov Theorem to be valid for a given problem, as stated above, the errors of our model must have a mean of zero (i.e., $E[\boldsymbol{\epsilon}] = 0$) and must be spherical, which is to say that the errors must be uncorrelated with each having the same (finite) variance (i.e., $V[\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{I}_n$). It should be noted, of course, that these conditions do not include a need for the errors to be Normally distributed or independent and identically distributed. However, if one wishes to apply the theorem to Wiener process type errors or any other error process having correlation between the errors, then the assumptions of the model will not be met and any resulting OLS regression estimates may not be the best linear unbiased estimates. Fortunately, the theorem was extended by Aitken in his 1936 work "On Least-Squares and Linear Combinations of Observations" to cases where the errors have some degree of correlation between them. Let us now explore Aitken's result in more detail.

Suppose that the output vector, $\boldsymbol{y}$, the design matrix, $\boldsymbol{X}$, and the regression coeffi-

cients, $\boldsymbol{\beta}$, are as described in the previous section, and that our wish to estimate $\boldsymbol{\beta}$ with the regression model is the same as in that section. Additionally, we assume, as we did before, that the inputs/design matrix and outputs are related via the following linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{89}$$

where $\boldsymbol{\epsilon}$ represents a vector of noise components having the following structure

- $E[\boldsymbol{\epsilon}|\boldsymbol{X}] = \mathbf{0}$
- $V[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{\Sigma}$

where $\boldsymbol{\Sigma}$ is assumed known and non-singular (as it would be in the case of the Wiener process, where $\Sigma_{ij} = E[w_{t_i} w_{t_j}] = \min(t_i, t_j)$).

Under these conditions, we obtain an estimate of the coefficient vector, $\boldsymbol{\beta}$, by minimizing the square of the Mahalanobis distance between the points of the output vector ($\boldsymbol{y}$) and the points expected by the linear regression ($\boldsymbol{X}\boldsymbol{\beta}$) while accounting for the covariance structure of the model (namely, the nature of $\boldsymbol{\Sigma}$). Thus,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{90}$$

Instead of attempting to prove the optimality of this estimate from this point, let us first observe that, since the covariance matrix, $\boldsymbol{\Sigma}$, is both symmetric and positive-definite, it has a unique Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$. Then, we may write an updated model as follows:

$$\boldsymbol{L}^{-1}\boldsymbol{y} = \boldsymbol{L}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{L}^{-1}\boldsymbol{\epsilon} \tag{91}$$

where the invertibility of L is implied by the invertibility of the covariance matrix $\boldsymbol{\Sigma}$.

Now, for this updated model, we may write the sum of squared residuals (noting that $E[\boldsymbol{L}^{-1}\boldsymbol{\epsilon}] = \boldsymbol{L}^{-1}E[\boldsymbol{\epsilon}] = \mathbf{0}$) in the manner displayed at the top of the following page, where step 1 uses factoring, step 2 uses association as well as a common property of the transpose, step 3 utilizes known relationships between transposes and inverses, and the

final step makes use of the relationship between the Cholesky decomposition of $\boldsymbol{\Sigma}$ and $\boldsymbol{L}$.

$$(\boldsymbol{L}^{-1}\boldsymbol{y} - \boldsymbol{L}^{-1}\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{L}^{-1}\boldsymbol{y} - \boldsymbol{L}^{-1}\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{L}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}))'(\boldsymbol{L}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})) \tag{92}$$

$$= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{L}^{-1})'(\boldsymbol{L}^{-1})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{93}$$

$$= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{L}\boldsymbol{L}')^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{94}$$

$$= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{95}$$

As mentioned above, two relevant properties of matrices have been used: $(\boldsymbol{A}\boldsymbol{B})' = \boldsymbol{B}'\boldsymbol{A}'$ and $(\boldsymbol{A}')^{-1} = (\boldsymbol{A}^{-1})'$. Thus, minimizing the sum of squared residuals for the updated model is equivalent to minimizing the squared Mahalanobis distance for the original model. Exploring the first two moments of the errors of this updated model, we have

$$E[\boldsymbol{L}^{-1}\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{L}^{-1}E[\boldsymbol{\epsilon}] \tag{96}$$

$$= \boldsymbol{0} \tag{97}$$

and

$$V[\boldsymbol{L}^{-1}\boldsymbol{\epsilon}|\boldsymbol{X}] = (\boldsymbol{L}^{-1})V[\boldsymbol{\epsilon}|\boldsymbol{X}](\boldsymbol{L}^{-1})' \tag{98}$$

$$= (\boldsymbol{L}^{-1})\boldsymbol{\Sigma}(\boldsymbol{L}^{-1})' \tag{99}$$

$$= (\boldsymbol{L}^{-1})(\boldsymbol{L}\boldsymbol{L}')(\boldsymbol{L}^{-1})' \tag{100}$$

$$= (\boldsymbol{L}^{-1}\boldsymbol{L})(\boldsymbol{L}'(\boldsymbol{L}')^{-1}) \tag{101}$$

$$= \boldsymbol{I}_n \tag{102}$$

which shows that the updated model meets the assumptions of the Gauss-Markov Theorem, specifically the zero mean condition for the noise vector of the process and the sphericity condition for the variance structure of the noise vector. As such, the best linear unbiased estimate of $\boldsymbol{\beta}$ is given by the following, where $\boldsymbol{y}_{updated} = \boldsymbol{L}^{-1}\boldsymbol{y}$, $\boldsymbol{X}_{updated} =$

$\boldsymbol{L}^{-1}\boldsymbol{X}$, and $\boldsymbol{\epsilon}_{updated} = \boldsymbol{L}^{-1}\boldsymbol{\epsilon}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'_{updated}\boldsymbol{X}_{updated})^{-1}\boldsymbol{X}'_{updated}\boldsymbol{y}_{updated} \tag{103}$$

$$= ((\boldsymbol{L}^{-1}\boldsymbol{X})'\boldsymbol{L}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{L}^{-1}\boldsymbol{X})'(\boldsymbol{L}^{-1}\boldsymbol{y}) \tag{104}$$

$$= ((\boldsymbol{X})'(\boldsymbol{L}^{-1})'\boldsymbol{L}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X})'(\boldsymbol{L}^{-1})'(\boldsymbol{L}^{-1}\boldsymbol{y}) \tag{105}$$

$$= ((\boldsymbol{X})'(\boldsymbol{L}\boldsymbol{L}')^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X})'(\boldsymbol{L}\boldsymbol{L}')^{-1}\boldsymbol{y}) \tag{106}$$

$$= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})\boldsymbol{y} \tag{107}$$

Note that step 1 uses the form of the best linear unbiased estimator of $\boldsymbol{\beta}$ given in the Gauss-Markov result of the previous section, step 2 deals with substitution for the sake of relating the updated model to the original model, steps 3 and 4 utilize the previously-mentioned transposition and inverse properties for suitable matrices, and the final step involves the substitution of the covariance matrix based on its Cholesky decomposition.

Thus, it has been established that the Generalized Least Squares form of regression, as presented here, provides a best linear unbiased estimator for the extension of the Gauss-Markov Theorem involving correlated random errors. To establish the *unbiased* nature of this result, let us compute the expected value of the given optimal estimator,

$$E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = E[(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\Sigma^{-1})\boldsymbol{y}|\boldsymbol{X}] \tag{108}$$

$$= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})E[\boldsymbol{y}|\boldsymbol{X}] \tag{109}$$

$$= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})\boldsymbol{\beta} \tag{110}$$

$$= \boldsymbol{\beta} \tag{111}$$

where the fact that $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$ are non-random has been used, as well substitution. Thus, it has been established that the Generalized Least Square estimator is unbiased. By the Gauss-Markov Theorem, we also know that this estimator has the lowest variance of all such unbiased estimators. For the sake of being thorough, let us compute the variance of

this estimator:

$$V[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = V[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})\boldsymbol{y}|\boldsymbol{X}] \tag{112}$$

$$= ((\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})V[\boldsymbol{y}|\boldsymbol{X}]((\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})' \tag{113}$$

$$= ((\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})V[\boldsymbol{\epsilon}|\boldsymbol{X}]((\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})' \tag{114}$$

$$= ((\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}((\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1})' \tag{115}$$

$$= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \tag{116}$$

where step 1 uses substitution, step 2 uses the property $V[\boldsymbol{A}\boldsymbol{y}] = \boldsymbol{A}V[\boldsymbol{y}]\boldsymbol{A}'$, step 3 uses the fact that the output vector receives all of its randomness from the vector of noise components, step 4 uses the form given for the variance/covariance of the errors, and step 5 involves the use of the transpose/inverse properties for suitable matrices already mentioned and used several times within this document previously.

While the Gauss-Markov Theorem result is useful for regression problems involving a linear trend component and uncorrelated noise components each having the same finite variance, it is not readily equipped to handle models involving more complicated error structures as well as situations involving periodicity in signal. Further, the extended form of Ordinary Least Squares known as Generalized Least Squares, while being equipped to handle correlated errors, does not handle situations with non-linear trend, such as models that have a periodic component. These limitations directly indicate the need for new modeling and estimation protocols for stochastic processes with both linear- and cycle-type trend in the presence of general noise structures, namely Levy process noise. To better understand the exact nature of these stated limitations, let us explore linear regression with Gaussian and non-Gaussian errors through examples and exposition.

1.4.3 Linear Regression with Gaussian Errors

Let us begin this section with the simplest and most well-known (and most ubiquitously taught) linear regression procedure, namely Ordinary Least-Squares (O.L.S.) regression. This elementary procedure seeks to determine the values of the regression coefficient

vector by minimizing the sum of squared errors. More specifically, we seek to minimize the quantity

$$SSE(\boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \tag{117}$$

which amounts to saying that we wish to minimize the vertical distance between our observed output values, $\boldsymbol{y}$, and what is expected (or predicted) based on the linear regression model, $\hat{\boldsymbol{y}} \coloneqq E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$. Note that we refer to the quantity $\boldsymbol{y} - \hat{\boldsymbol{y}}$ as the *residual* vector. The residual vector provides us with information about how well (or not-so-well) the chosen model conforms to the nature of the data. The larger the residual for a given data point, the further that point is from what is expected of it based on the chosen model. This is why residuals form the computational basis for many of the modeling protocols in statistical theory: they numerically indicate how wrong or right we are in our model choices. Let us now continue our treatment of linear regression with Gaussian noise.

Differentiating the sum of squares equation above (with respect to the vector $\boldsymbol{\beta}$), we find that the values of the regression coefficient vector may be estimated using the following equation

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{118}$$

which directly places an invertibility/rank condition on the design matrix (as discussed previously). Note that $\boldsymbol{X}'$ is the *transpose* of $\boldsymbol{X}$, meaning that $\boldsymbol{X}'$ is a matrix whose rows are the columns of $\boldsymbol{X}$ and whose columns are the rows of $\boldsymbol{X}$. Since the nature of a given regression procedure depends almost entirely on the conditions or assumptions we place on its parts, let us explore what is typically assumed in an O.L.S. regression context.

In general, we must assume that our regression model is correctly specified, which amounts to saying that we must assume that the relationship between our given inputs and outputs actually follows a linear model. In the context of O.L.S., we must assume that the expected value of the noise component vector conditioned on the design matrix has

a value of zero: $E[\boldsymbol{\epsilon}|\boldsymbol{X}] = \mathbf{0}$. By the *Law of Total Expectation* (i.e., $E(E[\boldsymbol{\epsilon}|X]) = E[\boldsymbol{\epsilon}]$), this condition further implies that $E[\boldsymbol{\epsilon}] = \mathbf{0}$. Consistent with what has been mentioned previously, we must also assume that the design matrix is of full rank (almost surely). We must assume that the noise vector is *Spherical*, meaning that $V[\boldsymbol{\epsilon}|\boldsymbol{X}] = \sigma^2\boldsymbol{I}_n$. This latter condition implies that the spread/variation in $\boldsymbol{\epsilon}$ is the same across the various values of the design matrix, $\boldsymbol{X}$, and that the noise components are uncorrelated with one another. Finally, for certain properties of the O.L.S. (as most know it) to be true, we usually assume that the noise components are Normally distributed when conditioned on the design matrix: $\boldsymbol{\epsilon}|\boldsymbol{X} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$.

