ABSTRACT

Some New Applications of Bayesian Longitudinal Models

Jonathon Vallejo, Ph.D.

Chairpersons: Matt D. Hejduk, Ph.D. and James D. Stamey, Ph.D.

In this dissertation we consider some novel applications of Bayesian longitudinal methods. As inference is generally focused on response of an individual, we work within the mixed model framework. The two applications are described below.

Our first application is to a data set containing measurements of the probability of collision between two space objects orbiting the Earth. These measurements are longitudinal in nature, as they are observed over time and vary according to which two satellites they are taken on. This application presents a number of specific challenges, such as measurements at irregular time intervals, sparse data, and a bounded response variable. The second application is that of longitudinal network meta-analysis. In clinical trials, one major question is how to compare treatments across trials. However, current methods usually only deal with comparisons at a single time point, discarding data at other time points. This problem presents different challenges from the previous, such as defining network treatment effects over time, developing diagnostic methods for choosing a correct model, and dual longitudinal models for the mean and variance.

Some New Applications of Bayesian Longitudinal Models

by

Jonathon Vallejo, B.A.

A Dissertation

Approved by the Department of Statistical Science

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

Matt D. Hejduk, Ph.D., Chairperson

James D. Stamey, Ph.D., Co-Chairperson

David J. Kahle, Ph.D.

Jack D. Tubbs, Ph.D.

Jerry Z. Park, Ph.D.

Accepted by the Graduate School
August 2016

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to acknowledge the committee for their thoughtful suggestions for improving the dissertation and for adding to the scholarship of this work.

I would like to thank Dr. Stamey specifically for helping recruit me to come to Baylor. Without you, I would probably still be working as a telemarketer in Omaha, agonizing over why it is so difficult to sell magazines to strangers. More broadly, I want to thank all of the professors in the statistics department. You have allowed me to carve out my own path, and trusted that I would responsibly use the generous length of rope afforded to me. Hopefully this allowance has produced novel research which justified the freedom.

I also owe a special thanks to Dr. Hejduk, who supervised my work on the first two chapters. Dr. Hejduk generously agreed to oversee my naive foray into the world of conjunction assessment. You devoted countless hours to helping me understand all of the different aspects of the problem, and helped me stay on track to meet all of the deadlines. More than that, you endured the painful learning process of publishing papers in Latex and poured over my many lines of MATLAB code. More than simply advising, you were often in the trenches of research with me, running simulations and making graphs.

Lastly, I would like to thank my family and friends for supporting my continued education. My parents have been behind every ambition, misstep, and flight of fancy I've had. These have not always been mutually exclusive. I've been incredibly fortunate to come in with a cohort from which I've made what promise to be lifelong friends. I'm proud to be finishing alongside RJ, an often partner in crime who embarked with me on various statistical and non-statistical endeavors, with varying

degrees of success. The last person I'd like to thank is my partner Michelle, whom I met in the program, and who has provided endless support throughout the process.

DEDICATION

To

My parents

CHAPTER ONE

Trending in Probabilities of Collision: Background and Motivation

1.1 Introduction

Satellites have become an integral part of modern life, supporting phone communication, television and radio broadcasting, internet access, and military activities. Indeed, it is difficult to imagine modern society without many of these technologies, especially in an age when the world is increasingly interconnected via long-distance communications. As of 2013, there were over one thousand operational satellites in orbit about Earth. About half of these active satellites are in Low-Earth Orbit (LEO, meaning an orbital period less than 225 minutes), which is where the International Space Station (ISS) conducts operations, along with other commercial missions such as earth observation and satellite telephone communications. An increasing amount of attention is being placed on protecting satellites in LEO, as the frequency of object launches and satellite fragmentation events has contributed to the proliferation of space debris, resulting in increased congestion. Kelly[47] notes that "the number of space objects has greatly increased in the past 15 years and is currently estimated to be 500,000 objects between 1 and 10 cm and 100 million objects less than 1 cm." She goes on to observe that these objects have three sources: 1) debris from satellites, 2) non-operational or "dead" satellites, and 3) operational satellites that may or may not be able to maneuver. In addition to these smaller objects, other sources suggest that the number of objects greater than 10 cm is roughly 20,000.

As a result of the growing amount of debris in the commonly used orbits, there has been an increased focus on protecting satellites from potential collisions with other objects. For example, NASA produced the first orbit debris mitigation

guidelines in 1995, although a formal policy was not promulgated until 2007. Though space objects had been cataloged since the late 1950's, these were the first procedural attempts to give guidance on how to manage close approaches between two objects and ultimately how to avoid collisions. In 2005, NASA established agency-wide protocol for performing collision analysis and reactions to close approaches. A project office called Conjunction Assessment Risk Analysis (CARA) was created to perform these analyses for robotic missions, and it currently provides this service to about 65 NASA and civil space satellites.

Although the probability of two space objects colliding is often negligible, collisions do occur; and their impact on future space congestion is often tremendous. Since 1991, eight on-orbit collisions have been reported, the last occurring in 2009 when an Iridium communications satellite was hit by an inactive Russian COSMOS satellite. This collision created two debris clouds of approximately 500 and 1,300 objects that have been subsequently cataloged. In addition to this debris, in 2007 the Fengyun 1C satellite was deliberately destroyed, creating another debris cloud of approximately three thousand cataloged objects. Though only the Iridium reflects the case of a collision that was avoidable by collision mitigation procedures, these two events exemplify how much impact collisions can have on the space debris population. For instance, 50% of detected close approaches in LEO involve a debris object from one of these three clouds. In effect, these two collisions doubled the number of close approaches tracked in LEO and consequently increased both the risk of further collisions and the amount of work needed to mitigate this risk.

The Iridium-COSMOS collision was the impetus for a significant increase in breadth and sophistication of conjunction assessment activities, as it represented a worst-case scenario for those attempting to mitigate collisions and the proliferation of space debris, as both satellites were completely intact and collided at hypervelocity (¿ 6,700 mph). In fact, these two satellites collided at the speed of 26,170 mph, which

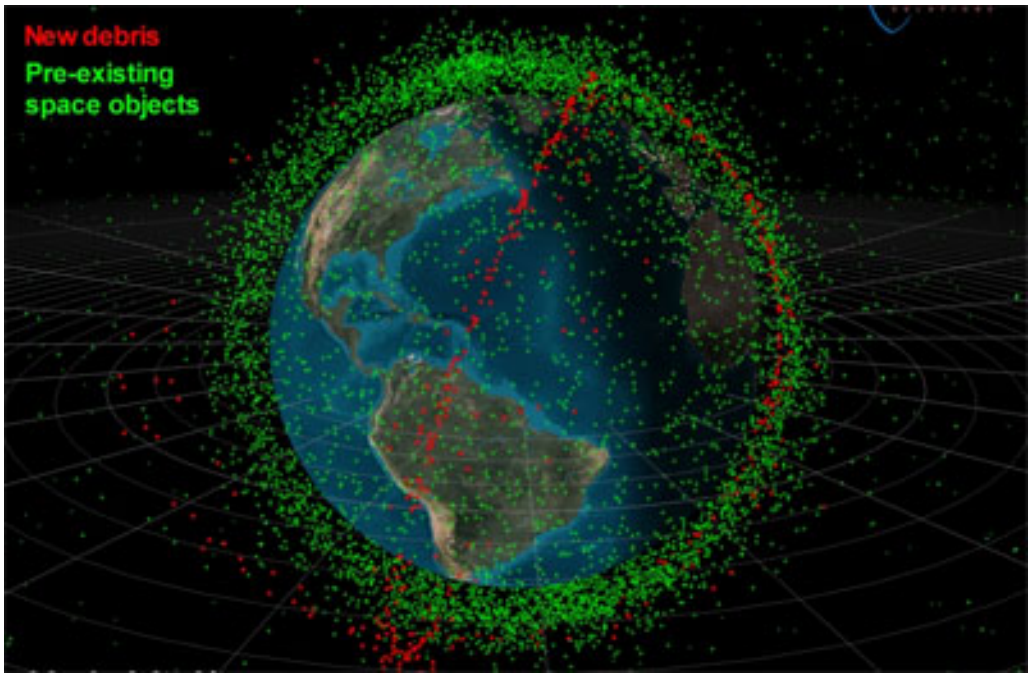

Figure 1.1. Visualization of new debris created by the Iridium 33/Cosmos 2251 collision (from Chan[19])

resulted in the creation of a huge debris cloud of the number of pieces outlined previously, with perhaps one hundred times that number of pieces too small to be tracked by current radars. The proliferation of additional conjunction events that this debris field generates will only increase the requirements for accurate and meaningful conjunction risk assessment.

For some years, this risk assessment was based only on the closest predicted miss distance between the two conjuncting objects. While this construct has immediate intuitive appeal and is easy to communicate conceptually to decision-makers, because it does not consider the uncertainties of the satellite trajectories, it tends to produce results that are difficult to interpret. Consider Figure 1.2, which for 14 reports leading up to the Iridium 33 collision, graphs the closest predicted approach of all tracked space objects, the closest predicted approach for any satellite within the Iridium constellation, as well as for the Iridium 33. The black line indicates the closest predicted approach between the Iridium 33 and the Cosmos 2251, which

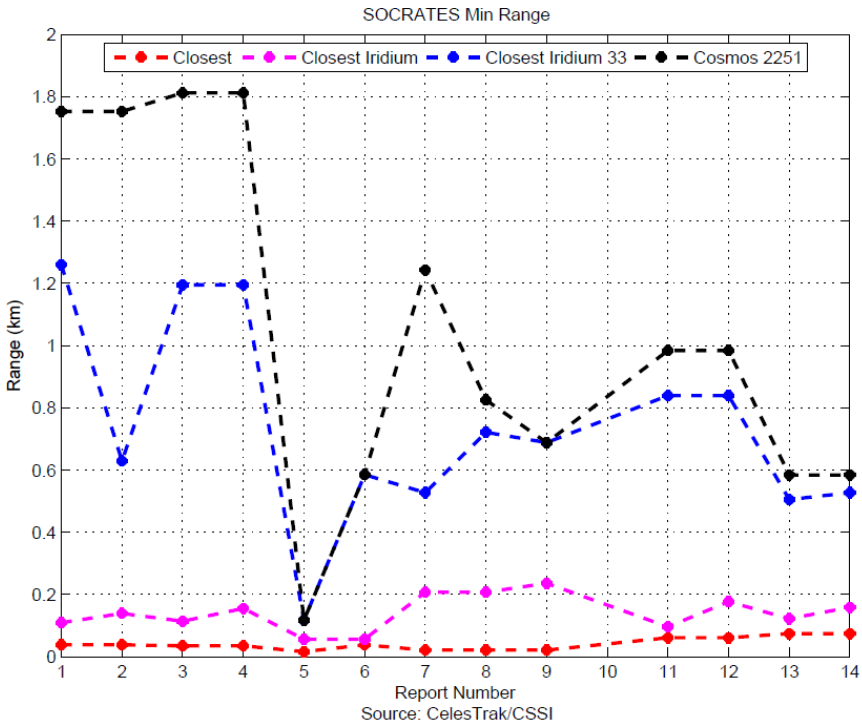

Figure 1.2. Closest approach distance for all tracked objects (from CelesTrak[1])

can be seen to be predicted farther than the closest object. This highlights a major difficulty in assessing risk by using miss distance: often there are multiple serious threats, and deciding which is most imminent is not straightforward, especially when based on a metric that is not actually a measure of conjunction likelihood.

Figure 1.3 gives further evidence of the poverty of this particular risk assessment paradigm. This figure shows what rank of risk the Iridium 33/Cosmos 2251 conjunction was of all conjunctions, those for just the Iridium constellation, and for the Iridium 33. Over the 14 reports, the Iridium 33/Cosmos 2251 conjunction varies from a rank of 1,611 of all potential conjunctions on report 3 to a rank of 11 on report 4. Thus, when compared to all possible events, the Iridium 33/Cosmos 2251 varies quite a bit on how comparatively "risky" it is. Even if one were to use its risk rating from the later reports, one would still conclude that it is less serious than 150 or 400 events.