Under the conditions and assumptions outlined above, we may determine certain distributional properties of the estimated/expected regression coefficient vector, $\hat{\boldsymbol{\beta}}$, which will help us to understand the true nature of O.L.S. regression. To this end, we have the following derivation proof, which has largely been presented elsewhere in this chapter.

Theorem 1.2. If the assumptions outlined above for O.L.S. are valid, then

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Proof. Let us begin by showing that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. In other words, let us show that $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] && (119)\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{y}] && (120)\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} && (121)\\ &= \boldsymbol{\beta} && (122) \end{aligned}$$

where line 2 made use of the fact that the design matrix is assumed fixed/constant, line 3 made use of the distributional properties of the noise vector, and line 4 made use of the unity definition for matrices and their inverses. Next, let us establish the variance of the

estimated coefficient vector:

$$V[\hat{\boldsymbol{\beta}}] = V[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \tag{123}$$

$$= V[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \tag{124}$$

$$= V[\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}] \tag{125}$$

$$= V[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}] \tag{126}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'V[\boldsymbol{\epsilon}]((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \tag{127}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I}_n)((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \tag{128}$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{129}$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{130}$$

where line 2 made use of the definition of $\boldsymbol{y}$, line 3 made use of unity property for matrices and their inverses, line 4 made use of the variance property $V[\boldsymbol{a} + \boldsymbol{x}] = V[\boldsymbol{x}]$ whenever $\boldsymbol{a}$ is constant, as $\boldsymbol{\beta}$ is in the current proof, line 5 made use of the property $V[\boldsymbol{A}\boldsymbol{X}] = \boldsymbol{A}V[\boldsymbol{X}]\boldsymbol{A}'$, line 6 made use of the assumptions placed on the noise vector variance, and finally, line 7 made use, once again, of the unity property for matrices. Thus, the expected value and variance of $\boldsymbol{\beta}$ have been established. The Normality of $\hat{\boldsymbol{\beta}}$ can easily be established from the Normality of $\boldsymbol{\epsilon}$, which induces all of the randomness present in the O.L.S. procedure (according to the assumptions of the procedure). □

In a more general context, we could take the phrase *Linear Regression with Gaussian Errors* to mean the following model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{131}$$

$$\boldsymbol{\epsilon} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{132}$$

which is, perhaps, the most general form for the phrase (if fewer assumptions are placed on the moments of the noise vector). However, *Linear Regression with Gaussian Errors* is usually taken as synonymous with *Ordinary Least-Squares Regression.* If only for the sake

of mathematical humoring and exploration, let us explore the nature of the coefficient vector under these more general conditions. Let us, still, assume that $\boldsymbol{X}$ is constant and full rank in the manner previously described. Let us also assume that the model is correctly specified for a given problem.

Beginning with the goal of minimizing the same objective function as before, namely the sum of squared errors, we obtain the following

$$\frac{\partial}{\partial \boldsymbol{\beta}} SSE(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 \tag{133}$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}} [(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})] \tag{134}$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}} [(\boldsymbol{y}' - (\boldsymbol{X}\boldsymbol{\beta})')(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})] \tag{135}$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}} [\boldsymbol{y}'\boldsymbol{y} + (\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{y}'(\boldsymbol{X}\boldsymbol{\beta}) - (\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{y}] \tag{136}$$

$$= 2(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} - 2\boldsymbol{X}'\boldsymbol{y} \tag{137}$$

Setting this final quantity to zero, for the sake of minimization, we obtain

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y} \tag{138}$$

which implies that we should set $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ to minimize the aforementioned objective function. It should be obvious from the quadratic nature of the objective function, that setting the partial derivative (with respect to $\boldsymbol{\beta}$) equal to zero *does* in fact minimize the function, as opposed to maximizing it. Further, it has been established that the optimal value $\hat{\boldsymbol{\beta}}$ is not influenced by our more general assumptions on the noise vector, since its value is the same as what was previously determined.

Continuing with the more general assumptions on the noise vector outlined above, we seek to determine the distributional properties of the regression coefficient vector, $\hat{\boldsymbol{\beta}}$. This includes the calculation of the mean and variance of the regression coefficient vector for this more general model. Here, we find that allowing the noise vector to have a non-zero mean induces biasedness on the expected value of the coefficient estimate vector. We also note that the variance of the estimate vector, while not having any additional

dependencies, does become more complicated than what was seen in any of the cases outlined so far. Let us compute the mean and variance now.

$$E[\hat{\boldsymbol{\beta}}] = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \tag{139}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{y}] \tag{140}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] \tag{141}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mu}_{\boldsymbol{\epsilon}}) \tag{142}$$

$$= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\mu}_{\boldsymbol{\epsilon}} \tag{143}$$

which establishes that, in the event that the mean of the noise vector is non-zero, we induce bias on the estimate of the regression coefficient vector, with the bias being equal to $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\mu}_{\boldsymbol{\epsilon}}$ (i.e., the bias is a function of both the design matrix and the expected value of the noise vector).

Continuing our calculations for the variance of the estimate, we obtain the following equivalences.

$$V[\hat{\boldsymbol{\beta}}] = V[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \tag{144}$$

$$= V[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \tag{145}$$

$$= V[\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}] \tag{146}$$

$$= V[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}] \tag{147}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'V[\boldsymbol{\epsilon}]((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \tag{148}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\Sigma_{\boldsymbol{\epsilon}})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \tag{149}$$

Note that in this formulation, the noise vector has not been assumed spherical, the noise components have not been assumed independent, nor have we assumed that the noise vector has a zero mean. Finally, since linear combinations of Normally distributed random vectors are also themselves Normally distributed, we have established the follow-

ing property (whenever $\boldsymbol{\epsilon} \sim N(\boldsymbol{\mu}_{\boldsymbol{\epsilon}}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$):

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\mu}_{\boldsymbol{\epsilon}}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\Sigma_{\boldsymbol{\epsilon}}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')') \tag{150}$$

Now that we have had a look at linear regression with Gaussian noise, let us look at several variations of this overall concept. To guide this discussion, let us focus on the assumptions for O.L.S. regression and how our model(s) must change to accommodate more flexible circumstances. In the event that our model is not correctly specified (i.e., we are attempting to use a linear regression to model a non-linear phenomenon), we will have to consider non-linear models, which will depend heavily on the nature of the non-linearity in the data. In the event that some or all of our input variables, which are manifested in the design matrix, are correlated with the noise components, our estimates (i.e., $\hat{\boldsymbol{\beta}}$) become invalid, as we have not modeled this relationship between inputs and noise. In some cases, a rank-deficient design matrix may completely prevent us from obtaining estimates, such as for the case when $(\boldsymbol{X}^T\boldsymbol{X})$ is not invertible, or may create a need for us to use a different type of regression, such as *Ridge Regression*. One of the current authors has had to employ ridge regression in a consulting project where most (if not all) of the inputs were heavily correlated with one another.

Continuing the discussion of generalities as they related to linear regression with Gaussian noise, let's look at the sphericity condition often imposed on the noise components. When the noise components of our model are heteroscedastic (i.e., they are inconsistent across the values of our design matrix), we can usually employ *Weighted Least-Squares*, which amounts to saying that we can adjust the sum of squared errors by the inverse of the residual variance for each noise component. Additionally, one may wish to use a more robust structure for the noise components to account for the lack of consistency among them, a practice often used in economic research (based on personal experience of the author). When our noise components are correlated, we may wish to use *Generalized Least-Squares*. The final assumption to be discussed more generally is the relaxing of the condition of Normality for the noise component vector distribution. This, is the point of the following section.

1.4.4 Linear Regression with Non-Gaussian Errors

Often, the noise component vector of our model *cannot* be Normally distributed, such as in cases when our output vector is nominal, ordinal, or some other measurement type that cannot be Normally distributed in the first place. Recall that we do not usually place our distributional assumptions on our output vector, but we *do* place them on our noise vector, which in turn, allows us to model the relationship(s) between our inputs and outputs. One of the most popular procedures for modeling input/output relationships in the presence of non-Gaussian noise is the *Generalized Linear Model* (G.L.M.). Let us begin with a formulation of this model.

Suppose that we are given an output vector, $\boldsymbol{y}$ and a design matrix, $\boldsymbol{X}$, which houses the input variable information for our problem. Suppose that we relate our inputs and outputs using what is often referred to as a *link* function:

$$E[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{\mu}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) \tag{151}$$

where g is such that its inverse exists. This formulation being more general than the O.L.S. case should be obvious to the reader. Our output vector, $\boldsymbol{y}$ is presumed to be generated by an exponential family distribution, which includes a wide array of distributions, such as the Normal, Exponential, Chi-Squared, Bernoulli, and Poisson, to name a few. The link function chosen for a given problem depends heavily on ones choice for the (exponential family) probability distribution. The term $\boldsymbol{X}\boldsymbol{\beta}$ is known as the *linear predictor* and is usually denoted by $\boldsymbol{\eta}$. Let us look at some special cases of the G.L.M. model.

1.4.5 Example 1. Linear Regression with Gaussian Errors

Suppose that we let $\boldsymbol{y} \sim N(\boldsymbol{\eta}, \sigma^2\boldsymbol{I}_n)$ and $\boldsymbol{y} = \boldsymbol{\mu}(\boldsymbol{\eta}) + \boldsymbol{\epsilon} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$ (i.e., the assumed form for O.L.S. regression). In this case, g is an identity function, $E[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta}$, and $V[\boldsymbol{y}|\boldsymbol{X}] = \sigma^2\boldsymbol{I}_n$.

1.4.6 Example 2. Binary Logistic Regression

Suppose that each element of our output vector, $\boldsymbol{y}$, follows a Bernoulli distribution: $y_i \sim Bernoulli(p, pq)$, such that the probability of success is the same for each y_i and that the y_i's are independent of one another, each having variance $p \cdot q$, where $q = 1 - p$. In this case, we have the following equivalences:

$$\eta_i = (\boldsymbol{X}\boldsymbol{\beta})_i = g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) \tag{152}$$

where our link function has been set to the familiar *Logit* function, a function useful for mapping the set $(0, 1)$ into the real numbers. Note that, being Bernoulli, our output vector will consist of a sequence of 0's and 1's, where a value of 1 is generally viewed as a success (i.e., a value representing the *presence* of some characteristic, such as a smoker, a win in a baseball game, an adopter of some technology, etc.). Under this formulation, we may write our mean function, μ_i in the following manner

$$\mu_i = \frac{1}{1 + \exp\left(-\boldsymbol{X}\boldsymbol{\beta}\right)} \tag{153}$$

which expresses nicely how the inputs are mapped to the outputs using an S-curve.

1.4.7 Example 3. Poisson Regression

Suppose that our output vector, $\boldsymbol{y}$, consists of a sequence of non-negative integers (i.e., $y_i \in \{0, 1, 2, 3, \ldots\}$ for each $i \in \{1, 2, \ldots, n\}$), each representing the number of occurrences in a fixed span of time or space. This being a typical situation for the use of the Poisson distribution, let us suppose that $y_i \sim Poisson(\lambda, \lambda)$, such that $\lambda \in (0, \infty)$. Under this scheme, one often used link function is the natural log. The following equivalences reflect this choice.

$$\eta_i = (\boldsymbol{X}\boldsymbol{\beta})_i = g(\mu_i) = \ln(\mu_i) \tag{154}$$

In an applied setting, an analyst using Poisson regression should make sure that each

y_i is not influenced by the value of any other y_i, but is instead (possibly) influenced by the inputs of the model, reflected in the design matrix. In the case where the data of a given problem do not meet the parametric requirement of the Poisson distribution, namely that the mean and the variance are equal, one may wish to instead use Negative Binomial regression, which is, in some ways, an extension of Poisson regression where overdispersion is allowed (i.e., the mean and variance are *not* equal).

More examples of linear regression with non-Gaussian errors can always be provided. However, the examples above should give the reader some idea of how the concept works and what type of problems can be managed with such a class of models. Let us now turn our attention to some concepts in optimal sampling theory.

1.5 Optimal Sampling

In general, observations on a stochastic process (i.e., the sample paths or realization of the process) are measured using *periodic sampling* (also referred to as *Riemann Sampling*), which is to say at a constant rate. In other words, the process is recorded at equally spaced points in time, such as $t_0, t_0 + h, t_0 + 2h, \ldots$. While this form of sampling is convenient, and as a result ubiquitous, it may not be the most efficient or unambiguous method of sampling the continuous signal of a stochastic process. For example, there may exist two or more possible processes that share the same sample path over the time horizon(s) considered, creating an issue of ambiguity for our accurate measurement of the process, which in turn may drastically alter the quality of our predictions based on the given sample path.

Going as far back as the late 1940's, Claude Shannon observed that a process that doesn't contain frequencies higher than B hertz (i.e., B cycles per second) is completely determined if the sampling rate is greater than $2B$ hertz, the quantity $2B$ being referred to as the Nyquist rate. Sampling a process at a rate less than this requirement may lead to a phenomenon known as *aliasing*, which as described above, creates a situation of ambiguity in the determination of the process. Of course, this sampling frequency requirement concerns sampling the process in a periodic way, so there is, in general, hope

for sparser sampling regimes that sample the process non-periodically based on some control rule(s). Regardless of this hope for more efficient sampling, non-uniform samples (i.e., those that are not periodic), must still have an *average* sampling rate greater than $2B$ hertz (Marvasti, 2001).

To explore the various types of non-uniform sampling, let us begin with the concept of *Lebesgue Sampling* if only for its simplicity. In this type of sampling protocol, the process is sampled only when its realized values pass some predefined threshold (usually in an absolute sense). By focusing only on points in the process with large enough values, we may capture the variation in the underlying process without having to sample periodically. In essence, Lebesgue sampling allows us to focus on the most information-rich points of the process, those points that tell us the most about the movements of the continuous process. Lebesgue sampling has been compared to Riemann/periodic sampling and has been found, under certain conditions, to reduce the necessary sampling rate required to maintain the same mean error variance as Riemann sampling (Astrom and Bernhardsson, 2002).

In its simplest form, optimal sampling theory seeks to utilize those points in a given process that provide the most information about that which is most important to the researcher. In this way, the theory seeks to isolate certain points of the process, creating a powerful subset of data that can then be used to model the phenomenon in question in a more targeted fashion. In the current research, this means, for instance, isolating those points where linear-type trend is important from those points where cycle-type trend is important. In other words, we believe that there are points that lean themselves toward the prediction of trend, X_t, while other points lean more toward the prediction of cycles in the process. Once such points have been isolated via some algorithm, one is then able to use these points in a way that is computationally more efficient or easier to understand than some more complex procedure being used on the overall process. This perspective, of points leaning to linear versus cyclical trend, is how we approach the current problem. Let us now wrap up this chapter with some remarks before venturing into the next chapter.

1.6 Chapter 1 Remarks

The concepts outlined in Chapter 1 largely represent what should be learned by a graduate student in pursuit of a degree in mathematical statistics. However, some rather general references are in line here. For a very thorough treatment of stochastic processes consult Gikhman and Skorokhod (1969). For less complicated yet decently rigorous expositions, see Billingsley (2008) or Ross (1996), which are commonly-used texts for graduate courses in probability theory. For a more detail explanation of most of the concepts of this first chapter, please consult Satō (1999). Any references more specific than these were included in the chapter where relevant. Now we move on to the more applied segment of this document, namely Chapter 2, which concerns the presentation of more specific statistical results for the goals of the current project. It is in this upcoming chapter that we see more specific results en route to our ultimate goal.

CHAPTER 2:
APPLIED STATISTICAL BACKGROUND

2.1 Stochastic Processes

A great deal of information about stochastic processes can be found in Chapter 1, where relevant preliminaries were outlined. Here, in Chapter 2, we present additional information about such processes, information more fundamentally related to the problem at hand. For a more thorough description of some of the topics of this chapter, please see Teodorescu (2013). Let us start by recalling what it means for a stochastic process to be *wide-sense stationary*, which we will merely refer to as *stationary* where there is no confusion.

We consider a stochastic process *stationary* if it has the following properties (where $m(t)$ and $K(s,t) = V[X_s, X_t]$ are the mean and covariance function of some generally stated stochastic process, X_t, respectively):

1. Constant Mean Condition:

   $$m(t) = m(t+h) \text{ for all } t, h \in \mathbb{R}$$

2. Time Difference Condition for Covariance:

   $$K(s,t) = K(0, t-s) \coloneqq K(t-s) \text{ for all } s, t \in \mathbb{R}$$

3. Finite Absolute Second Moment Condition:

   $$E[|X_t|^2] < \infty$$

For stationary (i.e., wide-sense stationary) stochastic processes, the Bochner-Lesbegue theorem implies that there exists a measure on $\mathbb{R}$, denoted below by μ and often referred to as the *spectral measure*, such that

$$K(t) = \int e^{-2\pi i t \nu} d\mu(\nu) \tag{155}$$

This result implies that the (auto-)covariance function a (zero-mean) stationary stochastic process is the Fourier transform of well-behaved densities. Generally, this result allows us to use the standard and well-understood tools of time-series analysis, such as *filtering, prediction and estimation* (see Section 2.2 below). This is about as good as one can hope to do when dealing with stochastic processes. Recall, however, that we are concerned with the prediction and estimation of *non-stationary* stochastic processes. As such, the question now becomes: how far away from stationarity can one go and still apply the standard tools? Along these lines, let us now explore a class of non-stationary processes, namely *harmonizable* stochastic processes.

2.1.1 Strongly Harmonizable Processes

As outlined in Loeve (1965) and detailed in Hurd (1989), a non-stationary stochastic process, X_t, is strongly harmonizable if there exists a measure μ on $\mathbb{R}^2$, such that

$$K(s,t) = \int_{\mathbb{R}^2} e^{i(s\nu - t\eta)} \mu(\nu, \eta) d\nu d\eta \tag{156}$$

which is to say that, like in the case of stationary processes, there exists a kind of Fourier transform representation for the covariance function of the process. Unlike the stationary case, however, the relationship for strongly harmonizable processes involves a measure on $\mathbb{R}^2$ as opposed to $\mathbb{R}$. This need for more free parameters to describe the spectral nature of the process has a lot to do with the fact that stationary processes have covariance functions which have only one input, namely the time difference $t - s$, as opposed to the covariance function having two inputs, t and s, as is the case for strongly harmonizable processes. This highlights the need for more involved spectral formulations to describe the nature of non-stationary processes, even those only slightly non-stationary, like is the current case. We turn our attention next to an inequality for stongly harmonizable processes, which is based on work done by Martin and Putinar (1989). This inequality is here to provide the interested reader with another perspective on the nature of strongly harmonizable processes.

Theorem 2.3. X_t is strongly harmonizable if there is a positive constant C such that:

$$\left| \int_{\mathbb{R}^2} f(s,t)K(s,t)\mathrm{d}s\mathrm{d}t \right| \leq C||f||_\infty \tag{157}$$

for any finite-norm function f.

Alternatively, we have the following result from Hurd (1989).

Theorem 2.4. X_t is strongly harmonizable if it is the Fourier transform of another stochastic process on the reciprocal space:

$$X_t = \int_{\mathbb{R}} e^{\mathrm{i}t\nu} Z(\nu)\mathrm{d}\nu, \tag{158}$$

where Z satisfies $E(Z(A) \cdot Z(B)) = \mu(A \times B)$, for any $A, B \subset \mathbb{R}_+$. If Z satisfies $Z(A) \perp Z(B)$ if $A \cap B = \emptyset$, then X_t is actually stationary. Let us now turn to a result which gives us hope for the analysis of strongly harmonizable processes.

Theorem 2.5. Any strongly harmonizable process is asymptotically stationary. In other words, there is a smooth limit such that:

$$K(t) = \lim_{T \to \infty} \left[\frac{1}{2T} \int_{-T}^{T} K(s+t, s)\mathrm{d}s \right] \quad (\forall) t \in \mathbb{R}_+ \tag{159}$$

This result certainly provides one with hope for analysis of strongly harmonizable processes, as the tools for estimation, prediction, and filtering of asymptotically stationary processes are the same as for stationary ones (with minimal supplementary conditions placed on the stochastic model).

Now that we have justified the existence of protocols (i.e., statistical techniques) for the treatment of certain non-stationary time series models, namely strongly harmonizable stochastic processes, let us now turn our attention to yet another form of non-stationary process, the weakly harmonizable process. For more information about weakly harmonizable processes, see (once again) Hurd (1989). For an even more detailed treatment of the concept than what is found in Hurd (1989), see Niemi (1975) or Chang and Rao (1987).

2.1.2 Weakly Harmonizable Processes

Definition 2.4. A function f is a *bi-measure* if it is the Fourier transform of any bounded function on $\mathbb{R}^2$. Then, X_t is *weakly harmonizable* if it is a bi-measure.

Theorem 2.6. The following two statements are equivalent:

- X_t is the Fourier transform of an arbitrary stochastic process on the reciprocal space
- X_t is the projection of a stationary process Y_t from higher-dimensional space

From Definition 2.1 and Theorem 2.4, we may infer that the properties needed for a stationary process analysis (i.e., the properties associated with the process being either stationary or asymptotically stationary) are lost during the projection from a higher-dimensional space. Seeing as the projection of a stationary stochastic process to a lower number of dimensions generically leads to a weakly harmonizable process, we do not have the same analytical luxuries afforded to us in the strongly harmonizable case, where the standard tools could be used in an asymptotic sense. As such, we have established that the existing tools do not address all non-stationary processes equally well, especially not processes as general as our current concern, a process with both linear- and cycle-type trend in addition to Levy process noise. Let us now explore some results related to the estimation, prediction, and filtering of an asymptotically stationary process. In many ways the following section is a continuation of the strongly harmonizable section above.

2.2 Applications: Estimation, Prediction and Filtering

Here, we briefly list some applications to estimation, prediction and filtering for asymptotically stationary processes. Recall from Theorem 2.3 that any strongly harmonizable process is also asymptotically stationary. The following results are presented without proof as they are important to but largely incidental in the overall goals of the current project. It is worth noting, however, that proof of these results is largely computational.

Theorem 2.7. Assume that X_t is asymptotically stationary and that the following two conditions are true (i.e., the "minimal supplementary conditions" mentioned previously):

$$\lim_{T\to\infty} \sup_{[-T,T]} \left| \frac{1}{2T} \int_{-T}^{T} K(s+t,s)\mathrm{d}s \right| = 0 \tag{160}$$

$$\frac{1}{2T} \int_{-T}^{T} [E(||X_t||^4)]^{1/2}\mathrm{d}t \leq M \quad (\forall) T \in R_+ \tag{161}$$

then

$$\mu_{n,N}(\nu) := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{4n^2} \int_{-n}^{n}\int_{-n}^{n} e^{-\mathrm{i}t\nu} E[X_i(s+t)X_i(s)]\mathrm{d}s\mathrm{d}t \tag{162}$$

is a consistent estimator of μ. By consistency, we mean that this estimator converges in probability to the true value of the parameter being estimated.

Theorem 2.8. The least-squares predictor of any asymptotically stationary process with limit auto-correlation K (such as that outlined in Theorem 2.3) is as good as the least-squares predictor of a stationary process with auto-correlation K.

Theorem 2.9. Adding any stationary noise term to an asymptotically stationary process with vanishing auto-correlation in the infinte-time limit and bounded fourth moment does not affect the estimators or predictors obtained based on the asymptotically stationary process without the additional noise.

Based on the results of the current section, we see that the tools available for stationary processes are largely available for asymptotically stationary processes, as was mentioned previously. Let us now move on to some of the limitations experienced so far.

2.3 Decomposition of Stochastic Processes

What can be done to model non-stationary processes when none of the conditions discussed above describe the system well enough? In other words, how can we generalize the currently available methods to handle more complicated (i.e., more non-stationary) trends within the data? Rigorously speaking, only a direct integration of the stochastic equations is justified in this case. For specific (linearizable) models (such as trend-cycle),

an exact time-dependent solution for the reduced variables may be obtained. However, these cases are *all* restricted to linear first-order equations, so that the case discussed above, that involving a stochastic process with both linear- and cycle-type trends and Levy process noise, cannot be treated with the approximation strategies outlined earlier in this chapter. Let us now focus our attention on a broader class of non-stationary stochastic process models, namely *Dynamic Linear Models*.