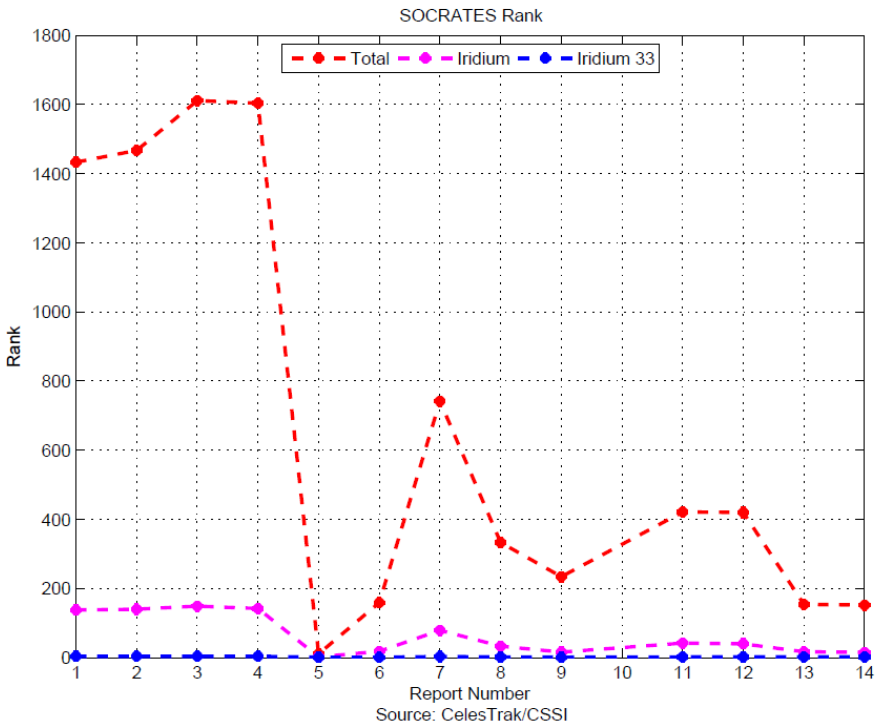

Figure 1.3. Closest approach distance for all objects in the Iridium constellation (from CelesTrak[1])

The level of threat is still ambiguous when one considers the comparative risk of the Iridium 33 to only satellites within the Iridium constellation. If one were to use the first four reports, one would conclude that the Iridium 33/Cosmos 2251 conjunction is less serious than about 150 other conjunctions. Even if one were to use later reports, this conjunction is never the most serious of all the satellites in the constellation. Thus, from the available data, it is hard to pinpoint that this is the event which needs most attention.

The situation can be ameliorated with a different type of calculation that considers the uncertainty in the state estimates, which will allow the significance of the miss distance to be assessed. If the uncertainties about both states are much smaller than the miss distance between the two objects, then the conjunction is not particularly worrisome even if the miss distance seems small in an absolute sense—the two objects' states are so well determined that one can have confidence that

the two objects really will pass each other with the calculated distance. Conversely, a larger miss distance with uncertainties about the same size as this miss distance could well have a probability of collision large enough to be of concern even though the miss distance might itself seem large. This calculation, called the Probability of Collision, will be discussed in depth in a subsequent section.

The result of these debris-producing events is the potential to beget more debris, a phenomenon known as Kessler syndrome. In 1978, Kessler[49] posited that, due to the increase in objects in space, satellite collisions would be inevitable and create further debris, in turn increasing the risk of future collisions. Kessler predicts that "the result would be an exponential increase in the number of objects with time, creating a belt of debris around the earth." He likened the process for creating this "debris belt" to the creation of the asteroid belt, though at an admittedly faster rate. Interestingly, Kessler predicted that the first satellite collision could be expected to occur in around 1989 (with a more conservative estimate placing the year at 1997), based on his estimate of an increase of 13%/year growth rate of debris. The first documented satellite collision occurred in 1991 between the Cosmos 1934 and debris from the Cosmos 296. The domino effect described by the Kessler syndrome is most likely in LEO, as this orbit regime contains by far the most space debris. Primack[64] notes that this would not only endanger the International Space Station and Hubble telescope, but also eventually GPS and other communications satellites.

The discussion above outlines the significant risks and ramifications associated with satellite collisions and makes it clear that there is a need for a systematic, sophisticated way of assessing and mitigating these risks. However, there are technical and logistical difficulties in implementing an effective system. The discussion surrounding the Iridium 33 collision gives a broad sense of both, suggesting the technical difficulty in calculating a reliable assessment of risk and the logistical difficulty associated with choosing among hundreds of similarly risky events. We consider

both types of difficulty further in order to give a perspective on limitations that add ambiguity to the process and argue for more sophisticated assessment techniques. The subsequent discussion is arranged in sections that address the definition of key terms, the data collection and distribution process, the sources of these data via the orbit determination process, and the calculation of the key parameter presently used in conjunction risk assessment, the probability of collision.

1.1.1 Definition of Terms

The ensuing discussion will benefit from defining certain key terms more precisely, so we provide such definitions here. The process of producing the model parameters to allow the prediction of a satellite's future position is called orbit determination (OD); it is a filter estimation process that combines actual sensor observations of satellite positions and other *apriori* information to generate a satellite state estimate (estimate of satellite position and velocity) at a given time, called an epoch time; a robust estimation process will also produce an estimation error covariance matrix, which will specify the expected uncertainties in the estimated parameters and the correlations among them. This information can be used by a satellite propagator to predict satellite positions and velocities at a future time.

When two objects are expected to pass within close proximity of each other, they are said to be in conjunction. A potential collision between two space objects is interchangeably referred to as an event or a conjunction. The primary space object is defined to be the satellite one is attempting to protect (generally speaking, an active satellite). The secondary is the object that is endangering the primary, and this object is usually a piece of debris, although as seen in the Iridium 33 collision, it can be an intact satellite. The time of closest approach (TCA) is the time at which these two objects are predicted to be closest, and the position of closest approach (PCA) is the corresponding position of each object at that time. Conjunction Analysis is the

process of determining which spacecraft will be in conjunction with a protected asset over a time period of interest, and Conjunction Risk Assessment is the process of determining the level of collision risk that each of these conjunctions presents. The probability of collision (P_c) is an empirically calculated probability of the two objects colliding, based on a few simplifying assumptions; and it serves as the principal parameter for assessing collision risk. An entire section is dedicated to explaining the calculation of this parameter.

1.1.2 Conjunction Data Distribution

Because a complete, up-to-date catalogue of the positions and velocities of all known satellites is needed for Conjunction Analysis, this portion of the daily calculation process takes place at the Joint Space Operations Center (JSpOC) at Vandenberg AFB, CA, where the Space Catalogue is actively maintained. Screening runs are executed in which each protected asset, with a box about it of carefully chosen dimensions, is "flown" several days into the future; and any other catalogued objects that penetrate this box are identified as conjunctors.

Once a screening run is complete, the results are further processed to generate the orbital information needed to perform Risk Assessment. The precise TCA is determined, and the two satellites PCAs are calculated. In addition to these position and velocity data that constitute the PCAs, the state error estimates at TCA, represented as covariance matrices, are also provided; these give a statement of the expected variance and covariance of each of the position and velocity components (the estimation process is presumed to be unbiased and therefore produce mean errors of zero), as well as the additional solved-for parameters of atmospheric drag and solar radiation pressure. Finally, information about the force model settings used in the OD is also provided so that, should the user of the data wish to propagate the solution, this can be done with the same model settings enabled. These data are

collected into what is called a Conjunction Data Message (CDM) and distributed as a discrete message to the owner or protector of the primary asset.

1.1.3 Conjunction Data Quality

Before considering methods for quantifying risk, we must first consider limitations that may be imposed due to the quality of the data used for this purpose; and such an investigation has two parts: the quality of the sensor observations that feed the OD process and the quality of the OD modeling itself. We will treat each of these in turn, beginning with the issue of sensor data quality.

Observation data are collected by a variety of space sensors that constitute the Space Surveillance Network (SSN). These include dish and phased-array radars, which provide range-to-target and two angles from the sensor to the spacecraft; optical telescopes, which observe two angles but cannot observe range-to-target; and occasionally other sensor types, such as interferometers or radio-frequency trackers, which typically provide angular data. Radar data are typically reasonably accurate in their range determination but not nearly as reliable in the angular measurements; angle measurements by optical sensors are frequently quite accurate (since they are calculated with reference to the star background, which is accurately known), but there is no range measurement provided.

Because it is difficult to track LEO satellites with telescopes and non-LEO satellites with most radars, it is unusual for a satellite to receive both types of tracking and allow the strengths of both sensor phenomenologies to complement each other. So there are errors in observational data due to inherent weaknesses in the different sensor types. Additionally, observations are typically taken in "tracks," or groups of observations all obtained during the same observing session; one might receive a set of, say, six observations in the span of one minute. While groups of data are certainly welcome, if correlation exists among the observations then the basic

premise of most estimation techniques—that measurements are uncorrelated—will not be strictly met.

A second issue with tracking data is irregularity of supply. The sensors in the SSN each have different detection and tracking capabilities, meaning that while all sensors can, for their orbit regime of specialty, track large objects, only a narrow subset can track the smallest objects. Because much of the Space Catalogue consists of debris objects and most debris objects are small, only a few of the SSN sensors are responsible for tracking a good bit of the Space Catalogue; and there is contention for the tracking resources of these sensors. The JSpOC possesses a software-managed sensor tasking paradigm that assigns the tracking of certain objects to certain sensors; but if only a few sensors are responsible for most of the catalogue maintenance, even with the priority scheme that this sensor tasking functionality allows, many debris objects receive far less tracking than one would wish. Furthermore, sensor outages, space weather phenomena, and radar energy misapplications can all conspire to encumber tracking throughput yet further. Lower tracking levels leave the OD process more vulnerable to sensor observation errors and provide a weaker fit overall.

Once tracking data are obtained, they are subjected to a batch minimum-variance estimation process to generate an updated set of orbital parameters for the satellite. The process begins by the appropriate choice of force model parameters, which includes the selection of the proper fidelity of a geopotential model (number of spherical harmonics to solve for in the Laplace equation series expansion that models the irregular Earth's gravity field), the effects of non-Earth gravity (such as the sun and moon), the geopotential irregularity introduced by liquid and solid Earth tides, and parameters that govern the solutions for atmospheric drag and solar radiation pressure. Next, a proper fit-span of observations needs to be chosen, as the batch technique does not correct for each observation sequentially but considers

the entire dataset as a "batch." Choosing a group of sensor observations that goes back too far in time (too long a fit-span) tends to weaken the solution for prediction into the future; choosing a group that does not go far back enough (too short a fit-span) tends to produce a poor atmospheric drag solution. Finally, the observation set must be reviewed for "outlier" data that can corrupt the estimated solution. While such an enterprise should of course be conducted with care since there is no *apriori* reason to suspect any particular observation, given the volume of objects and observations most such exclusions must be performed by computer, which is a much less robust process than a trained analyst's data exclusion through the visual review of residual plots. All of these areas are ripe for error that can weaken the correction and therefore the generated state estimate.

Finally, in order to produce data that can be used for conjunction risk assessment, the epoch states of the primary and secondary objects must be propagated to TCA. The principal source of error in this propagation is the inability to model the atmospheric density accuracies over the propagation interval, as the atmospheric drag acceleration on the satellite depends on the local atmospheric density. The atmospheric density is difficult to estimate because it requires estimating the atmospheric temperature, and a variety of factors influence this temperature. Since the temperature is generally governed by the extreme ultraviolet (EUV) heating of atmospheric gases by the sun, temperature is a function of time of day and latitude, as well as the sun's 27-day rotation cycle and 11-year cycle of activity. Acutely, the temperature is also affected by solar ejecta that enter the earth's atmosphere through the polar cusp and heat gases through the manipulation of the earth's magnetic field; this is a product of solar storm activity and is extremely difficult to predict. There can thus be considerable error associated with producing satellite future predicted positions; and given that the model cannot represent these processes well, it is unlikely that the covariance matrix emerging from the fit will model the error robustly.

It is often necessary, therefore, to add a consider parameter to the drag variance in the covariance matrix in order to try to represent the atmospheric density error more completely; while this approach is certainly welcome as an improvement over using the unaltered covariance directly, it is an imperfect compensation method.