2.4 Dynamic Linear Models

Let us begin this section by discretizing the time step and transforming all linear differential equations (in time) into difference equations with unit time step. Then a general dynamical linear model (in the manner covered in West and Harrison, 1997) is given by

$$\boldsymbol{Y}_t = \boldsymbol{F}_t\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \boldsymbol{V}_t) \tag{163}$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t,\ \boldsymbol{\omega}_t \sim N(\boldsymbol{0}, \boldsymbol{W}_t) \tag{164}$$

where $\boldsymbol{Y}_t$ represents the observed outputs to be modeled, $\boldsymbol{F}_t$ plays the role of the design matrix (see the Linear Regression section of Chapter 1), $\boldsymbol{\theta}_t$ represents the regression parameter vector at time t, $\boldsymbol{\nu}_t$ is the observational noise vector (with known covariance matrix $\boldsymbol{V}_t$), $\boldsymbol{G}_t$ is a matrix representing the evolution of the regression parameter vector in time, and $\boldsymbol{\omega}_t$ is the noise vector associated with the evolution of the regression parameter vector. The first equation is generally taken to be the *observation equation* for the modeling problem, while the second equation is taken to be the *evolution equation* for the same problem. The noise terms are taken to be uncorrelated, unbiased, possibly time-dependent Gaussians. Suppose, lastly, that

$$\boldsymbol{\theta}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \tag{165}$$

Under these conditions, the time-dependent solution of the DLM system outlined above is given via the following *updating equations*, where D_t represents all data observed

up to time t, such that $D_t := \{\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t\}$:

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim N(\boldsymbol{a}_t, \boldsymbol{R}_t), \quad \boldsymbol{a}_t = \boldsymbol{G}_t\boldsymbol{\mu}_{t-1}, \ \boldsymbol{R}_t = \boldsymbol{G}_t\boldsymbol{\Sigma}_{t-1}\boldsymbol{G}_t' + \boldsymbol{W}_t \tag{166}$$

$$(\boldsymbol{Y}_t|D_{t-1}) \sim N(\boldsymbol{\phi}_t, \boldsymbol{Q}_t), \quad \boldsymbol{\phi}_t = \boldsymbol{F}_t\boldsymbol{a}_t, \quad \boldsymbol{Q}_t = \boldsymbol{F}_t\boldsymbol{R}_t\boldsymbol{F}_t' + \boldsymbol{V}_t \tag{167}$$

$$(\boldsymbol{\theta}_t|D_t) \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad \boldsymbol{\mu}_t = \boldsymbol{a}_t + \boldsymbol{A}_t\boldsymbol{e}_t, \ \boldsymbol{\Sigma}_t = \boldsymbol{R}_t - \boldsymbol{A}_t\boldsymbol{Q}_t\boldsymbol{A}_t' \tag{168}$$

where $\boldsymbol{A}_t = \boldsymbol{R}_t\boldsymbol{F}_t'\boldsymbol{Q}_t^{-1}$, $\boldsymbol{e}_t = \boldsymbol{Y}_t - \boldsymbol{\phi}_t$, $(\boldsymbol{\theta}_t|D_{t-1})$ represents the prior for $\boldsymbol{\theta}_t$, $(\boldsymbol{Y}_t|D_{t-1})$ represents the single-step forecast for $\boldsymbol{Y}_t$, and $(\boldsymbol{\theta}_t|D_t)$ represents the posterior for $\boldsymbol{\theta}_t$.

These updating equations assume complete knowledge of the parameters of the noise vectors, which here means knowledge of the structure of the covariance matrices for $\boldsymbol{\nu}_t$ and $\boldsymbol{\omega}_t$. Yet, we can almost never rely on the data to give us this information directly. As such, we must now turn our attention to updating equations not conditioned on $\boldsymbol{V}_t$:

$$(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim T_{n(t-1)}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}) \tag{169}$$

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim T_{n(t-1)}(\boldsymbol{a}_t, \boldsymbol{R}_t) \tag{170}$$

$$(\boldsymbol{Y}_t|D_{t-1}) \sim T_{n(t-1)}(\boldsymbol{\phi}_t, \boldsymbol{Q}_t) \tag{171}$$

$$(\boldsymbol{\theta}_t|D_t) \sim T_{n(t)}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \tag{172}$$

where $n(t) = t$ (to avoid confusion). In these equations, T_n stands for the T distribution with n degrees of freedom. Now, we turn to the forecasting and prediction of DLM models as outlined above.

2.4.1 Forecasting and Prediction of Dynamic Linear Models

Let us introduce the forecast function, which represents what is expected of the observable process, $\boldsymbol{Y}_t$, k steps ahead of what has been observed (i.e., D_t):

$$\boldsymbol{f}_t(k) = E[\boldsymbol{Y}_{t+k}|D_t], \quad k \geq 1 \tag{173}$$

In light of this forecast function, we have the following:

$$(\boldsymbol{\theta}_{t+k}|D_t) \sim T_{n(t)}(\boldsymbol{a}_t(k), \boldsymbol{R}_t(k)) \tag{174}$$

$$(\boldsymbol{Y}_{t+k}|D_t) \sim T_{n(t)}(\boldsymbol{f}_t(k), \boldsymbol{Q}_t(k)), \tag{175}$$

where

$$\boldsymbol{f}_t(k) = \boldsymbol{F}_{t+k}\boldsymbol{a}_t(k) \tag{176}$$

$$\boldsymbol{Q}_t(k) = \boldsymbol{F}_{t+k}\boldsymbol{R}_t(k)\boldsymbol{F}'_{t+k} + \boldsymbol{V}_{t+k} \tag{177}$$

$$\boldsymbol{a}_t(k) = \boldsymbol{G}_{t+k}\boldsymbol{a}_t(k-1) \tag{178}$$

and

$$\boldsymbol{R}_t(k) = \boldsymbol{G}_{t+k}\boldsymbol{R}_t(k-1)\boldsymbol{G}'_{t+k} + \boldsymbol{W}_{t+k} \tag{179}$$

Note that covariance matrices may also be forecasted in a similar manner. Now that we have some background, definitions, and forecast equations for dynamic linear models, which are (once again) a type of non-stationary stochastic process model, let us move on to the concept of Bayesian estimation, which while very general itself, will prove quite useful in our current aims and goals.

2.5 Bayesian Estimation

Let us begin this section with some background on Bayesian estimation, first generally and then in the context of non-stationary stochastic processes. Bayesian estimation is to be contrasted with what is referred to as *Fisher Estimation*, in the sense that the latter is data-driven and wholly dependent on availability of large data sets, while the former isn't so much. Fisher inference makes use of the Central Limit Theorem and its variants (Maximum Likelihood, Delta method, etc) to provide either point estimates (MLE, UMVUE) or interval estimates (confidence intervals), whose statistical relevance and effectiveness generally increase with increases in sample size. Seen as both a strength and weakness, Bayesian estimation has a greater reliance on subjective information through its use of

distributions chosen by the researcher (and not based on the data) which may or may not yield drastically different estimates in a given statistical analysis context.

In the Bayesian paradigm we have a data vector $\boldsymbol{Y}$ with density p_θ for some unknown yet variable parameter $\theta \in \Theta$. One of the initial goals of a Bayesian estimation problem is to put a prior density on θ. The family of available prior densities can be denoted as $\{\nu_h, h \in \mathcal{H}\}$, where h is called a hyperparameter, which is usually seen as a fixed parameter for the distribution of the variable parameter, θ. Typically, the hyperparameter is multivariate and choosing it can be difficult. However, this choice is also very important and can have a rather large impact on subsequent inferences. As such, there are two issues to consider, the first dealing with sensitivity analysis and the second with model selection:

- Suppose that we fix a quantity of interest, say, $f(\theta)$, where f is some function. How, then, may we assess changes in the posterior expectation of $f(\theta)$ as we make changes to h? In other words, how sensitive are the results to changes in the hyperparameter, h?

- How do we determine if a given subset of $\mathcal{H}$ constitutes a class of reasonable choices?

Once again, when comparing the two approaches, Bayesian versus Fisher estimation, we must focus our attention on the relative importance of sample size. While the Fisherian analyst can almost always say, "the more data, the better the estimates," the Bayesian analyst does not see sample size with the same importance. It is perfectly possible to obtain a very good, precise, accurate Bayesian estimate from a small sample, provided that the prior distribution was properly chosen. Conversely, with a bad choice of prior, estimation based on a large sample will give bad Bayesian estimates, which is not the case in Fisher inference. In some ways, we can see this difference as representing a shift in importance from the rigid, objective nature of sample size to the more open, more subjective nature of prior distribution choice. Next, we turn our attention to a comparison of the ordinary and empirical Bayesian approaches, which may be seen as more parametric and non-parametric Bayesian approaches, respectively.

2.5.1 Ordinary Versus Empirical Bayesian Estimation

In ordinary Bayesian estimation, it is known (or assumed) that the distribution we aim to determine is parametric, meaning that it belongs to a class of explicit functions characterized by parameters, such as the Gamma, Normal, Uniform, etc. In empirical Bayes, such an assumption is not needed, as we simply aim to determine the empirical c.d.f. (or some other equivalent probabilistic description) for that population, based on the sample available, the data of the problem.

Consider the problem of variable selection in Bayesian linear regression. Here, we have a response variable Y and a set of predictors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$, each a vector of length m. For every (predictor index) subset $\gamma \subseteq \{1, \ldots, q\}$ we have a potential model $\mathcal{M}_\gamma$ given by

$$Y = 1_m \beta_0 + \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\varepsilon}, \tag{180}$$

where 1_m is the vector containing m 1's, $\boldsymbol{X}_\gamma$ is the design matrix whose columns consist of the predictor vectors corresponding to the subset γ, $\boldsymbol{\beta}_\gamma$ is the vector of coefficients for that subset, and $\boldsymbol{\varepsilon} \sim N_m(0, \sigma^2 \boldsymbol{I})$. Let q_γ denote the number of variables in the subset γ. The unknown parameter vector for this model is $\boldsymbol{\theta} = (\gamma, \sigma, \beta_0, \boldsymbol{\beta}_\gamma)$, which includes the predictor index subset for the subset of predictors that go into the linear model, γ. A very commonly used prior distribution on $\boldsymbol{\theta}$ is given by the hierarchical process in which we first choose the indicator γ using an *Independent Bernoulli Prior*, where each predictor goes into the model with a certain probability, say w, independently of all the other variables. We then determine the vector of regression coefficients corresponding to the selected predictors. Under these conditions, we have the following model:

$$\begin{aligned}
Y &\sim N_m(1_m \beta_0 + \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}) && (181)\\
(\sigma^2, \beta_0) &\sim p(\sigma^2, \beta_0) \propto 1/\sigma^2 && (182)\\
(\sigma, \boldsymbol{\beta}_\gamma) &\sim N_{q_\gamma}(0, g\sigma^2 (\boldsymbol{X}_\gamma' \boldsymbol{X}_\gamma)^{-1}) && (183)\\
\gamma &\sim w^{q_\gamma} (1-w)^{q - q_\gamma} && (184)
\end{aligned}$$

where the third distribution (183) is Zellner's g-prior, which was introduced by Zellner in 1986 and is indexed by the parameter g. Although this prior is improper, meaning that it is not a proper probability density (i.e., does not have a total measure of 1), its use does result in a posterior distribution which is proper. Let us now continue the description of this model.

The prior on the parameter $\boldsymbol{\theta} = (\gamma, \sigma, \beta_0, \boldsymbol{\beta}_\gamma)$ is given by the two-level hierarchy (184) and (182, 183), and is indexed by $h = (w, g)$. Loosely speaking, when w is large and g is small, the prior encourages models with many predictors and small coefficients, whereas when w is small and g is large, the prior concentrates its mass on parsimonious models with large coefficients. Therefore, the hyperparameter $h = (w, g)$ plays a very important role, and in effect determines the model that will be used to carry out variable selection.

A standard method for approaching model selection involves the use of what are known as *Bayes Factors*. For each $h \in \mathcal{H}$, let $m_h(y)$ denote the marginal likelihood of the data under the prior ν_h, that is, $m_h(y) = \int p_\theta(y)\nu_h(\theta)\, d\theta$. We will write m_h instead of $m_h(y)$ where there is no confusion in doing so. The Bayes factor of the model indexed by h_2, compared to the model indexed by h_1, is defined as the ratio of the marginal likelihoods of the data under the two models, m_{h_2}/m_{h_1}, and is denoted throughout by $B(h_2, h_1)$. Bayes factors are widely used as a criterion for comparing models in Bayesian analyses. In terms of selecting the best models from the family of models indexed by $h \in \mathcal{H}$, the strategy is usually to compute and subsequently compare all the Bayes factors $B(h, h_1)$, where $h \in \mathcal{H}$ and h_1 is a fixed hyperparameter value. We could then consider as good candidate models those with values of h that result in the largest Bayes factors.

Suppose now that we fix a particular function f of the parameter θ. For instance, we may choose to fix the indicator set so that predictor 1 is included in the regression model. It is of general interest to determine the posterior expectation $E_h(f(\theta) \mid Y)$ as a function of h and to determine whether or not $E_h(f(\theta) \mid Y)$ is very sensitive to the value of h. If it is not, then two individuals using two different hyperparameters should reach approximately the same conclusions and the results of the analysis should be less controversial in general. On the other hand, if for a function of interest the posterior

expectation varies considerably as we change the hyperparameter, then we will want to know which aspects of the hyperparameter (e.g., which components of h) produce the biggest changes. As such, we may wish to see a plot of the posterior expectations as we vary those aspects of the hyperparameter. Except for extremely simple cases, posterior expectations cannot be obtained in closed form, and are typically estimated via the Markov-Chain Monte Carlo (MCMC) method. Seeing as it is slow and inefficient to run Markov chains for every hyperparameter value h, we should not place too much hope in obtaining a full numerical perspective or picture for all hyperparameter values. Before moving on to a brief section comparing the Fisherian and Bayesian approaches to large sample theory, some general references are in line here.

For good expositions of the concepts of Bayesian analysis in general see *Bayesian Data Analysis* by Gelman et al. (2004) or *Bayesian and Frequentist Regression Methods* by Wakefield (2013). The former is often used as a text for graduate courses in mathematical statistics, while the latter does a good job of comparing the Fisherian/Frequentist approaches to the Bayesian approaches.

2.5.2 Bayesian Approach in Large Sample Theory

The main conceptual advantage of performing Bayesian inference even when data are plentiful rests in the ability of the researcher to guess distribution functions whose analytical structure is significantly different from the unimodal, mean-dominated kind of inference afforded by the Fisherian approach. In other words, it is possible to obtain a perfectly good Bayesian estimate with low convergence properties such as a Cauchy distribution, which would be quite challenging for Fisher-based inference, even with large samples.

A second important distinction is that the Bayes approach naturally allows us to perform hypothesis testing in the framework of decision (game) theory, where the loss functions, risk functions, etc. are far more general than those used in (classical) Fisherian hypothesis testing methods (such as the UMP or LRT methods). Now that we have some familiarity with model selection, let us turn next to sensitivity analysis.

2.5.3 Sensitivity Analysis in Bayesian Inference

As mentioned previously, sensitivity analysis and model selection are issues often at the core of Bayesian inference problems. One possible approach involves running Markov chains corresponding to a few values of the hyperparameter, say, $h_1, \ldots, h_k$, and using these to estimate $E_h(f(\theta) \mid Y)$ and the Bayes factors $B(h, h_i)$ for all $h \in \mathcal{H}$. The difficulty we face is that there is a severe computational burden caused by the requirement that we handle a very large number of values of h. Another approach for estimating large families of posterior expectations and Bayes factors is based on a combination of MCMC, importance sampling, and the use of control variates.

The idea of using importance sampling to investigate data streams from multiple densities has been studied repeatedly, as we see throughout the remainder of the current section.

Suppose that we have a sample $\theta_1, \ldots, \theta_n$ (of iid or ergodic Markov chain output) from the posterior density $\nu_{h_1,y}$ for a fixed h_1 and we are interested in the posterior expectation

$$E_h\big(f(\theta) \mid Y = y\big) = \int f(\theta) \frac{\nu_{h,y}(\theta)}{\nu_{h_1,y}(\theta)} \nu_{h_1,y}(\theta)\, d\theta \tag{185}$$

for different values of h. Using the fact that

$$\int \frac{p_\theta(y)\nu_h(\theta)/m_h}{p_\theta(y)\nu_{h_1}(\theta)/m_{h_1}} \nu_{h_1,y}(\theta)\, d\theta = 1 \tag{186}$$

we see that this expectation (185) may be written as

$$\int f(\theta) \frac{p_\theta(y)\nu_h(\theta)/m_h}{p_\theta(y)\nu_{h_1}(\theta)/m_{h_1}} \nu_{h_1,y}(\theta)\, d\theta = \frac{\int f(\theta)(\nu_h(\theta)/\nu_{h_1}(\theta))\nu_{h_1,y}(\theta)\, d\theta}{\int (\nu_h(\theta)/\nu_{h_1}(\theta))\nu_{h_1,y}(\theta)\, d\theta} \tag{187}$$

where the right-hand side of (187) does not involve the ratio m_h/m_{h_1} (i.e., the Bayes factor comparing h to h_1). The idea to express $\int f(\theta)\nu_{h,y}(\theta)\, d\theta$ in this way was proposed in a different context by Hastings, in 1970. The right-hand side of (187) is the ratio of two integrals with respect to $\nu_{h_1,y}$, each of which may be estimated from the sequence

$\theta_1, \ldots, \theta_n$. We may estimate the numerator and the denominator by

$$\frac{1}{n}\sum_{i=1}^{n} f(\theta_i)[\nu_h(\theta_i)/\nu_{h_1}(\theta_i)] \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}[\nu_h(\theta_i)/\nu_{h_1}(\theta_i)] \tag{188}$$

respectively, and $\int f(\theta)\nu_{h,y}(\theta)\,d\theta$ is estimated by the ratio of these two quantities.

The disappearance of the likelihood function on the right-hand side of (187) is very convenient because its computation requires considerable effort in some cases (e.g., when we have missing or censored data, the likelihood is a possibly high-dimensional integral). Note that the second average in (188) is an estimate of m_h/m_{h_1}, that is, the Bayes factor $B(h, h_1)$. Ideally, we would like to use the estimates in (188) for multiple values of h using only a sample from the posterior distribution corresponding to the fixed hyperparameter value h_1. But, when the prior ν_h differs from ν_{h_1} greatly, the two estimates in (188) are unstable because of the potential that only a few observations will dominate the sums. Their ratio suffers the same defect.

A natural approach for dealing with the instability of these simple estimates is to choose k values $h_1, \ldots, h_k \in \mathcal{H}$ and in (185) replace $\nu_{h_1,y}$ with a mixture $\sum_{s=1}^{k} a_s\nu_{h_s,y}$, where $a_s \geq 0$, for $s = 1, \ldots, k$, and $\sum_{s=1}^{k} a_s = 1$. For concreteness, consider the estimate of the Bayes factor. Let $\overline{\nu}_{\cdot y} = \sum_{s=1}^{k} a_s\nu_{h_s,y}$, and let $d_s = m_{h_s}/m_{h_1}, s = 1, \ldots, k$. Note that if $\nu_h(\theta) = 0$ whenever $\nu_{h_s}(\theta) = 0$ for all s, then we have the following

$$B(h, h_1) = \int \frac{\nu_h(\theta)}{\sum_{s=1}^{k} a_s\nu_{h_s}(\theta)/d_s}\overline{\nu}_{\cdot y}(\theta)\,d\theta \tag{189}$$

and

$$\begin{aligned}\int f(\theta)\nu_{h,y}(\theta)\,d\theta &= (B(h, h_1))^{-1}\int f(\theta)\frac{\nu_h(\theta)}{\sum_{s=1}^{k} a_s\nu_{h_s}(\theta)/d_s}\overline{\nu}_{\cdot y}(\theta)\,d\theta &(190)\\ &= \frac{\int f(\theta)(\nu_h(\theta)/\sum_{s=1}^{k} a_s\nu_{h_s}(\theta)/d_s)\overline{\nu}_{\cdot y}(\theta)\,d\theta}{\int(\nu_h(\theta)/\sum_{s=1}^{k} a_s\nu_{h_s}(\theta)/d_s)\overline{\nu}_{\cdot y}(\theta)\,d\theta} &(191)\end{aligned}$$

Suppose that, for each $l = 1, \ldots, k$, we have Markov chain samples $\theta_i^{(l)}, i = 1, \ldots, n_l$, from the posterior density $\nu_{h_l,y}$. Letting $n = \sum_{s=1}^{k} n_s$: if $a_s = n_s/n$, then the pooled

sample is a stratified sample from $\overline{\nu}_{\cdot y}$. In the case where the vector $d = (d_2, \ldots, d_k)'$ is known, the right-hand side of (189) is the integral of a known function with respect to the mixture density $\overline{\nu}_{\cdot y}$. Then, under certain regularity conditions, the estimate of $B(h, h_1)$ obtained by replacing the right-hand side of (189) by its natural Monte Carlo estimate using the pooled sample is consistent and asymptotically normal.

In virtually all applications, the value of the vector d is unknown. The estimates of $B(h, h_1)$ and $\int f(\theta)\nu_{h,y}(\theta)\, d\theta$ usually considered in this case are constructed by first forming an estimate $\hat{d}$ of d, and then using the natural Monte Carlo estimates of the integrals in (189, 190, 191) with $\hat{d}$ substituted for d.

While it may feel as though we have ventured off a bit from the goals of this project, this is not the case, as we will utilize the concepts of the current chapter to bolster our understanding of the contents of Chapter 3. So far in this section, we have looked at many results and concepts somewhat related (be they directly or tangentially related) to the modeling of non-stationary stochastic process with both linear- and cycle-type trend in the presence of Levy process noise. While the discussion has grown somewhat armchair, we have not lost sight of the goals of the project. In the next section, we will learn more about *Nonparametric Inference*, which consists of a class of procedures that do not make the same distributional or parametric assumptions as *Parametric Inference*.

2.6 Nonparametric Inference

As indicated previously, parametric analysis assumes that the distributions of interest belong to certain classes, and therefore the data, models, etc. can be fitted according to some exact analytical expressions. By contrast, nonparametric analysis makes no such assumptions and works exclusively with the empirical cumulative distribution function and its derived quantities. Let us now explore some of the basic concepts of nonparametric inference.

Let $\boldsymbol{X}$ be a random vector with distribution function F and let $\mathbf{x} = (x_1, \ldots, x_n)'$ be an observed sample from F. Suppose $R(\mathbf{x}, F)$ is a statistical quantity that depends in general on both the unknown distribution F and on the sample $\mathbf{x}$. For example,

$R(\mathbf{x}, F)$ could be an estimator of an unknown parameter. If F is unknown, then the exact distribution of the random variable $R(\mathbf{x}, F)$ is generally unknown.

A well-known method to nonparametrically estimate the distribution of $R(\mathbf{x}, F)$ consists of the following steps:

(i) From the observed sample $\mathbf{x}$, use the empirical distribution function, $\widehat{F}_n$, as an estimate of the probability function F. The empirical distribution function is defined by $\widehat{F}_n(x) = \frac{n(x)}{n}$, where $n(x)$ is the number of values x_i in $\mathbf{x}$ that are less than or equal to x.

(ii) Draw B samples of size n from $\widehat{F}_n$ conditional on $\mathbf{x}$. Denote these as $\mathbf{x}_j^*$, for $j = 1, \ldots, B$.

(iii) For each sample $\mathbf{x}_j^*$, compute $R_j^* = R(\mathbf{x}_j^*, \widehat{F}_n)$ and approximate the distribution of $R(\mathbf{x}, F)$ with the empirical distribution of $R_1^*, \ldots, R_B^*$.

The empirical distribution function can also be computed, based on the sample available. Denote this function by $\widehat{F}$. A $(1-\alpha)100\%$ confidence interval based on the percentile method of Efron (1979) is given by $[\widehat{F}^{-1}(\alpha), \widehat{F}^{-1}(1-\alpha)]$. Here, $x_L = [\widehat{F}]^{-1}(\alpha)$ is the largest value of x such that the number of elements in the sample that are less than x is smaller than αn. Likewise, $x_U = [\widehat{F}]^{-1}(1-\alpha)$ is the smallest value of x such that the number of elements in the sample that are smaller than x is larger than $(1-\alpha)n$. Specifically,

$$x_L = \max\left\{x : \widehat{F}_n(x) \leq \alpha\right\} \tag{192}$$

$$x_U = \min\left\{x : \widehat{F}_n(x) \geq 1-\alpha\right\} \tag{193}$$

2.6.1 Nonparametric Kernel Density Approach

Assume that if $X_1, \ldots, X_n$ are iid random variables having a common probability density function $f(x)$. Then the kernel estimate of $f(x)$ is defined by

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{194}$$

where h is the bandwidth and $K(u)$ is the kernel function. The kernel estimate of the cumulative distribution function $\widehat{F}(x)$ and reliability function $R(x)$ are, respectively, given by

$$\hat{F}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{x} K\left(\frac{y - X_i}{h}\right) dy \tag{195}$$

and

$$\hat{R}_n(x) = 1 - \hat{F}_n(x) \tag{196}$$

It is usually assumed that $K(y)$ is a symmetric function, which can be taken to be normalized to 1, centered (zero first moment), and positive-definite (positive second centered moment). Using a kernel approach has the advantage that it is often possible to optimize the analysis and obtain reliable results relatively quickly. However, the kernel method ignores any interactions between the data, missing entirely any unwanted correlations or higher-order effects. Therefore, it is better to use a hierarchical approach, where we first use the usual kernel approach and then use a function that has dependence on both x and predictor pairs, such as X_i, X_j, etc. Note that here K represents the kernel function and not the covariance function outlined previously.