For all of these reasons, the information contained in a CDM to describe a conjunction is uncertain and subject to change with future tracking and OD updates; this is the reason that a single estimate of the situation taken several days from the expected event is not adequate for Risk Assessment and that the more elaborate trending approaches explored by this research are warranted. Before turning directly to these methods, however, it is necessary to explain the calculation of the probability of collision (P_c), as it is the parameter used by the industry as the single encapsulation of collision risk. Having just discussed the potential issues with the data, one can observe how these issues work their way through the calculation.

1.1.4 Calculation of Probability of Collision

As mentioned above, provided in the CDM is a parameter called the probability of collision (P_c), which is generally considered to be the best possible measure for quantifying the risk for an event. Here, we briefly outline the methods used for calculating this value. As noted previously, each conjuncting satellite has an estimated position at TCA, about which a 3-dimensional error covariance is estimated. This covariance is ellipsoidal, and, for near-Earth orbits, usually oriented in such a way that the semi-major axis is close to the direction of the velocity of the object.

An image depicting these assumptions is given in Figure 1.4. This ellipse is usually defined in terms of radial, in-track, and cross-track (RIC) coordinates, which is a satellite-centered coordinate system. The radial direction is the direction of the position vector emanating from the earth, the in-track direction is the direction along the trajectory of the object, and the cross-track direction is perpendicular to

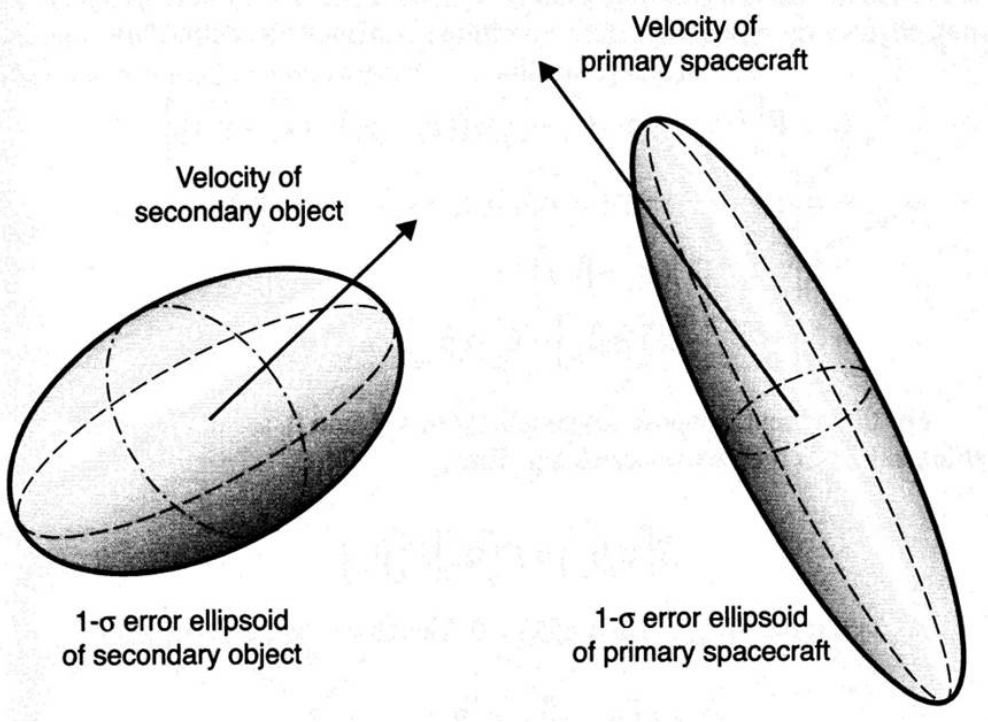

Figure 1.4. Visualization of a conjunction (from Chan[19])

these two vectors (using the right-hand rule). Though the velocity vectors depicted are not aligned with their respective in-track vectors, in practice they are usually observed to be closely aligned, and thus are assumed to be aligned.

These vectors are sometimes given the alternative designation UVW, as seen in figure 1.5. It is usually the case that the covariance ellipsoid is longest in the in-track direction, so that it is most difficult to be accurate about where on its trajectory a satellite is when TCA occurs.

This phenomenon is depicted in 1.6, where we see the ellipsoid longest in the "V" (in-track) direction. In practice, we further assume that the ellipsoidal errors associated with positional uncertainty are trivariate Gaussian. This implies that the mean of the distribution of each object is taken to be the calculated position at TCA. These covariances are presumed to be uncorrelated, implying that the total positional uncertainty can be calculated simply by the sum of the two covariances (after having

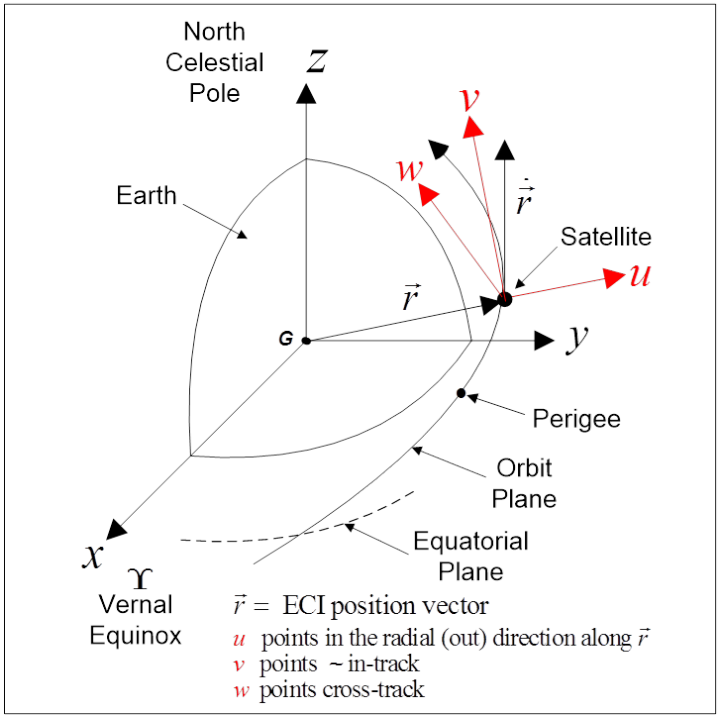

Figure 1.5. Visualization of trajectory using UVW coordinates (from Barker[7])

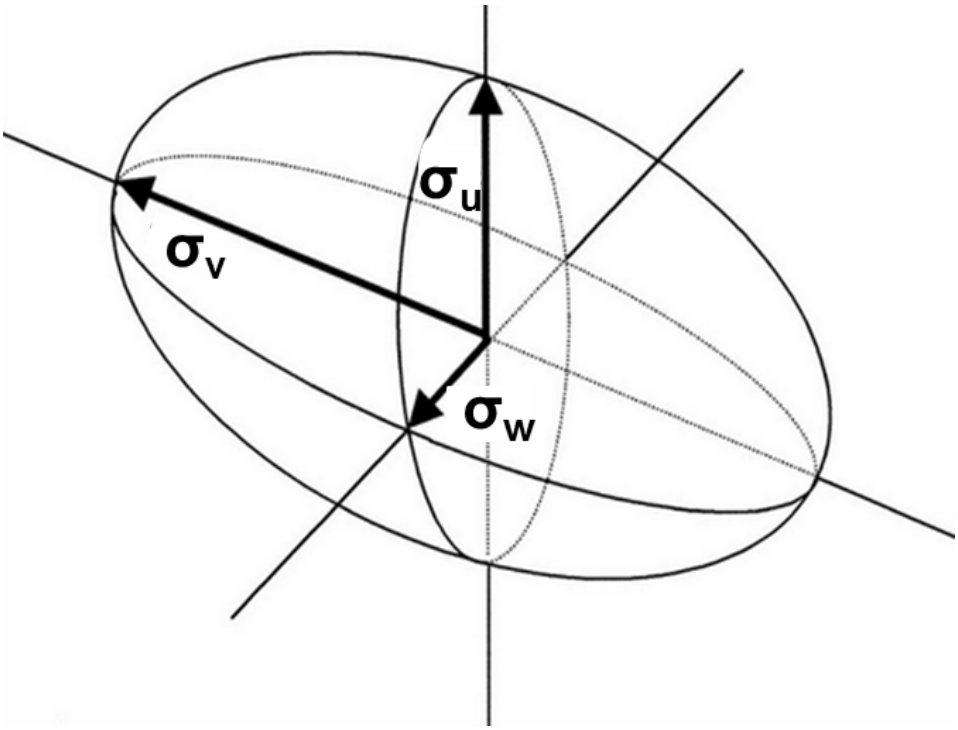

Figure 1.6. Positional error covariance ellipsoid defined by UVW coordinates (from Chan[19])

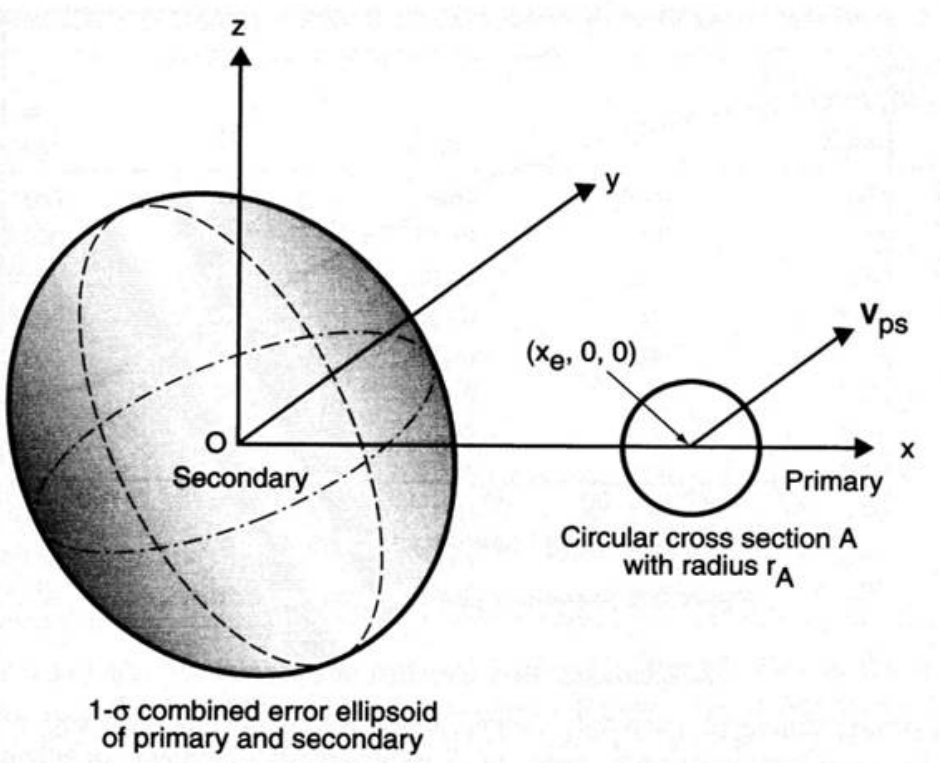

Figure 1.7. Projection of conjunction into the conjunction plane (from Chan[19])

been rotated to be expressed in the same coordinate system). Traditionally, we take this combined covariance and center it about the secondary object. Likewise, the radii of circumscribing spheres about each object are summed to create a single combined hard body sphere, which is placed at the location of the primary object. This problem is equivalent to the original problem involving two separate Gaussian densities due to the assumption that the covariances are uncorrelated. The result of these assumptions is visualized in Figure 1.7, though in two dimensions as opposed to three.

One last assumption generally made is that of rectilinear motion near the time of conjunction, so that the dimensionality of the problem may be reduced. If the conjunction between the two satellites takes place at high velocity, then the relative motion in the neighborhood of the conjunction will be rectilinear; and a collision, should it take place, will occur in a plane normal to the relative velocity vector

between the two objects. We can then project the combined covariance and the hard body sphere into this plane and consider the situation as a two-dimensional problem: a circle, resulting from the projection of the hard body sphere, and a covariance ellipse, resulting from the projected combined covariance. We are then interested in the probability of the area swept out by the circle in the probability density formed by the ellipse.[19]

The process for calculating P_c outline above is generally used for events which are approached at a high velocity. If the velocity is sufficiently small (¡ 10 m/s), many of the assumptions above break down (such as rectilinear motion). In such a case, we use a Monte Carlo approach, simulating millions of trajectories for both space objects and counting the number of times the miss distance is below a pre-specified threshold. This is obviously more time consuming than the above approach and therefore is employed only when necessary.