Properties of the kernel function $K(u)$ partially determine the properties of the kernel density estimates, such as differentiability and continuity. For example, if $K(u)$ is a proper density function, that is if it is non-negative and it integrates to one, then the kernel density estimate is also a proper density function. If $K(u)$ is n times differentiable, so is $\hat{f}_n(x)$. Early works on kernel density estimation include Rosenblatt (1956), Hodges and Lehmann (1965), and Epanechnikov (1969). In their work, Hodges and Lehmann showed that the Epanechnikov kernel (which was not yet fully defined at the time) optimizes the expression used in finding the optimal bandwidth, thus making it the most efficient kernel.

When evaluating subject kernels by comparing them with the Epanechnikov (i.e., parabolic) kernel, the optimal bandwidth is given by

$$h_o = \left[\frac{||K||^2}{nM_2R(f'')}\right]^{1/5} \tag{197}$$

where $||K||^2 = \int K^2(t)dt$ and $M_2 = \int t^2 K(t)dt$. The optimal bandwidth value is determined by minimizing the mean square error (MSE) for the estimate,

$$MSE(\hat{f}) = \mathbb{E}(\hat{f} - f)^2 = \text{Bias}(\hat{f})^2 + \text{Var}(\hat{f}) \tag{198}$$

If $h \to 0$, $nh \to \infty$, and the underlying density f is a sufficiently smooth L^2 function, then it can be shown that Bias $\to h^2 M_2 f''(x)/2$ and Var $\to f(x)||K||^2/(nh)$.

Thus, we can infer that if the bandwidth decreases, the bias of the kernel estimate also decreases but the variance increases, resulting in a rough and unacceptable estimate of the kernel density. Conversely, if the bandwidth increases, the variance of the kernel estimate decreases but the bias increases. This means that there is significant smoothing and the underlying characteristics of the probability density are smoothed out. Combining these results and integrating over the entire real line gives us an estimate of the global accuracy of $\hat{f}(x)$, the asymptotic significant mean integrated square error (AMISE):

$$\text{AMISE}(f) = \frac{h^4 M_2^2 R(f'')}{4} + \frac{||K||^2}{nh} \tag{199}$$

Thus, we can conclude that AMISE depends on four quantities: the bandwidth h, the sample size n, the kernel function K, and the target density $f(x)$. The target function and the sample size are largely out of our control. However, we can minimize AMISE by choosing the appropriate kernel and the bandwidth. If we fix the kernel function $K(u)$ and minimize AMISE with respect to the bandwidth we obtain the following optimal forms:

$$h_o = \left[\frac{||K||^2}{nM_2 R(f'')}\right]^{1/5} \tag{200}$$

and

$$\text{AMISE}_o = \frac{5}{4}\left[\frac{\sqrt{M_2}||K||^2}{n}\right]^{4/5} (C(f''))^{1/5} \tag{201}$$

To calculate the optimal kernel function, we minimize AMISE_o with respect to K. The optimal kernel function was derived by Epanecnikov in 1969 (in the same work cited

above) and is given by

$$K(u) = \frac{3}{4}(1 - u^2)\chi_{|u|\leq 1} \tag{202}$$

The value of $\sqrt{M_2}||K||^2$ for the Epanechnikov kernel is $3/(5\sqrt{5})$, so that the ratio $\frac{\sqrt{125M_2}}{3}||K||^2$ provides a measure of inefficiency for other kernels. We then have a measure of the relative effectiveness of other kernels relative to the Epanechnikov kernel.

2.7 Chapter 2 Remarks

Throughout the current chapter, we have seen a number of different approaches to optimization and estimation. At the beginning of the chapter, we explored harmonizable processes, which form a class of non-stationary models theoretically close to stationary processes. There, we found that asymptotically stationary processes may be treated quite similarly to stationary processes in terms of estimation and prediction.

After briefly touching on the topic of stochastic processes decomposition, we then explored dynamic linear models, which form a rather general class of non-stationary models. In that section, we also explored the forecasting and prediction of such models, focusing our attention again on those points where the existing methods fall short of our current goals, namely the goals associated with the prediction, estimation, and modeling of stochastic processes having both linear- and cycle-type trend in addition to Levy process noise.

Following our treatment of dynamic linear models, we then explored Bayesian estimation, if only as a candidate approach among candidate approaches. There, optimization was detailed in a regression context, with importance placed on the limitations, pitfalls, and benefits of using a Bayesian approach to modeling. Among the treatment of Bayesian inference concepts, we explored the differences between ordinary and empirical Bayesian approaches and looked at the primary focus points of a Bayesian analysis, namely model selection and sensitivity analysis, which provide us with a means of selecting and optimizing our hyperparameters.

Lastly, we explored certain results and concepts in nonparametric inference, which will prove fruitful in our longer-term goals for the current project.

CHAPTER 3:
LINEAR TREND SIGNAL DETECTION IN THE PRESENCE OF PERIODIC SIGNALS AND LEVY PROCESS NOISE

3.1 Statement of the Problem

Suppose that we have an additive stochastic process, S_t, such that

$$S_t = X_t + Y_t + Z_t \tag{203}$$

where X_t is a (deterministic) linear trend process, monotonic in t; Y_t is a (deterministic) Fourier series in t; and Z_t is a purely stochastic process (i.e., Z_t is a random variable for each time, t).

The goal of the current project is to determine, without knowledge of the components of S_t, the optimal sampling protocol such that a linear regression conducted on $\{t_k, S_{t_k}\}$ (i.e., the set of points chosen by some protocol) is optimally close to the trend sub-component, X_t.

To state this goal more clearly, suppose that we are given a number of measurements $N \in \mathbb{Z}$, a target time $T \in \mathbb{R}_{\geq 0}$, an initial time $t_0 < T$. Our goal is then to determine a subset of N such that

$$t_0 < t_{k_1} < t_{k_2} < \cdots < t_{k_n} < T \tag{204}$$

and $S_{t_k} \approx X_t$

Toward this goal, the optimal linear regression conducted on the sub-sample data should produce a model with

$$\beta \approx \frac{X_T - X_{t_0}}{T - t_0} \tag{205}$$

where β is the slope of the line generated by (linearly) regressing S_{t_k} on t_k.

3.2 Statistical Properties of the Problem

Before attempting to construct a solution to the current problem, let us venture to explore some of the properties necessary to understanding what such a solution should look like. Let us begin with an exploration of the moments of the processes involved. Since the moments of a generally-stated Levy Process are quite non-trivial, let us begin by assuming that the noise is a Wiener Process, which is a type of Levy Process. Once this noise assumption has had it's useful properties exhausted, we will then generalize the results to the more general class of processes (i.e., Levy Processes).

The mean, $\mu_{S_t} = E[S_t]$, of the overall (i.e., additive) process is given by

$$\mu_{S_t} = E[X_t + Y_t + Z_t] \tag{206}$$

$$= X_t + Y_t + E[Z_t] \tag{207}$$

$$= X_t + Y_t \tag{208}$$

where step 1 uses substitution, step 2 uses the linearity property of expectation as well as the property that a deterministic process has a mean equal to the value of that (deterministic) process, and step 3 uses the fact that our chosen type of noise (i.e., Wiener Process) has a mean of 0. Given that this derived expectation depends on t, the process S_t is non-stationary. Next, we explore the variation of the overall process.

The covariance, $K(s,t) = E[(S_s - \mu_{S_s})(S_t - \mu_{S_t})]$, of the overall process is given by

$$K(s,t) = E[(S_s - \mu_{S_s})(S_t - \mu_{S_t})] \tag{209}$$

$$= E[(X_s + Y_s + Z_s - (X_s + Y_s))(X_t + Y_t + Z_t - (X_t + Y_t))] \tag{210}$$

$$= E[Z_s Z_t] \tag{211}$$

$$= \min(s,t) \tag{212}$$

where step 1 uses the definition of the (auto-)covariance function, step 2 uses substitution for the processes involved, step 3 is a simplification of the previous step, and step 4 uses the familiar autocovariance/autocorrelation result for Wiener processes. Since this result

depends t but not merely through the difference between s and t, we see, once again, the process is non-stationary. Therefore, in its full generality, our problem concerns an additive stochastic process that is most certainly non-stationary.

In general, any computational results associated with the current problem formulation will come in the form of a sequence of discrete measurements (i.e., a time series), as computers are not yet equipped to handle continuous problems in their natural state. As such, let us explore the mean and autocovariance of the differenced process.

Let us define the (first) differenced process as follows:

$$\begin{aligned}
\Delta S_t &= S_t - S_{t-1} && (213)\\
&= (X_t + Y_t + Z_t) - (X_{t-1} + Y_{t-1} + Z_{t-1}) && (214)\\
&= (X_t - X_{t-1}) + (Y_t - Y_{t-1}) + (Z_t - Z_{t-1}) && (215)\\
&= (\alpha + \beta t - (\alpha + \beta(t-1))) + \Delta Y_t + \Delta Z_t && (216)\\
&= \beta + \Delta Y_t + Z_1 && (217)
\end{aligned}$$

where substitution, collection of terms, and the linearity of X_t (in t) are used. By the Gaussian increment properties of the Wiener process, Z_1 is a standard normal random variable, regardless of the value of time used in the differencing. Now that the differenced process has been defined and identified, let compute the mean and autocovariance of this process.

The mean of the differenced process is given by

$$\mu_{\Delta S_t} = \beta + \Delta Y_t \tag{218}$$

where similar steps as before were used to obtain this result. If ΔY_t is a function of t, then the differenced process, like the overall process, S_t, is non-stationary.

We may observe so far that it is the inclusion of both the linear- and cycle-type trend that makes the problem new and more difficult. In the mere presence of linear trend and noise, we may use the Generalized Least Squares approach outlined in Chapter 1. In

the mere presence of cycle-type trend and noise, we may use a modified Wiener filtering approach. It is in the presence of both types of trend that we must explore, entertain, and create new methods for the estimation, prediction, and modeling of such a process.

The auto-covariance of the differenced process is given by

$$K_{\Delta S}(s,t) = E[(\Delta S_s - \mu_{\Delta S_s})(\Delta S_t - \mu_{\Delta S_t})] \tag{219}$$

$$= E[(Z_s - Z_{s-1})(Z_t - Z_{t-1})] \tag{220}$$

$$= \begin{cases} 1 & |t-s| = 0 \\ 0 & |t-s| > 0 \end{cases} \tag{221}$$

which indicates that the autocovariance function for the differenced process meets the conditions for wide-sense stationarity, since the autocovariance function depends only on the difference between the input times, s and t. As mentioned in the mean calculation above, however, the process does not meet all of the conditions for wide-sense stationarity, as the mean derived is generally a function of time (through ΔY_t).

In order to obtain a stationary variant of the overall process, let us now explore a period-differenced version of the process. To this point, let P be the (single) period of the periodic signal of the problem, namely the period of Y_t. Defining the period-differenced process as follows, we may determine the mean and autocovariance of this differenced process.

$$\Delta_P S_t = S_t - S_{t-P} \tag{222}$$

$$= (X_t + Y_t + Z_t) - (X_{t-P} + Y_{t-P} + Z_{t-P}) \tag{223}$$

$$= (X_t - X_{t-P}) + (Y_t - Y_{t-P}) + (Z_t - Z_{t-P}) \tag{224}$$

$$= (\alpha + \beta t - (\alpha + \beta(t-P))) + \Delta_P Y_t + \Delta_P Z_t \tag{225}$$

$$= \beta P + \Delta_P Z_t \tag{226}$$

The mean of this period-differenced process is given by

$$\mu_{\Delta_P S_t} = \beta P \tag{227}$$

which meets the mean condition for wide-sense stationarity, due to its lack of dependence on time. The steps to calculate this value should be obvious to the reader, given the previous computations of this section.

The auto-covariance of the period-differenced process is given by

$$\begin{aligned}
K_{\Delta_P S}(s,t) &= E[(\Delta_P S_s - \mu_{\Delta_P S_s})(\Delta_P S_t - \mu_{\Delta_P S_t})] && (228)\\
&= E[\Delta_P Z_s \Delta_P Z_t] && (229)\\
&= E[(Z_s - Z_{s-P})(Z_t - Z_{t-P})] && (230)\\
&= E[(Z_s - Z_{s-P})(Z_{s+\tau} - Z_{s+\tau-P})] \qquad (t-s=\tau) && (231)\\
&= \begin{cases} P - |t-s| & |t-s| < P \\ 0 & |t-s| \geq P \end{cases} && (232)\\
&= K_{\Delta_P S}(0, t-s) && (233)
\end{aligned}$$

which establishes that the auto-covariance function of the period-differenced process meets the conditions for wide-sense stationarity. The most important property used to determine this result is the independent increment property of the Wiener process, which allows us to focus only on the interval on which $(Z_s - Z_{s-P})$ and $(Z_{s+\tau} - Z_{s+\tau-P})$ overlap, which happens to have length $P - \tau = P - |t-s|$. Since one of the times, s or t must be smaller than or equal to the other, the reader may focus their attention on the case where $s \leq t$, noticing that the result is symmetric about 0 (i.e., $t = s$).

Since wide-sense stationarity has been established for the period-differenced process for both the mean and auto-covariance function, the final step is to determine that the absolute value of the variance of the process is finite. By the Cauchy-Schwarz inequality

and the derivations above, we have that

$$|K_{\Delta_P S}(0, t-s)| \leq K_{\Delta_P S}(0,0) \tag{234}$$

$$= V[\Delta_P S_t] \tag{235}$$

$$= P \tag{236}$$

$$< \infty \tag{237}$$

where the equivalence between steps 2 and 3 is true for all $t \in \mathbb{R}$. Thus, we have established that the period-differenced process is wide-sense stationary. Of course, differencing is a type of linear filter, so let us describe this period-differencing filter in more detail before moving on to the next section, which involves the construction of a solution for the aims and goals of the problem.

The period-differenced filter can be written as

$$y_t = \Delta_P S_t \tag{238}$$

$$= S_t - S_{t-P} \tag{239}$$

$$= \sum_{j=-\infty}^{\infty} a_j S_{t-j} \tag{240}$$

where $a_0 = 1$, $a_P = -1$, and $a_j = 0$ for all other j. It should be obvious to the reader that the coefficients, a_j (known collectively as the *impulse response function*) satisfy the absolute summability condition, with their absolute sum being equal to 2, which is obviously finite. The frequency response function for this filter is given by

$$A_{yS}(\omega) = \sum_{j=-\infty}^{\infty} a_j e^{-2\pi i \omega j} \tag{241}$$

$$= 1 - e^{-2\pi i \omega P} \tag{242}$$

Since the period-differenced process, here defined as y_t, is wide-sense stationary, we have the following result by the Wiener-Khinchin theorem. More information about this

theorem may be found in Wiener (1930) or Chapter 11 of Champeney (1987).

$$s(\omega) = \int_{\mathbb{R}} r_y(\tau) e^{-2\pi i \omega \tau} d\tau \tag{243}$$

where $s(\omega)$ is the power spectral density of the period-differenced process (which is a function of the frequency, $\omega \in [0, 1]$), r_y is the auto-correlation function for the process y_t, and $\tau = t - s \geq 0$. To evaluate this relationship even further, let us first compute the auto-correlation function of the period-differenced process.

$$\begin{aligned}
r_y(\tau) &= E[y_t y_{t+\tau}] && (244)\\
&= E[(S_t - S_{t-P})(S_{t+\tau} - S_{t+\tau-P})] && (245)\\
&= E[(\beta P + \Delta_P Z_t)(\beta P + \Delta_P Z_{t+\tau})] && (246)\\
&= (\beta P)^2 + E[\Delta_P Z_t \Delta_P Z_{t+\tau}] && (247)\\
&= (\beta P)^2 + \chi_{|\tau|<P}(\tau)(P - |\tau|) && (248)
\end{aligned}$$

where step 1 uses the definition of the autocorrelation function, step 2 uses the definition of the period-differenced process, step 3 uses the previously-derived form for the aforementioned process, step 4 uses the fact that $E[\beta P \Delta_P Z_t] = \beta P E[\Delta_P Z_t] = 0$ for any t, and the final step uses the previously-derived form for the auto-covariance function of the process y_t.

Substituting this result into the Wiener-Khinchin theorem statement, we arrive at the following form for the power spectral density of the period-differenced process:

$$\begin{aligned}
s(\omega) &= \int_{\mathbb{R}} r_y(\tau) e^{-2\pi i \omega \tau} d\tau && (249)\\
&= \int_{\mathbb{R}} [(\beta P)^2 + \chi_{|\tau|<P}(\tau)(P - |\tau|)] e^{-2\pi i \omega \tau} d\tau && (250)\\
&= (\beta P)^2 \int_{\mathbb{R}} e^{-2\pi i \omega \tau} d\tau + \int_{\mathbb{R}} \chi_{|\tau|<P}(\tau)(P - |\tau|) e^{-2\pi i \omega \tau} d\tau && (251)\\
&= (\beta P)^2 \delta(\omega) + \int_{-P}^{P} (P - |\tau|) e^{-2\pi i \omega \tau} d\tau && (252)\\
&= (\beta P)^2 \delta(\omega) + P \int_{-P}^{P} e^{-2\pi i \omega \tau} d\tau - \int_{-P}^{P} |\tau| e^{-2\pi i \omega \tau} d\tau && (253)
\end{aligned}$$

which may be further simplified to the following form:

$$s(\omega) = (\beta P)^2\delta(\omega) + P\int_{-P}^{P} e^{-2\pi i\omega\tau}d\tau - \int_{-P}^{P}|\tau|e^{-2\pi i\omega\tau}d\tau \tag{254}$$

$$= (\beta P)^2\delta(\omega) + P\int_{-P}^{P}(\cos(-2\pi\omega\tau) + i\sin(-2\pi\omega\tau))d\tau - \int_{-P}^{P}|\tau|e^{-2\pi i\omega\tau}d\tau \tag{255}$$

$$= (\beta P)^2\delta(\omega) + P\int_{-P}^{P}\cos(2\pi\omega\tau)d\tau - iP\int_{-P}^{P}\sin(2\pi\omega\tau)d\tau - \int_{-P}^{P}|\tau|e^{-2\pi i\omega\tau}d\tau \tag{256}$$

$$= (\beta P)^2\delta(\omega) + 2P\int_{0}^{P}\cos(2\pi\omega\tau)d\tau - \int_{-P}^{P}|\tau|e^{-2\pi i\omega\tau}d\tau \tag{257}$$

$$= (\beta P)^2\delta(\omega) + 2P\left(\frac{\sin(2\pi P\omega)}{2\pi\omega}\right) - \int_{-P}^{P}|\tau|e^{-2\pi i\omega\tau}d\tau \tag{258}$$

$$= (\beta P)^2\delta(\omega) + 2P\left(\frac{\sin(2\pi P\omega)}{2\pi\omega}\right) - \left(\int_{0}^{P}\tau e^{-2\pi i\omega\tau}d\tau - \int_{-P}^{0}\tau e^{-2\pi i\omega\tau}d\tau\right) \tag{259}$$

Based on the statistical properties of the problem outlined previously in this section, we see that our process is non-stationary, that the first difference for our process is non-stationary, and that our period-differenced process is wide-sense stationary. Following these conclusions, we obtained a form for the power spectral density of the differenced process. These calculations should provide the reader with some insights into the nature of the overall additive stochastic process being considered. Let us now consider an optimization approach to the problem.

3.3 Optimization Theory for the Decomposition Problem

For the practitioner, the main difficulty presented by the general analysis of the decomposition problem stems from the fact that none of the existing approaches provides an *efficient* algorithm for the approximate trend-cycle-noise decomposition of actual data, in the sense of computational complexity relative to the size of the input. To illustrate why, regarded as an algorithmic problem, decomposition is in fact intractable, consider the case of a signal given by

$$S(t) = \beta_0 + \beta_1 t + \sum_{k=1}^{M}[a_k\cos(k\omega t) + b_k\sin(k\omega t)] + Z(t) \tag{260}$$

where the random variable $Z(t)$ is assumed to have the infinite divisibility property. Realistically, the decomposition problem starts with a sample of $N \in \mathbb{N}$ observations,

$$\Lambda_N = \{S(t_1), S(t_2), \ldots, S(t_N)\} \tag{261}$$

and we wish to find the best approximation of (260) on the Hilbert space

$$\mathcal{H}_\nu = \mathbb{C} \oplus L^2([0, 2\pi/\nu]) \tag{262}$$

that is to associate to Λ_N a vector of coefficients corresponding to the linear part (trend) $(\widehat{\beta}_0, \widehat{\beta}_1) \in \mathbb{C}$ and the oscillatory part

$$\sum_{n\in\mathbb{Z}} \widehat{c}_n e^{in\nu t} \in L^2([0, 2\pi/\nu]), \quad (\widehat{c}_n)_{n\in\mathbb{Z}} \in \ell^2(\mathbb{Z}) \tag{263}$$

Evidently, the main difficulty resides is the proper choice for the parameter $\nu \in (0, \infty)$, absent any information about the period of the oscillatory part of the signal, $2\pi/\omega$. Formally, the decomposition is found by projecting the derivative of the signal (260) onto $L^2([0, 2\pi/\nu])$, which then leaves only the determination of the constant terms $\widehat{\beta}_0$ to be accomplished by the Gauss-Markov theorem. Algorithmically, the numerical derivative of the signal would be computed from Λ_N in $o(N)$ operations, and the projection would add (according to Wiener's theorem) another $o(MN\omega/\nu)$ operations.

However, this is only true under the assumption that the true period of the signal, $2\pi/\omega$, is known. While many practitioners may find it reasonable to assume the value of the period based on a visual inspection of the data (or worse yet, intuition), this is not justified in general, as the process may contain various forms of cycling behavior, potentially with multiple period values and complex interplay between cycles. Therefore, all the theoretical results surveyed thus far may be seen as impractical in the real-world.

To illustrate this, consider the simple computation of the Fourier coefficients for the component $\cos(\omega t)$ from the oscillatory part of the signal.

The projection coefficients

$$\int_0^{2\pi/\nu} \cos(\omega t)\cos(n\nu t)dt = \frac{\omega[\cos(2\pi\omega/\nu) - 1]}{\omega^2 - n^2\nu^2} \tag{264}$$

will require computations up to n corresponding to the closest rational approximation of ω by $n\nu$, assuming that $\nu \in \mathbb{Q}$. But this is to say that for arbitrary values of ω, the sample selection may be insufficient and cause the algorithm to fail. In other words, there is no way to predict the minimum sample size N and the observation times $\{t_k\}_{k=1}^N$ needed for the application of the method. At the very least, these results highlight the inherent complexity of the problem, displaying once again where the existing methods and theory fall short of what is needed.

3.3.1 Large Deviations Functional and Optimal Sampling

The aim of the approach to be described in this section is to provide the mathematical framework for an algorithm to select the distribution of observation times (in a Bayesian sense) in such a way that the iterative process is proven to converge and that asymptotically it has for a limit the optimal distribution yielding the period of the oscillatory part of the signal and the best estimators for both trend and cycle components of the signal. This, of course, is no small task.