1.2 Problem Specification

As made clear by the preceding section the problem of deciding whether to maneuver a satellite which is in conjunction with another space object is often not straightforward, and a serious collision threat often involves the deliberation and cooperation of various parties[33]. Quantifying the risk for any such conjunction is typically accomplished through the use of the calculated probability of collision P_c at TCA. This calculation is generally performed with each received CDM, and such messages typically are received throughout the seven days leading up to TCA. The calculated P_c value is affected by the uncertainty in the positions of the space objects, an uncertainty that generally decreases as one approaches TCA. This decrease in uncertainty typically yields a particular kind of behavior in P_c values, which we shall refer to as the "canonical behavior". We seek to incorporate the shape of this canonical behavior into our understanding of the P_c values, with the goal of making

predictions about future P_c values, as well as making inferences about the location of the highest P_c value.

There has been considerable work in calculating the probability of collision, see for example Akella (2000)[4], Patera (2000)[60], Chan (2003)[20], or more recently Xu (2013)[82]. Though methods for improving the P_c are relatively well developed, far less work has focused on detecting trends in repeated measurements of the P_c. Notably, Carpenter and Markley have proposed various implementations of Wald's Sequential Probability Ratio Test (WSPRT) in deciding whether to accept the hypothesis that a new measurement on the P_c is identical in information content to the previous measurement[14][13][15]. Among the advantages of this method are its simplicity and its inherent modeling of false alarms and missed detections. While a considerable advance in P_c predictive methods, this approach is not without limitations. For instance, although the WSPRT tests consecutive measurements, it has no way of directly incorporating the times at which the measurements were taken; it considers measurement time only indirectly through the accumulation of data in forming the total information matrices from which it works. In general, P_c measurements are not taken at equidistant time intervals, suggesting a potential loss of information in the WSPRT approach.

In this manuscript, we propose a simple method to detect the trend in repeatedly-measured P_c values. Our approach has the advantage of directly incorporating the time between observations, which is allowed to be irregular. Additionally, we use the Bayesian paradigm in order to incorporate prior information gathered from past conjunctions. More sophisticated methods are certainly possible, but we wished to determine how much predictive power could be rendered by a simple and straightforward foundational approach.

1.2.1 "Dilution" of the P_c

It is clear that the probability of collision depends heavily on the size and shape of the combined covariance ellipsoid. Alfano[5] investigated this relationship for various miss distances, hard body volumes, covariance sizes and shapes. He reported that for a given miss distance, hard body volume, and covariance shape, there is a covariance size which maximizes the probability of collision, with the probability decreasing slowly if uncertainty is increased (that is, the size of the objects' covariances are increased) and decreasing very rapidly if this uncertainty is decreased. We seek to incorporate this known behavior into a statistical model in order to better calibrate each measured probability of collision. In practice, one typically observes a decrease in the size of the covariance as the event moves closer to TCA, producing what we will refer to as a "canonical behavior". We aim to try to recover this behavior beneath all the other "noise" of the problem and ultimately identify the point of maximum probability of collision, in order to make better judgments regarding the degree of continued monitoring that the conjunction merits.

1.2.2 Canonical Behavior

As noted above, changes in P_c generally follow a canonical behavior with respect to a decreasing state estimate uncertainty; and the the parameter used to illustrate this phenomenon is the ratio of covariance radius to miss distance (for the present we have used a spherical covariance for convenience, but this ratio can be generalized as the Mahalanobis distance and applied to the general case). Figure 1.8 depicts what we have called the "canonical behavior" of an event's P_c: an initial increasing change in order of magnitude in P_c as uncertainty decreases, followed by a subsequent drop off when the uncertainty becomes even smaller. The decrease in probability as uncertainty increases is what Alfano[5] referred to as "dilution in probability" because it was caused not by improvements in knowledge of satellite

positions but by a lack of positional knowledge that renders any conclusion of high risk more difficult. Note that P_c values are generally particularly small probabilities, and consequently one is usually concerned with changes in orders of magnitude. That is, one is interested in changes in $\log_{10} P_c$ as opposed to simply changes in the P_c value. In the following development, we let y denote the $\log_{10} P_c$ value.

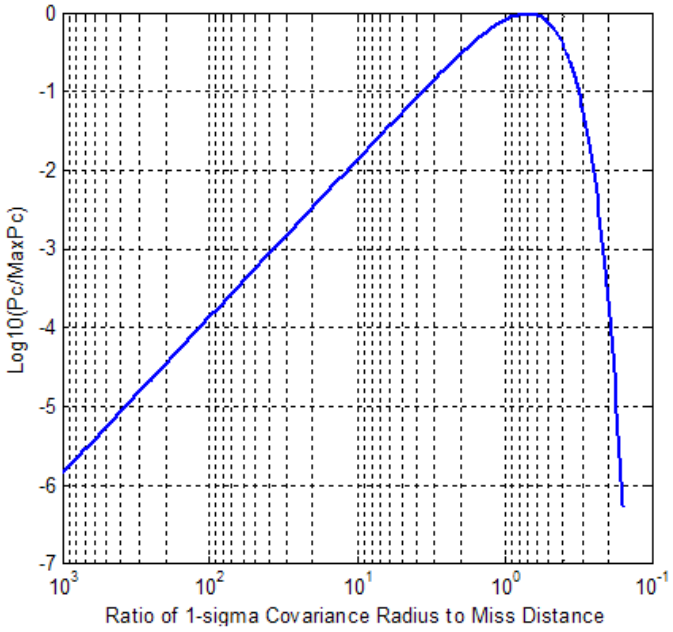

Figure 1.8. Theoretical canonical behavior (from Hejduk[39])

Though informative, using the ratio of covariance size to miss distance as a predictor variable is difficult in practice. Although the size of the combined covariances tend to shrink over time, the rate is not the same for each event. In some cases, the value of this ratio, which appears monotonic before and after the peak point in the figure above, actually increases and decreases several different times before reaching its final value, making modeling a trend even more difficult. Furthermore, the miss distance calculated on the initial CDM is subject to change on subsequent CDMs, and there is often no obvious trend in these updates. As a result, one never knows

what the next ratio value will be, even if one knows at which time a CDM would be received. Thus, to use this ratio as a predictor in a statistical model, one would need to regress the ratio on some quantity one could predict, such as time. This is especially difficult because the relationship between the ratio and the $\log_{10} P_c$ value is different for each event, as is the relationship between the ratio and time. To see the difficulty in this kind of modeling, let x_{kt} be the ratio of covariance radius to miss distance for the k^{th} event at time t. We assume that both y_{it} and x_{it} is measured with errors e_{it} and ϵ_{it}, respectively. Then this model is a state-space model[65] and can be written as

$$y_{it} = f(x_{it}) + e_{it}$$

$$x_{it} = g(t) + \epsilon_{it}$$

$$e,\, \epsilon \sim h(e, \epsilon|\alpha)$$

where e and ϵ are error terms with a joint distribution $h(\cdot|\alpha)$. It is clear that when attempting to calculate y via the ratio, in practice one needs to specify not only the relationship f between the ratio x_{it} and y_{it} but also specify the relationship g between x_{it} and t.

We leave the exploration of this kind of hierarchical model for later research. In this investigation, we use time as the predictor variable. We assume that the y values still follow a similar canonical behavior with respect to time as they do to with respect to the ratio. Note that this assumes that, as time nears TCA, the ratio decreases. We expect this kind of behavior, as one generally has more accurate information as one approaches TCA. This approach inherently encapsulates the interplay of how the covariances, miss distance, and positional approximations change over time. Because each of these exhibit a high variability across events, we seek to model to overall ensemble progression of these effects over time, as opposed to how each effect impacts the calculated P_c over time.

For simplicity, we attempt to model this behavior with a downward opening parabola, with the aim of correctly predicting the location and, less critically, the magnitude of the peak y value. Though model fit is important, our main goal is to correctly identify the peak y location and, secondarily, value, whether or not the other y values are predicted accurately. For this purpose, the parabola is the simplest curve that can provide a reasonable match to the behavior shown in the previous figure, given the particular attributes of interest here.

1.2.3 Bayesian Inference

Bayesian inference relies on the posterior distribution of the parameters. To see how the posterior distribution is calculated, let $\boldsymbol{\theta}$ be a vector of unknown parameters, defined on the parameter space Θ. Suppose one has data $\mathbf{y}$, with joint distribution $f(\mathbf{y}|\boldsymbol{\theta})$. Let $\pi(\boldsymbol{\theta})$ be a prior distribution on $\boldsymbol{\theta}$ with CDF $P_{\boldsymbol{\theta}}$. Treated as a function of θ for fixed $\mathbf{y}$, the joint distribution becomes the likelihood, $l(\boldsymbol{\theta}|\mathbf{y})$, defined on Θ. The posterior distribution of $\boldsymbol{\theta}$, given by Bayes' theorem, is

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{l(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})}{\int l(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

This is the distribution of the parameters after having seen the data vector $\mathbf{y}$. Thus, the prior beliefs about parameters and their distributions are updated after encountering the actual data. The posterior distribution often does not have a closed form and must be approximated using numerical methods such as Markov Chain Monte Carlo (MCMC).

1.2.3.1 Inference for a new data point Suppose one has data $\mathbf{y}$ and one wishes to predict $y^* = \log P_c$ at a new data point at time t^*. One can make an inference on y^* by using the predictive distribution

$$g(y^*|x^*, \mathbf{y}) = \int_\Theta g(y^*|\theta, x^*, \mathbf{y})\pi(\theta|\mathbf{y})d\theta,$$

which can be estimated by using the posterior samples from the MCMC draws. In practice, we usually do not know t^* in advance, as it is determined by the exigencies of any particular event. However, for simulation purposes, we use the next time point at which a CDM was received and make a prediction. We construct a 95% credible interval for y^* and check to see if the actual value of y is contained in the interval. The percentage of credible intervals which contain the true y value is known as coverage. If the coverage is close to the nominal value of 95%, we can assume that these predictions are reliable.

CHAPTER TWO

Methods & Results for Trending in Probabilities of Collision

2.1 Methods

In this section we present the statistical methodology underlying our models for capturing the trend of $\log P_c$ value over time.

2.1.1 Last Observation Carried Forward

A common technique for imputing missing data in a longitudinal data is Last Observation Carried Forward (LOCF)[63]. In this approach, one replaces any missing value for a given subject with their last observed value. Many have offered criticisms of this approach. For instance, Saha[69] argues that this method induces bias under informative dropout, Kenward[48] lambastes the method, and argues that it is only appropriate under unrealistic special cases. Nevertheless, this method is widely used in missing data, as it is simple and intuitive.

In conjunction risk assessment, LOCF more or less represents the current practice for interpreting P_c values. For instance, the most recent P_c value is generally taken to be the most reliable measurement, and decisions are currently based on these values. In effect, operators "predict" all future P_c values to be the last observed P_c value, though they certainly expect some variability in future values. One of our objectives is to quantify how much variability operators can expect in future values. In addition, we consider whether the last P_c value really is the "best" prediction, or if other information can be used to improve this prediction. Underlying these objectives are the empirical findings by operators that some events exhibit more variability in P_c than others, and that P_c values seem to follow a general trend over time.

2.1.2 Look-Up Tables

Look-up tables have been in wide use in statistics since at least 1903, due to Sheppard's definitive tables for the standard normal cdf and pdf[71]. These tables have historically been used to avoid having to repeatedly calculate difficult quantities, such as $\int_0^x \exp(-t^2)dt$, which is involved in the normal cdf and pdf[24]. These look-up tables serve as an important example for our problem, as they are in wide use due to the popularity of the normal distribution as a modeling distribution. In the problem of P_c trending, we consider the distribution $log_{10}P_c$ by time to TCA. Similar to the normal distribution, this distribution is to be referred to repeatedly for inference, so that the idea of a look-up table might be useful in practice.

Creating a look-up table for the distribution $log_{10}P_c$ by time to TCA involves estimating the conditional distribution of $y = log_{10}P_c$ at a given time t to TCA. One of the first important approaches to this problem was given by Stone[76], who suggested neighbor-type estimates. To illustrate ideas, let $(X_1, Y_1), ..., (X_n, Y_n)$ be a random sample from the joint distribution of (X, Y). Stone suggested estimates of $F(y|x)$ of the form

$$\hat{F}(y|x) = n^{-1} \sum_{i=1}^{n} W_i(x) I(Y_i \leq y), \quad -\infty < y < \infty \tag{2.1}$$

where $W_i(x) = W_i(x; X_1, ..., X_n)$ weights more heavily Y-values for which X_i is closer to x. Stute[77] developed asymptotic properties for estimators of this type based on kernel weights.

Perrachi[61] notes that, in estimating the distribution of a random variable, one can choose to estimate either the conditional quantile function or conditional distribution function. For simplicity, suppose that Z is a random variable with an absolutely continuous distribution with strictly positive density. Then the cumulative distribution function (CDF) is defined on R by $F(z) = P(Z \leq z)$, while the quantile function is defined on $(0, 1)$ as $Q(u) = \{z \in R : F(z) = u\}$. Thus, Q and

F are inverses of each other, so that $Q(F(z)) = z$ and $F(Q(u)) = u$. As a result, either function could be used to estimate the distribution of a random variable.

These estimators can be easily extended to conditional distributions. Suppose instead one observes a random vector (X, Y), where X is k-dimensional and Y is a real-valued continuous random variable with strictly positive density, as before. One may characterize the conditional probability distribution through the conditional distribution function $F(y|x) = P_x(Y \leq y)$, or through the conditional quantile function $Q(u|x) = \{y \in R : F(y|x) = u\}$.

Characterizing a conditional probability distribution through the quantile function has a few notable difficulties. To explore these, note that $Q(u)$ may be characterized as the unique solution to the problem

$$\min_{z \in R} E l_u(Z - z) \tag{2.2}$$

where l_u denotes the asymmetric loss function

$$l_u(v) = [u - I(v < 0)]v. \tag{2.3}$$

To describe a conditional distribution $f(y|x)$, one typically employs *quantile regression*. Quantile regression seeks to estimate the parameters of a function $g(\cdot)$, which is the unique solution to the problem

$$\min_{g \in G} E l_u(Y_i - g(X_i)), \quad 0 < u < 1 \tag{2.4}$$

where G is the class of real-valued functions defined on R^k. In practice, $g(\cdot)$ is often taken to be linear, so that $Q(u|x) = x^T \beta(u)$, as described in Koenker and Bassett[52]. Then the estimate of the k-dimensional parameter $\beta(u)$ is any solution to the problem

$$\min_{b \in R^k} n^{-1} \sum_{i=1}^{n} l_u(Y_i - X_i^T b), \quad 0 < u < 1. \tag{2.5}$$

The resulting estimate $\hat{\beta}(u)$ can be used to define $\hat{Q}(u|x)$, which in turn can be used to estimate the conditional distribution function

$$\hat{F}(y|x) = \sup\left\{u \in (0,1) : \hat{Q}(u|x) \leq y\right\}. \tag{2.6}$$

Perrachi shows that if the conditional distribution of Y depends on x through both a linear location parameter $\mu(x) = \alpha + x\beta$ and a scale parameter $\sigma(x) > 0$, then the conditional quantiles are no longer linear in x. That is, if the data Y exhibit heterogeneity of variance across varying levels of x, then the quantiles are not necessarily linear. Thus, linear quantile regression may yield poor estimates in this case. Furthermore, this non-constant variance may produce estimates of linear models for conditional quantiles which cross each other, violating a basic assumption about quantiles. Some nonparametric estimators have been proposed, based on kernel or nearest neighbor methods (Antoch and Janssen[6]; Samanta[70]; Truong[78]; Bhattacharya and Gangopadhayay[8]; Chaudhuri[21]), regression splines with a fixed number of knots (Hendricks and Koenker[40]), smoothing splines (Koenker et al[52]) and penalized likelihood.

A difficulty particular to quantile regression methods, mentioned above, is the so-called "no-crossing" condition. This is the condition that, for all values of a covariate x, one should have $\hat{Q}(u_1|x) \leq \hat{Q}(u_2|x)$ when $u_1 < u_2$. That is, the regression line of a lower quantile should not up-cross the regression line of an upper quantile. To see why this must be the case, note that when $u_1 < u_2$, the solutions z_1 and z_2 which satisfy

$$F(z_1) = u_1$$

$$F(z_2) = u_2,$$

must also satisfy

$$z_1 < z_2,$$

since the CDF is a monotonically increasing function. Thus, we must have $Q(u_1) < Q(u_2)$ and $Q(u_1|x) < Q(u_2|x)$ by extension. Thus, we seek estimators of $Q(\cdot|x)$ which adhere to this constraint. Koenker[51] avoided this issue by considering parallel quantile functions.

There have been some solutions proposed for the "no-crossing" condition, notably He[38], Wu and Liu[81], Neocleous and Portnoy[58]. Bondell[10] notes that several authors have proposed to first estimate the conditional cumulative distribution function via local weighting, and then invert it to obtain the quantile curve. He notes that this method is suitable for estimation of the conditional quantile, but that it is not suited for estimation of linear predictor effects. As our concern is only estimation of conditional quantiles, this is not a limitation for our application. Because our data exhibits non-linearity and heterogeneity with respect to the covariates, direct estimation of the quantile function is difficult. We proceed with methods based on the empirical CDF, which circumvents these issues.

2.1.3 Constrained Bayesian Inference

Gelfand[35] introduced MCMC methods for constrained parameter problems. He notes that if one has the full conditionals for each parameter in the model, producing MCMC draws adhering to the constraint simply involves modifying the full conditional density. For instance, suppose one has data $\mathbf{y}$ which has density $f(\mathbf{y}|\theta)$, where θ is a k-dimensional vector constrained to lie in a subset $S^k_{\mathbf{Y}}$ of R^k. Furthermore, in Bayesian models, we specify a prior distribution for θ, say $p(\theta|\lambda)$. Then, Gelfand shows that the posterior distribution of any element of θ is

$$f(\theta_i|\mathbf{Y}, \lambda, \theta_j, j \neq i) \propto f(\mathbf{Y}|\theta)p(\theta|\lambda), \quad \theta_i \in S^k_i(\theta_j, j \neq i),$$

so that the posterior distribution follows its usual form, only with the specified constraints.

As Gelfand notes, there are a few simple ways to simulate draws from this distribution. One way is to generate the full distribution, not accounting for the constraints, and then to only keep the variates which satisfy the constraints. This can also be accomplished by simulating draws U from a uniform(0,1) distribution and following $\theta_i = F_i^{-1}[F_i(a) + U(F_i(b) - F_i(a))]$ where F_i is the full conditional CDF of θ_i. This produces a draw of θ_i which adheres to the constraints, and is due to Devroye[25].

2.1.4 Bayesian Beta Regression Models

Generalized linear models were introduced by Nelder[57] in 1972 to solve the problem of regression for responses which have a non-normal distribution. These models often are used for binary and count data, so that the usual models involve the Binomial, Poisson, Negative-Binomial, or Multinomial distributions. Of course, this general class of models includes those which handle continuous data, such as Normally-distributed responses (the usual linear regression case), and Weibull-distributed responses (common in survival analysis). Notice that none of the models handle responses with bounded support. To handle such responses, methods for Beta regression were introduced by Paolino[59], Kieschnick and McCullough[50], and Ferrari and Cribari-Neto[31].

Though Beta regression seems to be the most popular regression method for bounded responses, it should be noted that other methods exist. For instance, one may choose to transform the data using the logit function and use linear regression on the transformed responses. Let $\mathbf{y}$ be a vector of bounded responses such that $y_i \in (a, b)$. Furthermore, let $\mathbf{X}$ be an $n \times p$ design matrix containing the corresponding covariates. Then one may choose to model

$$\log\left(\frac{\boldsymbol{y}}{1-\boldsymbol{y}}\right) = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}),$$

so that one may proceed to apply usual linear regression techniques. Define $z_i = \log(y_i/(1-y_i))$, so that one may write the response vector as $\boldsymbol{z}$. Suppose y_i is a probability, so that z_i can be interpreted to be the log odds of an event. Thus, this model has the advantage of being able to utilize known techniques in linear regression, but has the disadvantage of forcing one to interpret the results in relation to the response on a different scale.

Other techniques are possible, such as the fractional logistic model, various nonlinear models, and models based on other bounded distributions (e.g. the simplex distribution). Beta regression is attractive because it builds upon the GLM framework and it is based on a familiar distribution. Though these other models are possible, we proceed with Beta regression because it has a recently developed framework for mixed and mixture models, and because it can be fit in a rather straightforward manner with Bayesian techniques. Thus, it is sufficiently flexible and is able to incorporate prior information.

The primary aim of GLMs is to model some transformation of the expected value of the response variable with a linear model. This indirectly specifies a relationship for the mean, and one often treats parameters related to variability, such as the so-called "precision parameter", as a nuisance parameter. This parameter can also be modeled, and such models were explored in the context of Beta regression by Cepeda[18], Cepeda and Gamerman[17], and Simas et. al[72].

The aforementioned developments all extended methods of inference for independently and identically distributed Beta variables. Our data consists of observations which are likely dependent, as they are longitudinal in nature. Furthermore, we are interested in making predictions about individual events, such as the $\log_{10} P_c$ behavior for an event as it nears TCA. When one has longitudinal data and the goal of inference is prediction of individual responses, the most common technique is the mixed model. Mixed models were popularized for longitudinal data by Laird and

Ware[53], and a general approach for fitting GLMMs was introduced in Stiratelli, Laird, and Ware[75]. As the Beta GLM was late to be developed, mixed models are fairly new to this setting. For instance, one of the first introductions to Beta mixed models was provided by Verkuilen and Smithson[79]. Bayesian implementations of the Beta mixed model soon followed, described in Figeuroa-Zuniga et. al[32], and Bonat et. al[9].

2.1.5 Functional Data Analysis

Functional Data Analysis is an extension of longitudinal data analysis, a field which attempts to explain the effect of time on various subjects with multiple measurements. For example, consider a study which follows the blood pressure of a number of individuals over time. Each subject's blood pressure will change depending on each individual's genetic predisposition, thus implying that the analyst must account for effect of the subject as well as time. The case of trending P_c over time is similar in that one observes multiple measurements of P_c for a single event, implying correlation among measurements within the event. Additionally, each event seems to have a slightly different effect on the trend of P_c values over time, reinforcing the need for a longitudinal model. In previous sections, we have introduced parametric models, one of which (downward-opening parabola) is of the form

$$y \sim F(\mu, \phi)$$
$$g^{-1}(\mu) = \boldsymbol{X}\boldsymbol{\beta},$$

where $F()$ is the CDF of a distribution from the exponential family. However, we noted that this parametric model may be too restrictive for the overall trend of P_c values, or a within-event trend in P_c values. Instead, we might consider modeling a trend non-parametrically by specifying a general unknown function $f(x)$:

$$y = f(x) + \epsilon.$$

Specifically, let y_{ij} represent the $\log P_c$ value of the i^{th} event from the j^{th} CDM of that event. Let $X_i(T_{ij})$ be a random function for the i^{th} event at the j^{th} CDM at time T_{ij}. The model we consider is

$$y_{ij} = X_i(T_{ij}) + \epsilon_{ij} = \mu(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ij}\phi_k(T_{ij}) + \epsilon_{ij},$$

where ϕ_k are eigenfunctions and ξ_{ij} are uncorrelated random variables with zero mean and variances λ_k, the eigenvalues corresponding to the eigenfunctions.[83] Utilizing eigenfunctions to express an overall function is often called functional principal component analysis. This method allows us to draw inferences about the overall trend and the within-event trend based on only a few observations. This, in fact, was the reason for our choice in pursuing a functional data approach to the problem. It is well known that longitudinal data analysis models tend to need many observations for each subject in order to estimate their numerous parameters. In contrast, the current method estimates only a few parameters. In practice, the sum of eigenvectors is limited to the first few, those which explain the majority of the variability in the model. The need for only a few parameters and a high level of functional variability makes the functional data analysis ideal for the trending of P_c values.