To begin, consider a simpler case of the signal:

$$S(t) = \beta_0 + \beta_1 t + a\cos(\omega t) + b\sin(\omega t) + B(t), \tag{265}$$

where the parameters $\beta_0, \beta_1, a, b, \omega$ and the variance σ^2 of the martingale $B(t)$ are not known. Let $p(t)$ denote the discrete distribution of N observation times $\{t_k\}_{k=1}^N$, and $\zeta > 0$ an auxiliary variable to be used in the asymptotic approximation of the large deviations functional for the process $S(t)$. Computing the characteristic function of the signal $S(t)$ with respect to the distribution $p(t)$, we have (by the independence assumption):

$$\phi_S(\lambda) = E(e^{i\lambda S(t)}) = e^{i\lambda\beta_0}E(e^{i\lambda B(t)})\sum_{k\in\mathbb{Z}}[i^k J_k(a\lambda) + J_k(b\lambda)]E(e^{i(k\omega+\lambda\beta_1)t}) \tag{266}$$

Denoting by $\varphi_p(\lambda)$ the characteristic function of the distribution of observation times $\{\tau_k = t_k - t_1\}_{k=1}^N$ reset to begin at 0, we arrive (in the case of Gaussian noise) at the expression

$$\phi_S(\lambda) = e^{i\lambda(\beta_0+\beta_1 t_1)} e^{-\frac{\sigma^2\lambda^2}{2}} \varphi_p(-i\lambda^2) \sum_{k\in\mathbb{Z}} [i^k J_k(a\lambda) + J_k(b\lambda)] \varphi_p(k\omega + \lambda\beta_1) \tag{267}$$

where we have used generating function formulas for the Bessel functions of the first kind. As the distribution $p(\tau)$ has compact support, we can use analytic continuation and write the moment-generating function for this distribution, $m_S(x) = \phi_S(-ix)$, by setting $\lambda = -ix$ in (267):

$$m_S(x) = e^{x(\beta_0+\beta_1 t_1)} e^{\frac{\sigma^2 x^2}{2}} m_p(x^2) \sum_{k\in\mathbb{Z}} [i^k J_k(-iax) + J_k(-ibx)] m_p(\beta_1 x - ik\omega) \tag{268}$$

Clearly, we have the conjugation identities

$$m_p(\beta_1 x - ik\omega) = \overline{m_p(\beta_1 x + ik\omega)} \tag{269}$$

which ensure, together with the properties of integer-index Bessel functions, the reality of $m_S(x)$ for real argument, and the following time - translation invariance result:

Theorem 3.10. If $\beta_1 = 0$, the model (265) with Gaussian noise has a stationary moment-generating function $m_S(x)$ independent of our choice of starting time t_1.

Moreover, in this case, uniform sampling of τ over the interval $[0, \frac{2\pi n}{\omega}], n \in \mathbb{N}$, allows us to retrieve the Wiener filtering result explicitly:

Theorem 3.11. If $\beta_1 = 0$, the model (265) with Gaussian noise and uniform sampling $\tau \sim U[0, \frac{2\pi n}{\omega}], n \in \mathbb{N}$ has the moment-generating function

$$m_S(x) = e^{\beta_0 x + \frac{\sigma^2 x^2}{2}} m_p(x^2)[I_0(ax) + I_0(bx)] \tag{270}$$

where $I_n(z)$ is the modified Besel function of the first kind.

The proofs of both theorems are straightforward and largely computational.

For the general case, we expand the Bessel functions into power series as entire functions of the argument, which leads to the formula

$$m_S(x) = e^{\beta_0 x + \frac{\sigma^2 x^2}{2}} m_p(x^2) \left[m_p(\beta_1 x) R_0(x) + 2\Re \sum_{k=0}^{\infty} R_k(x) m_p(\beta_1 x - ik\omega) \right] \tag{271}$$

where we have introduced the notation

$$R_k(x) = \left(\frac{x}{2}\right)^k \sum_{m=0}^{\infty} \frac{[a^{2m+k} + (-i)^k b^{2m+k}]}{(m!)(m+k)!} \left(\frac{x}{2}\right)^{2m} \tag{272}$$

for all positive integers k.

3.3.2 Asymptotic Expansions of the Large Deviations Functional

For $|x|$ sufficiently small, we can approximate the moment-generating function as follows:

$$m_S(x) \simeq e^{\beta_0 x + \frac{\sigma^2 + 2\langle T \rangle}{2} x^2} \left\{ 1 + [(\beta_1 - b\omega)\langle T \rangle + a]x + \frac{1}{2}\left(a^2 + 2a\beta_1\langle T \rangle + \langle\langle T \rangle\rangle \beta_1^2\right) x^2 \right\} \tag{273}$$

with $\langle T \rangle$ and $\langle\langle T \rangle\rangle$ representing the first and second moments of the distribution of sampling times, respectively. By completing the square for the quadratic term in parenthesis, we can summarize this asymptotic expansion of the moment-generating function in a distributional sense:

$$S(t) \sim N\left(\beta_0 + a + (\beta_1 - b\omega)\langle T \rangle, \sigma^2 + 2\langle T \rangle + (a + \beta_1\langle T \rangle)^2 + \beta_1^2 V[T]\right) \tag{274}$$

which (asymptotically) expresses the distribution of the process in terms of the parameters $\beta_0, \beta_1, a, b, \omega$, and σ, as well as the first and second moments of the distribution of the sampling times.

3.3.3 Optimal Sampling Distribution by Jeffreys Priors Bayesian Inference

We may arrive at an iterative converging procedure, which in turn, leads to the optimal sampling distribution, by considering the problem of Bayesian inference of the parameters

$$\mu_S = \beta_0 + a + (\beta_1 - b\omega)\langle T \rangle \tag{275}$$

$$\sigma_S^2 = \sigma^2 + 2\langle T \rangle + (a + \beta_1 \langle T \rangle)^2 + \beta_1^2 V[T] \tag{276}$$

characterizing the Normal approximation conditioned on the sampling distribution for the signal $S(t)$, by recalling that the Bayesian inference will be self-conjugate when using Jeffreys priors, that is uniform sampling for the parameters $\langle T \rangle \in \mathbb{R}$, $V[T] > 0$. This is consistent with the fact that for uniform distributions, sampling times form a two-parameter family, where

$$T \sim U[\tau, \tau + L] \Rightarrow \langle T \rangle = \tau + \frac{L}{2} \in \mathbb{R} \tag{277}$$

and

$$V[T] = \frac{L^2}{12} > 0 \tag{278}$$

which is to say we have the following iterative procedure:

Theorem 3.12. Jeffrey's Sampling for the Decomposition Problem

For the process (265), let $\mathcal{H}[\mathbb{R} \times (0, \infty)]$ be the space of Haar measures on the product space $\mathbb{R} \times (0, \infty)$. For any $L > 0$, consider a sequence of measures $p_j \in \mathcal{H}[\mathbb{R} \times (0, \infty)]$ uniform on $[\tau_j, \tau_j + L]$ with mean $\tau_j + \frac{L}{2}$ and fixed variance $\frac{L^2}{12}$. Then the conditional processes

$$\left(S_j \middle| \langle T \rangle = \tau_j + \frac{L}{2}, V[T] = \frac{L^2}{12} \right) \tag{279}$$

have unbiased sample estimators $\widehat{\mu}_{S_j}, \widehat{\sigma^2}_{S_j}$ whose linear regression on τ_j and τ_j^2 yields estimates for the curvature $\widehat{\beta}_1^2$, slope $\widehat{a}\widehat{\beta}_1$, and the intercepts $\widehat{\beta}_0 + \widehat{a}$ and $\widehat{\sigma}^2 + \widehat{a}^2$ (i.e. the trend-noise components of (265)). So far in this process, we have estimates for all parameters except for b and ω, which are to be estimated in the following steps.

Once the trend-noise components of the process have been isolated, the remaining oscillatory component will be optimally estimated by the choice of L which matches a multiple of the period $\frac{2\pi}{\omega}$ and uniform sampling. This brings us to the final theorem of this document, which will find computational treatment and further exploration in future research.

Theorem 3.13. For the signal $S_0(t) = a\cos(\omega t) + b\sin(\omega t)$, a sample of N observations with distribution p will yield zero mean if and only if $\frac{\omega L}{2\pi} \in \mathbb{N}$.

The proof is immediate using the fact that the sample average of $S(t_n)$ can be written as

$$\overline{S(t_n)} = a\mathbb{R}\phi(\omega) + b\Im\phi(\omega), \tag{280}$$

and the characteristic function on $[\tau, \tau+L]$ vanishes identically if and only if $\exp(\omega L) = 1$.

3.4 Concluding Remarks

Throughout this document, we have focused our attention on the goal of modeling and estimating an additive stochastic process with both linear and oscillatory signals along with Levy process noise. To this end, we have discussed a variety of topics, some more on base than others. Wherever possible within this document, we have also focused our attention on those points where the existing theory is inadequate for our purpose. In terms of limitations not yet discussed, there is the issue of theoretical breadth within this document, mostly borne about by time constraints on the project. In the always relevant words of Blaise Pascal, "I would have written a shorter letter, but I did not have the time." This is not to say that the project is unfocused. Instead, these words are meant to reflect the inherent conundrum of modern research, in that we are given time schedules for problems that don't always match the schedules. Regardless of this issue, the end of this document contains new results which provide some hope for the analysis of more complicated stochastic processes. We hope that you have enjoyed this rather complicated expression of mathematical ability and hope also that you decide to chase giant dissertation-like goals if that's what you're into. Please, be well.

REFERENCES

[1] Aitken, A. C. (1936). On Least-Squares and Linear Combinations of Observations. *Proceedings of the Royal Society of Edinburgh*. 55, 42-48.

[2] Gikhman, I. I., Skorokhod, V. A. (1969). *Introduction to the Theory of Random Processes*. Courier Corporation.

[3] Billingsley, P. (2008). *Probability and Measure* (3rd Ed.). Wiley.

[4] Ross, S. (1996). *Stochastic Processes* (2nd Ed.). Wiley.

[5] Satō, K. (1999). *Lévy Processes and Infinitely Divisible Distributions* (English Ed.). Cambridge University Press.

[6] Teodorescu, I. (2013). *Optimization in Non-parametric Survival Analysis and Climate Change Modeling*. ProQuest Dissertations Publishing.

[7] Shannon, C. E. (2001). A Mathematical Theory of Communication. *Mobile Computing and Communications Review*, 5(1), 3–55. https://doi.org/10.1145/584091.584093

[8] Marvasti, F. (2001). *Nonuniform Sampling Theory and Practice* (1st Ed.). Springer US. https://doi.org/10.1007/978-1-4615-1229-5

[9] Astrom, K. J. and Bernhardsson, B. M. (2002). Comparison of Riemann and Lebesgue Sampling for First Order Stochastic Systems. *41st IEEE Conference on Decision and Control, Las Vegas, NV, United States, 2002-12-10 - 2002-12-13, 2*, 2011–2016 vol.2. https://doi.org/10.1109/CDC.2002.1184824

[10] Loeve, M. (1965). *Probability Theory*. Van Nostrand, New York.

[11] Hurd, H. L. (1989). Representation of Strongly Harmonizable Periodically Correlated Processes and their Covariances. *Journal of Multivariate Analysis, 29*(1), 53–67. https://doi.org/10.1016/0047-259X(89)90076-6

[12] Martin, M. and Putinar, M. (1989). *Lectures on Hyponormal Operators*. Birkhäuser Verlag.

[13] Niemi, H. (1975). Stochastic Processes as Fourier Transforms of Stochastic Measures. *Ann. Acad. Sci. Finn. A I Math.* 591, 1-47.

[14] Chang, D. K. and Rao, M. M. (1987). Bimeasures and Nonstationary Processes. In *Real and Stochastic Analysis* (M. M. Rao Ed.), Chapter 1. Wiley, New York.

[15] Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions. *The American Statistician.* 49: 327-335.

[16] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd Ed.). Chapman and Hall/CRC.

[17] Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods* (1st Ed.). Springer New York. https://doi.org/10.1007/978-1-4419-0925-1

[18] Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov chains and Their Applications. *Biometrika, 57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

[19] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics, 7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

[20] Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics, 27*(3), 832–837. https://doi.org/10.1214/aoms/1177728190

[21] Epanechnikov, V. A. (1969). Non-Parametric Estimation of a Multivariate Probability Density. *Theory of Probability and Its Applications, 14*(1), 153–158. https://doi.org/10.1137/1114019

[22] Hodges, J. L. and Lehmann, E. L. (1963). Estimates of Location Based on Rank Tests. *The Annals of Mathematical Statistics, 34*(2), 598–611. https://doi.org/10.1214/aoms/1177704172

[23] Wiener, N. (1930). Generalized harmonic analysis. *Acta Mathematica, 55*, 117–258. https://doi.org/10.1007/BF02546511

[24] Champeney, D. C. (1987). *A Handbook of Fourier Theorems*. Cambridge University Press.