2.1.5.1 Principal Component Analysis. The model above is a functional extension of a standard dimension-reduction technique called principal component analysis. Consider a set of n observations of a p-dimensional random variable, i.e. $\mathbf{x_i} = \{x_{i1}, x_{i2}, ..., x_{ip}\}$ for $i = 1, ..., n$. The basic idea behind principal components is to explain the variance-covariance structure for a large number of variables (in this case, the dimensions of $\mathbf{x_i}$) through a few linear combinations of these original variables. This method allows for dimension reduction and for interpretation. Let $\vec{x_i}$ denote the n observations of the i^{th} dimension. Then the idea is to find a new set

of vectors $\vec{y_1}, \vec{y_2}, ..., \vec{y_p}$ where

$$\vec{y_i} = \sum_{j=1}^{p} l_i j \vec{x_j},$$

where $var(\vec{y_i}) = \vec{l_i'} \Sigma \vec{l_i}$ and for $cov(\vec{y_i}, \vec{y_k}) = \vec{l_i'} \Sigma \vec{l_k} = 0$ and $var(\vec{y_1}) \geq var(\vec{y_2}) \geq ... \geq var(\vec{y_p})$ for $\vec{l_i'} = (l_{1i}, l_{2i}, ..., l_{pi})$. This problem has the following solution:

(1) Suppose that the matrix Σ has associated real eigenvalue-eigenvectors given by $(\lambda_i, \vec{e_i})$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, then the i^{th} principal component is given by

$$\vec{y_i} = \vec{e_i}^i X = e_{i1}\vec{x_1} + e_{i2}\vec{x_2} + ... + e_{ip}\vec{x_p},$$

and $var(\vec{y_i}) = \lambda_i$ for $i = 1, 2, ..., p$, $cov(\vec{y_i}, \vec{y_k}) = \vec{e_i}' \Sigma \vec{e_k} = 0$ for $i \neq j$. Note, the eigenvalues λ_i are unique, however, the eigenvectors (and hence the vectors $\vec{y_i}$) are not.

(2) The total variance for the p dimensions is $tr[\Sigma] = \sum_{i=1}^{p} \lambda_i$. Hence, the proportion of variance explained by the k^{th} principle component is $\lambda_k / \sum_{i=1}^{p} \lambda_i$.

(3) If the matrix X is centered and scaled so that Σ is the correlation matrix, then $\sum_{i=1}^{p} \lambda_i = p$.

In practice, we choose a suitable "cutoff" for what percentage of the variance we want the new bases $\vec{y_i}$ to explain (say, 95%). Then we choose to use these $q < p$ new bases to re-express the original data in a smaller dimension. In functional data analysis, we replace the vector observations with functional observations.

2.1.5.2 Functional Data for Sparse Longitudinal Data. As mentioned earlier, the data we consider is sparse longitudinal data. We have many events, but each event has only a few $\log_{10} P_c$ values. This sparseness presents difficulty in estimating

functional data models, as one must estimate the functional principal component scores $\xi_{ik} = \int (X_i(t) - \mu(t))\phi_k(t)dt$, which are usually estimated via numerical integration. When the data is sufficiently sparse, a common estimate for this parameter is $\hat{\xi}_{ik}^S = \sum_{j=1}^{N_i}(Y_{ij} - \hat{\mu}(T_{ij}))\hat{\phi}_k(T_{ij} - T_{i,j-1})$, setting $T_{i0} = 0$. As noted by Yao[83], this estimator will not yield reasonable approximations to ξ_{ik} when the data is sparse.

Yao[83] overcomes this difficulty in estimating ξ_{ik} by assuming that ξ_{ik} and ϵ_{ij} are jointly Gaussian. As a result, the expectation of ξ_{ik} is tractable, and is given by

$$\tilde{\xi}_{ik} = E(\xi_{ik}|\tilde{\boldsymbol{Y}}_i) = \lambda_k \boldsymbol{\phi}_{ik}^T \boldsymbol{\Sigma}_{Y_i}^{-1}(\tilde{\boldsymbol{Y}}_i - \boldsymbol{\mu}_i),$$

where $\boldsymbol{\Sigma}_{Y_i}^{-1} = cov(\tilde{\boldsymbol{Y}}_i, \tilde{\boldsymbol{Y}}_i)$. Yao proceeds with inference using this expected value in lieu of the less reliable estimators based on sums. This procedure, which he calls Principal Component Analysis through Conditional Expectation (PACE), is now a common method for handling sparse functional data.

2.2 P_c Trending

In this section we present the models we propose for assessing the trend in $\log P_c$ values over time.

2.2.1 Last Observation Carried Forward

We briefly describe our implementation for LOCF. In our simulations, we predict future $\log_{10} P_c$ values with the previously observed value for that event. Thus, for the j^{th} OCM of the i^{th} event, we predict

$$\hat{y}_{i(j+1)} = y_{ij},$$

which ignores all other past values, as well as the time to prediction. We construct prediction intervals for this method using Repeated Cross-Validation, in the same way as discussed below for the Look-Up Method.

2.2.2 The Look-Up Method

We next consider a simple extension of the LOCF method, called the Look-Up method. This method utilizes the previous quantile rather than using the previous $\log P_c$ value directly.

2.2.2.1 Intuition. We propose a simple method as basis of comparison for our more sophisticated models, which we call the Look-Up Method. The Look-Up Method is based on the common expectation that, when an event is observed with relatively high P_c values, we can suppose this event to behave similarly to other events with other similarly high values. In order to formalize this intuitive approach into an explicit model, we need to establish how high "relatively high" is. A natural way to quantify this notion is in terms of quantiles. That is, we expect events with $\log P_c$ values in the qth quantile to behave similarly to other events with $\log P_c$ values in the qth quantile. The method we describe below is similar to methods involving "look-up tables," where quantiles for various scenarios are used to find the probability of an event within the table.

2.2.2.2 Method. Let x and y be the time and P_c value from the most recent observation. Furthermore, let x^{new} and y^{new} represent the time of prediction and the true P_c value at this time. The algorithm for the Look-Up Method is as follows

We find that $w = 2$ days to be a reasonable window length. This length depends on how much prior data is available, as w may need to be smaller for large datasets, allowing for more precise estimation of the CDF. Note that this method only predicts an estimate of y^{new} and does *not* by default generate a prediction interval or any other confidence information.

The method above is simple: find the sample quantile of the observed P_c value at the given time, and assume that future P_c values will be at the exact same

Algorithm 1 Look-Up Method

procedure LOOK-UP
    Choose an historical data set $\mathbf{Y}_h$ such that the events contained in $\mathbf{Y}$ are believed to behave similarly to the event of interest.
    Choose a window w.
    Calculate the empirical CDF $\hat{F}(y)$ of the $\log P_c$ values in the interval $(x - w, x + w)$.
    Calculate the sample quantile $\hat{q}$ of y
    Calculate the empirical CDF $\hat{F}(y)$ of the $\log P_c$ values in the interval $(x^{new} - w, x^{new} + w)$.
    Predict y^{new} to be $\hat{q}(x^{new})$
end procedure

quantile. As the model is simple, it also discards potentially useful information. For instance, the predictions are made based only on the sample quantile of the most recent observation, and makes no use of previous observations other than, of course, the historical P_c behavior information. However, one could argue that the most recent observation is the most (or only) meaningful observation, and thus one should make inferences based on this value rather than more immediate past values.

2.2.2.3 Prediction Intervals. As noted above, the Look-Up Method does not automatically generate prediction intervals; this is a consequence of the method making no distributional assumptions. However, one may still construct prediction intervals via bootstrapping or cross-validation[74]. Recently, these methods were compared[11], and the results from this comparison indicate that estimators based on Repeated Cross Validation (RCV) tend to outperform other estimators (e.g. bootstrap estimators). As a result, we implement RCV to generate prediction intervals. The method was initially proposed by Burman (1989)[12], which describes the algorithm in detail.

We use RCV to estimate the distribution of prediction errors. This will allow us to construct prediction intervals at any confidence level for the Look-Up method.

Our procedure is as follows.

Algorithm 2 RCV algorithm

procedure RCV
  for Each repetition $r \in nRep$ **do**
    for Each fold $i \in nFolds$ **do**
      for Each event $j \in nEvents$ **do**
        for Each OCM $k \in nOcms - 1$ **do**
          Predict y_{k+1} using Look-Up Method
          Estimate prediction error $e_{k+1} = y_{k+1} - \hat{y}_{k+1}$
        end for
        Collect prediction errors across OCMS $pVecj = e_2, ..., e_{nOcms}$
      end for
      Collect prediction errors across events $pStore = (pVec_1, ..., pVec_{nEvents})$
      Calculate estimated percentiles $\hat{p}_i$ using $pStore$ for $i = 1, ..., 99$
    end for
    Calculate mean of estimated percentiles
  end for
  Calculate mean of estimated percentiles
  Return estimated percentiles
end procedure

Notice that for each event, predictions are made for all but the first CDM using the previous CDMs. Recall that CDMs are received at varying intervals, so that the procedure above results in making predictions at varying intervals into the future. Thus, the procedure implicitly assumes that the distribution of prediction errors does not depend on time to prediction. While this assumption is generally untenable, the data is such that the time between consecutive CDMs is generally 2 days or less. As one is generally concerned with making predictions no sooner than one day into the future, these prediction errors are conservative for their operational use, meaning that they predict a more worrisome P_c than the actual value. This is because predictions are generally more variable as time to prediction increases. Thus, this procedure results in prediction intervals which hold reasonably for all

prediction times of 2 days or less. The procedure is simple and generally conservative. Future research may focus on different bootstrapping procedures which improve the accuracy of these errors.

2.2.3 Vertex Model

In order to compute this predicted P_c inverted parabola, we use constrained optimization[54] to enforce a downward-opening behavior. There is also precedent for constrained inference in the Bayesian paradigm, as Gelfand[35] introduced an approach to Gibbs sampling in constrained parameter and truncated data problems. Specifically, Gelfand considers problems with ordered parameters, constrained parameters, and censored data. Considering the general equation for a parabola below, our problem is seen as one involving constrained parameters, as we know that $\beta_2 < 0$ and (as discussed subsequently) $\beta_0 < 0$.

$$y = \beta_0 + \beta_1 t + \beta_2 t^2$$

We also show that this induces a constraint on β_1. Implementing these constraints is another way in which we can "inform" the model. Utilizing these constraints along with an informative prior structure allows us to include a maximal amount of prior information, which we believe to be essential, as many of the events we consider contain only 3 or 4 data points, and we wish a reasonable prediction as early as possible within the event.

To allow our model to incorporate prior information from past events, we use the Bayesian paradigm[36]. Let y_{ij} be the $\log_{10} P_c$ from the j^{th} CDM from the i^{th} event. Similarly, let t_{ij} be the time (in days) until TCA for the j^{th} CDM from the i^{th} event. We assume that the observed P_c values over t follow the relationship

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$. Furthermore, we assume this parabola will be downward opening, to attempt to model the expected canonical behavior. Utilizing the Bayesian paradigm will allow us to incorporate information about where the peak y value usually is, and how quickly the y values tend to drop off. This incorporation is accomplished by specifying informative prior distributions for the parameters, which are considered random variables in the Bayesian paradigm. Another consequence of treating parameters as random variables is that one can make probabilistic statements about functions of parameters, a Bayesian feature that is not fully possible with a frequentist approach. Thus, one can make statements about the probability of the peak P_c value occurring at a particular time and magnitude. Finally, using the Bayesian paradigm allows us to make predictions in this four-parameter model even with only two or three observations by utilizing the prior distributions of the parameters to help identify the likely values of the parameters.

2.2.3.1 Prior Structure. The usual prior structure for regression coefficients in linear regression is an independent normal prior for each regression coefficient[36]. We amend this structure to incorporate the constraints we know to exist in our problem. We know that the parabola must open downwards, so that $\beta_2 < 0$. As a consequence of this constraint and the fact that all y values are less than or equal to 0 by definition (since they represent the base 10 logarithm of values between 0 and 1), we also know that $\beta_0 < 0$.

We show that the parameter β_1 must also be constrained. Because $y_{ij} \leq 0$ for all i, j, it follows that the peak y value should also be less than or equal to zero. It is easy to show that the location of the peak is $h = -\beta_1/2\beta_2$, and that the magnitude of the peak is $b = \beta_0 - \beta_1^2/4\beta_2$. In order to force the magnitude of the peak b to be less than or equal zero, we must have

$$\beta_0 - \beta_1^2/4\beta_2 \leq 0$$

$$4\beta_2\beta_0 - \beta_1^2 \geq 0$$

$$\beta_1^2 \leq 4\beta_2\beta_0,$$

where the second line follows since $\beta_2 < 0$. This implies that $\beta_1 \in [-2\sqrt{\beta_0\beta_2}, 2\sqrt{\beta_0\beta_2}]$. Implementing these constraints in conjunction with the usual prior structure, we have

$$\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \epsilon_{ij} \\
\epsilon_{ij} &\sim N(0, \sigma_i^2) \\
\beta_0 &\sim Normal(\mu_0, \sigma_0^2) I_{(-\infty,0)} \\
\beta_1 &\sim Normal(\mu_1, \sigma_1^2) I_{(-2\sqrt{\beta_0\beta_2}, 2\sqrt{\beta_0\beta_2})} \\
\beta_2 &\sim Normal(\mu_2, \sigma_2^2) I_{(-\infty,0)} \\
\sigma^2 &\sim InverseGamma(a, b),
\end{aligned}$$

where $I()$ is the indicator function. Thus, we fit a downward opening parabola to the $\log P_c$ values over time for each event. This implies that each event has $\log P_c$ values which will rise and fall over time and that each event is allowed to have its own rate of increase/decrease. Eliciting informative priors on the regression coefficients will allow us to borrow information about what the shape of this parabola is for most events, and how much it is prone to vary. Though there are other ways to borrow information, *e.g.* a mixed model, we find this to be a simple and straightforward way to allow the model to be flexible enough to fit all of the events. On a more technical note, attempting a mixed model in this setting is not particularly straightforward, as any random effects specified in the model would also have to be constrained. Furthermore, at least two random effects would be necessary (a random intercept and a random slope), as we desire a model which can have a different peak location and value for each event. Ultimately, we favor a more simple model that is interpretable and flexible.

We note a few additional attributes of this model here. First, although we know that the regression coefficients are necessarily correlated, we choose not to incorporate this correlation in our prior structure, principally because the prior for β_1 depends on other regression coefficients. Although estimates may be slightly more efficient by including more information, we believe that independent priors are sufficient in this case. Additionally, it is worth noting that because the regression coefficients are defined on half the real line (β_0 and β_1) and a closed interval (β_1), other prior distributions could be chosen. For instance, the Gamma distribution is defined on $(0, \infty)$, so theoretically it could be used as a prior distribution for $-\beta_0$ or $-\beta_2$. Similarly, the Beta distribution could be considered for β_1. However, our testing of these priors showed problems with their use. The sampling generally exhibited a high amount of autocorrelation and/or slow convergence, which is not the case with the truncated normal distributions.

Because P_c values can assume very small values, including the value of 0 to machine precision, using these data in an unbounded way introduces a very large dynamic range in the observed values. Operationally, there is little interest in events with a P_c below 1E-07 and essentially none with a P_c below 1E-10; so it is quite reasonable to truncate (left-censor) the dataset by resetting the values of P_c data $<$ 1E-10 to the 1E-10 value. Of course, in such a case one must accept the cognitive dissonance of the model predicting P_c values less than 1E-10. However, this is acceptable to us for a few reasons. One reason is that we are mainly concerned with the time point at which the peak y value occurs and, to a lesser degree, its predicted value. The other reason is more practical: we are not particularly concerned with prediction for smaller values of y. Because the y values represent orders of magnitude, we are far less worried about prediction error for small values of y than we are for large values of y. Experts generally agree on the threshold of 1E-10 as representing a probability of "essentially zero", though because collisions are

rare, this threshold is impossible to justify with empirical evidence. Nonetheless, we proceed with this threshold, and note that others may reasonably explore different thresholds with these models.

Lastly, we admit that our model cannot capture the rare occurrence that the y values initially decrease and then increase, i.e. an upward opening parabola. We do not concern ourselves with this case, as in such a case our model would fit essentially a horizontal line, indicating no discernible peak value. Though the shape of the data is not preserved, our end goal is: we seek significant statistical evidence of the size and location of the peak, and in this situation its size and location are unclear.

2.2.3.2 Inference for the Peak Value. The supposed canonical behavior suggests that the order of magnitude of the P_c value increases as the uncertainty decreases and drops off after a certain point. In general, uncertainty tends to decrease with time. Thus, we expect that this relationship holds with reference to time as well. Though some events exhibit this behavior, many events only exhibit the decline in order of magnitude of the P_c value. That is, if we believe the $\log_{10} P_c$ truly increases in time initially, this increase is censored within many events–the earlier small $\log_{10} P_c$ values lie outside of the 7-day screening window or outside of the physical screening volume and were thus not reported. Similarly, because we are only able to observe a few $\log_{10} P_c$ values, we are unlikely to observe the true peak. Thus, it is difficult to measure the accuracy of any prediction of the peak we might make. Because we are not certain of being able to observe the true peak, we take the highest observed $\log_{10} P_c$ value to be the peak.

We can infer the distribution of the location of the peak by utilizing the well-known identity that the peak is located at $x_{max} = -\beta_1/2\beta_2$. We estimate this distribution by collecting the posterior samples of β_1 and β_2 from the MCMC output and transforming them as x_{max} is defined. From the empirical distribution of x_{max},

we can compute a point estimate and a 95% credible interval for x_{max}. For the point estimate, we utilize the posterior mode of x_{max}, which is found by fitting a kernel density to the samples of x_{max} and finding the most likely value. For the credible set, since we define time as time until TCA, we are mainly concerned with the lower bound. Here, the lower bound represents, with 95% probability, the latest time at which our model predicts a peak will occur. This is operationally useful, as we are often interested in whether the peak will occur before 48 hours until TCA. Thus, if we can say that the peak will occur before this time with 95% probability, then the operator may be able to use this information to make a more informed decision regarding the importance of continuing to follow the event. Similarly, we can construct bounds for the magnitude of the peak. The distribution for the magnitude of the peak y_{max} is computed in the same way as for the location but instead using the transformation $y_{max} = \beta_0 - \beta_1^2/4\beta_2$.

2.2.4 The Bayesian Beta Regression Model

As discussed above, Beta regression is natural for bounded random variables. Below, we discuss various types of Bayesian Beta regression models used in modeling $\log_{10} P_c$ values.

2.2.4.1 The Distribution of $\log_{10} P_c$ *Values.* When modeling the trend in P_c values, one is generally concerned with changes in order of magnitude, thus one generally models $\log_{10} P_c$ as opposed to the observed P_c values. This poses an interesting statistical question, namely the distribution of $\log_{10} P_c$ values. Distribution selection is more obvious for the P_c values, as they are bounded between 0 and 1; thus a statistical modeler generally chooses a beta distribution to model these values (although there are a few other less commonly used distributions, such as the simplex distribution, that could be deployed). Theoretically, there is no lower bound on

$\log_{10} P_c$ values, as P_c values can be arbitrarily close to zero. Operationally, however, one often considers P_c values below 1E-10 to be effectively 0. To account for this, in a previous section we "floored" the $\log_{10} P_c$ values at -10, so that the large number of small $\log_{10} P_c$ values did not overly influence the model. This allows one to focus inference on the operationally relevant $\log_{10} P_c$ values, which tend to be around -5 and greater. We follow suit here, flooring all $\log_{10} P_c$ values at -10. Therefore, even in modeling the $\log_{10} P_c$ values, we have bounded data (between -10 and 0) that yield a mixture of discrete and continuous outcomes. In this case, the data are -10-inflated, but when the variable is rescaled to fit the Beta distribution, the data are zero-inflated. This can be accommodated by the zero-inflated beta distribution; the form of the equation is given below, with the specific symbology explained in the subsequent sections:

$$f(y|\mu,\phi,p) = (1-p)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} I_{(0,1)}(y) + pI_{[0]}(y). \quad (2.7)$$

Here, $I_A(\cdot)$ is the indicator function, so that the first term corresponds to values falling between 0 and 1 (or $\log P_c$ values falling between -10 and 0) and the second term corresponds to values equal to 0 (or $\log P_c$ values equal to -10). The parameter μ is the mean of the Beta distribution, which will be modeled in the GLM framework, and the parameter ϕ is the corresponding dispersion parameter, which is a measure of variability. The parameter p can be interpreted to be the probability that one observes a 0 (or a $\log P_c$ of -10). To our knowledge, no one has investigated a joint model for an Bayesian inflated Beta regression model.

As noted previously, the P_c values for each event tend ultimately to decrease with time but at a different rate in each conjunction. This suggests approaching the problem within a mixed model framework, allowing random terms for each conjunction. This is a natural approach to take, as the data are longitudinal in nature: one observes an overall trend in time, yet each subject (in this case, each conjunction) deviates somewhat from this trend, and observations within a subject are correlated

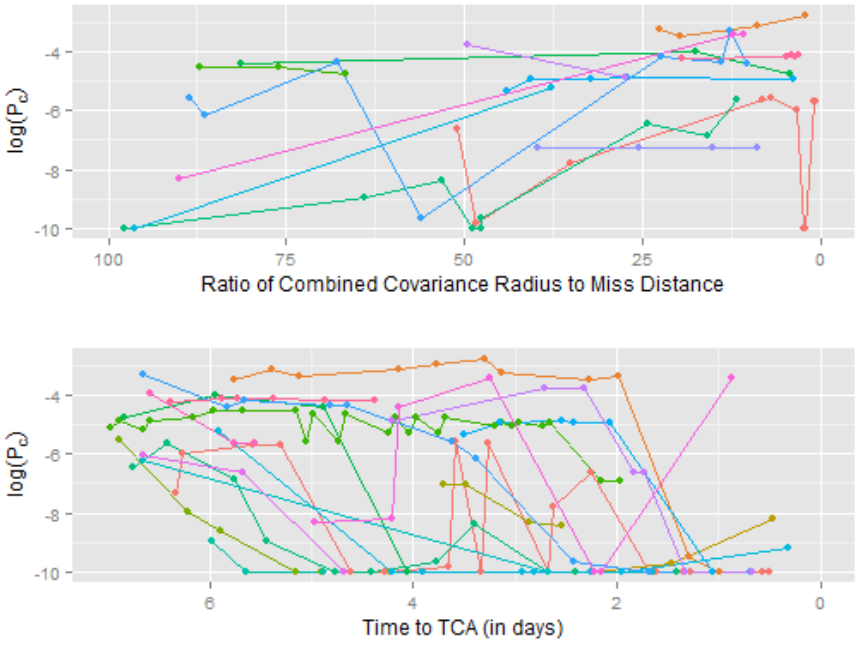

Figure 2.1. Plot of $\log_{10} P_c$ vs. ratio of covariance radius to miss distance

with each other. In Figure 1, we visualize the longitudinal nature of the data. We plot the $\log_{10} P_c$ values of ten events over time, with each events' values connected by a line; and we also plot these values versus the so-called ratio of combined covariance radius to miss distance. This is to expose the canonical trend in P_c development: as the event moves closer to TCA, the covariance shrinks, bringing this ratio slowly to a peak and then a marked drop-off.

It is not clear that the trend in $\log P_c$ values is stronger for ratio of covariance of radius to miss distance than days to TCA. In fact, the $\log P_c$ values exhibit a slight increasing trend with respect to the ratio, while the expected decreasing trend occurs with respect to time. It is possible that the $\log P_c$ values "drop off" at some small value of the ratio, but it is not clear from the data when this might occur, and how frequently one would observe it. Additionally, as was discussed in the introduction, this value is not monotonically increasing or decreasing with time (due to unpredictable changes in the covariance size and the estimate of the mean

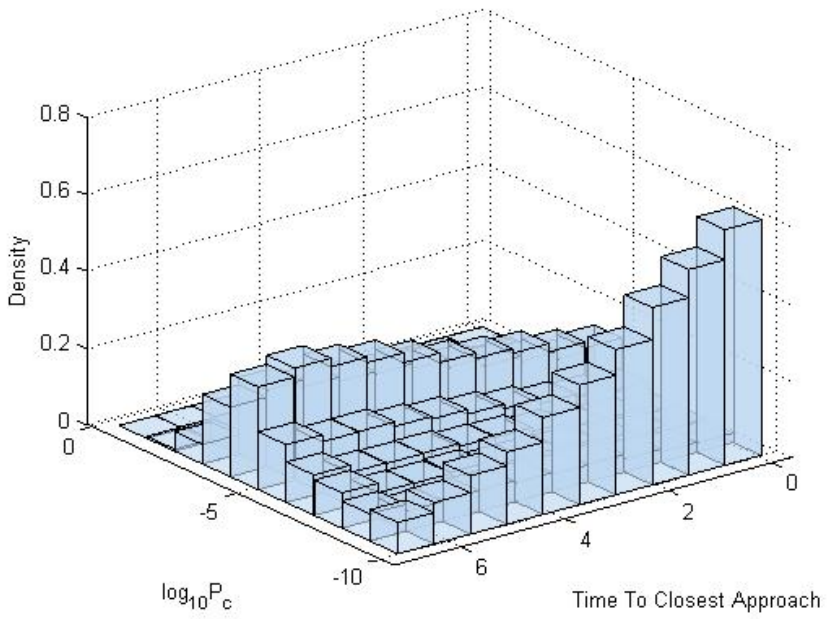

Figure 2.2. Two-dimensional histogram of $\log_{10} P_c$ values vs. TTCA

miss distance between the two satellites); so despite its theoretical linkage to the actual phenomenology of the situation, it is actually a less desirable independent variable for performing trending and prediction. As previously, model construction for P_c trending and prediction will use time to TCA as the independent variable.

We can visualize the trend of the $\log_{10} P_c$ values over time by considering a two-dimensional histogram, as displayed in Figure 2.2.

Recall that we have replaced all $\log_{10} P_c$ values below -10 with -10. Then the figure above indicates that the probability of observing a P_c value of 1E-10 or lower increases as one approaches TCA. In fact, at 2 days to TCA, about 40% of events observed have a P_c of 1E-10 or lower. At 7 days until TCA, the most observed value is about -5, which becomes less frequent over time, as more events observe a $\log_{10} P_c$ of -10. Interestingly, -5 seems to be the most likely value when one does *not* observe a -10, regardless of the time. We can use this information to construct a prior distribution for the model in Equation (1), as the increase of observed -10

values gives us an idea of how p behaves over time, and the observed mode of -5 of the $\log_{10} P_c$ values above -10 gives us some information about the mean of the Beta distribution.

2.2.4.2 Beta Regression. To model a Beta-distributed random variable with reference to a covariate (such as time), one must use a generalized linear model (GLM). Although GLM's for many other members of the exponential family (Normal, Gamma, etc.) have been developed since 1972[57], the GLM for the beta distribution is relatively new, being introduced in 2001[59]. The reason for this late development is due to the fact that one must reparameterize the beta distribution in order to model the mean sufficiently, an expansion that was not explored until recently. We develop this reparameterization for completeness. The probability density function (pdf) of a random variable X with a beta distribution is generally given as

$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where $\Gamma(\cdot)$ is the gamma function. The mean of this distribution is $E(X) = \frac{\alpha}{\alpha+\beta}$. GLM's are generally specified by setting some function $g(\mu)$ of the mean equal to a linear combination of covariates. For instance, logistic regression uses the logit link $g(\mu) = \log(\frac{\mu}{1-\mu})$, which is then set equal to a linear combination of covariates, e.g. $\beta_0 + \beta_1 X$, where X is a covariate, such as time. However, as the beta distribution is specified, it is unclear how to model the mean. To facilitate direct modeling of the mean, let $\mu = \frac{\alpha}{\alpha+\beta}$ and $\phi = \alpha + \beta$. Then we can rewrite the beta pdf as

$$f(x|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}.$$

Now we may model the mean μ directly. For instance, we may choose the logit link and model

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x + ... + \beta_p x^p,$$

so that the log-odds of the mean has a linear relationship to X. Various link functions are possible, such as the probit link, the complementary log-log link, and the log link. Our simulations have shown that there is no significant advantage in choosing one over the other, so we proceed with the logit link, as it is easy to interpret.

Recall that for each conjunction, one observes a different progression of P_c values. Sometimes the P_c values drop off quickly before TCA, other times they drop off much nearer TCA, and sometimes not at all. To model such a behavior, we may include a random intercept for each event as follows. Let μ_{ij} be the mean of the j^{th} P_c value in the i^{th} event, scaled to be between 0 and 1. Since we have $\log_{10} P_c$ values bounded between -10 and 0, a suitable transformation is $\mu_{ij} = E(Y_{ij})/10 + 1$, where Y_{ij} is the $\log_{10} P_c$ value of the j^{th} P_c value in the i^{th} event. We may consider the model

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + ... + \beta_p t_{ij}^p + b_i$$
$$b_i \sim N(0, \sigma_b^2),$$

where b_i is a random effect allowing for an intercept for the i^{th} event, and t_{ij} is the time until TCA of the j^{th} P_c value within the same event. One may additionally consider a random slope or other random effects for higher order terms, but given the amount of data, these would be difficult to fit and depend strongly on choice of prior distribution.

Recall that in Equation (1) we also introduced the parameter p. This parameter controls what percentage of the time we observe a zero. In our case, since about a third of our data are zeros, p might be close to 1/3. However, we also know that the closer an event approaches TCA, the more likely one is to observe a P_c value that is 0. As a result, we can also let p depend on our covariate. This parameter is also bounded between 0 and 1, so we again use a logit link function here (or any of the other aforementioned link functions). Additionally, we may consider a random

term for this model for each event, as the probability of observing a zero is higher for some events than others. Thus, we may consider a regression such as

$$\log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 t_{ij} + ... + \alpha_p t_{ij}^p + a_i$$
$$a_i \sim N(0, \sigma_a^2)$$

which is similar to the regression for μ above. Again, more random terms could be introduced if necessary.

2.2.4.3 Model Selection. Given below are some selected results from an exploratory model selection. To evaluate the relative merits of different levels of model complexity, we use the penalized deviance construct[73], where lower values indicate a better fit. Specifically, define $D(\theta)$ to be the "Bayesian Deviance", with form

$$D(\theta) = -2\log p(y|\theta) + 2\log f(y), \tag{2.8}$$

where $p(y|\theta)$ is the likelihood of y given θ and $f(y)$ is the saturated model, where $f(y) = p\{y|E(Y) = y\}$. We can rewrite $D(\theta)$ as

$$D(\theta) = -2\left(\log p(y|\theta) - \log f(y)\right), \tag{2.9}$$

which shows that $D(\theta)$ is -2 times the difference between the fitted model and the saturated model. Put simply, $D(\theta)$ measures how well a model fits the data relative to a model that fits the data perfectly. We estimate $D(\theta)$ with $\overline{D(\theta)}$, which can be written as

$$\overline{D(\theta)} = D(\bar{\theta}) + p_D, \tag{2.10}$$

where $p_D = \overline{D(\theta)} - D(\bar{\theta})$. The estimate $\overline{D(\theta)}$ is known as the *penalized deviance*, as it is computed as the sum of $D(\bar{\theta})$, the mean deviance, and p_D, the penalty term. The term $D(\bar{\theta})$ measures how well one fits the data, with lower values indicating

better fit, and the term p_D penalizes this fit for more parameters, where higher values indicate a larger penalty. The penalty term p_D is also known as the *effective number of parameters*, so that one may interpret this term as an estimate of how many parameters the model is actually estimating in order to describe the data. This is to account for models with more parameters fitting the data better, or over-fitting the data.

In Table 2.1 we provide the calculated mean deviance, penalty, and consequent penalized deviance for various models. These values justify how we came to our final model, as we chose the model with the lowest penalized deviance. The variables Y_c and Y_d represent the continuous and discrete parts of the model given in Equation (1), respectively. That is, Y_c are the values produced by the beta distribution, and Y_d are the 0-1 variables that either indicate a zero (1) or a continuous variable (0). All added complexities are in addition to the baseline linear model specified below:

$$Y_{ij} \sim f(y_{ij}|\mu_{ij}, \phi, p_{ij}) \tag{2.11}$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + b_i \tag{2.12}$$

$$b_i \sim N(0, \tau_b) \tag{2.13}$$

$$\tau_b \sim Gamma(0.001, 0.001) \tag{2.14}$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + \alpha_1 t_{ij} + a_i \tag{2.15}$$

$$a_i \sim N(0, \tau_a) \tag{2.16}$$

$$\tau_a \sim Gamma(0.001, 0.001), \tag{2.17}$$

$$\phi \sim Gamma(0.001, 0.001) \tag{2.18}$$

$$\beta_k, \alpha_k \sim Normal(0, 1), \quad k = 0, 1, 2. \tag{2.19}$$

Note that adding a random slope to either Y_c or Y_d did not produce a better fit, nor did specifying a correlation between the random effects.

Based on these results, we propose the following model. Let Y_{ij} be the j^{th}

Table 2.1: Model Selection Output: Beta regression

Model	Mean Deviance	Penalty	Pen. Deviance
Linear	-17.23	74.93	57.7
Quad Term for Y_c	-23.02	77.32	54.31
Quad Term for Y_d	-26.78	76.45	49.67
Quad Term for Y_c and Y_d	-32.52	79.23	46.71
QuadTerm for both, RanSlope for Y_c	-31.12	81.07	49.95
QuadTerm for both, RanSlope for Y_d	-36.91	85.54	47.63
Cubic Term for Y_c	-31.89	81.1	49.21
Quadratic, linear for phi	-27.76	80.87	53.11

scaled $\log_{10} P_c$ value of the i^{th} event. Also, let t_{ij} be the corresponding time until TCA (in days).

$$Y_{ij} \sim f(y_{ij}|\mu_{ij}, \phi, p_{ij}) \tag{2.20}$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + b_i \tag{2.21}$$

$$b_i \sim N(0, \tau_b) \tag{2.22}$$

$$\tau_b \sim Gamma(0.001, 0.001) \tag{2.23}$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 t_{ij}^2 + a_i \tag{2.24}$$

$$a_i \sim N(0, \tau_a) \tag{2.25}$$

$$\tau_a \sim Gamma(0.001, 0.001), \tag{2.26}$$

$$\phi \sim Gamma(0.001, 0.001) \tag{2.27}$$

$$\beta_k, \alpha_k \sim Normal(0, 1) \quad k = 0, 1, 2. \tag{2.28}$$

For this and all other models proposed in this section, we generate predictions for the next OCM by conditioning on the random effects. This yields the best linear unbiased predictor (BLUP), as discussed in Diggle[28]. Thus, we generate the predictive distribution

$$Y_{i(j+1)} \sim f(y_{i(j+1)}|\mu_{i(j+1)}, \phi, p_{i(j+1)}, a_i, b_i),$$

where $\mu_{i(j+1)}$ and $p_{i(j+1)}$ are produced by using $t_{i(j+1)}$ in their respective models. Using this distribution, we can construct credible sets. As in the other Bayesian models, we use the posterior mode of this predictive distribution as our estimate for $Y_{i(j+1)}$.

2.2.4.4 Issues of Identifiability. The model proposed in equations (10)-(17) has a total of 7 parameters and 2 random effects, which suggests one must estimate a total of 9 quantities in order to make inferences and hence predictions. However, this issue can be ameliorated by using informative priors in a Bayesian framework. To acquire these informative priors, we run the proposed model on a training dataset of a large number of events, which are not themselves used for model evaluation. We use the posterior distribution of the parameters as informative prior distributions by matching sample moments of the posterior samples with its prior distribution family. We do this for all of the population-level parameters, which are ϕ, α_k, and β_k for $k = 0, 1, 2$. Then we are left with two parameters to estimate, the random effect variances, τ_a and τ_b. Because we begin making predictions beginning with the second observation, these parameters are identifiable when making inference on a single event.

Motivated by the large number of events in our testing data set, we investigated whether prediction could be improved by making inferences on more than one event at once. In order to test this, we followed the mean squared prediction error (MSPE) when making predictions on one event, 5 events, 10 events, and 25 events. Including more events did not improve the MSPE, likely due to the fact that, in reference to a single event, other events only contribute to the population-level parameters, which are already well-known due to the informative prior distributions. Thus, we make predictions on a single event at a time.

2.2.4.5 An Aside: Coverage from an Initial Simulation. In an initial exploratory simulation, we found that 97.5% prediction intervals constructed in the Beta model had 86% coverage. Though coverage is often lower than the nominal rate with real data, we found this coverage to be too low to have any meaningful operational use. In exploring this phenomenon, we found that splitting up the dataset into three parts, a high-, medium-, and low-risk group, ameliorated the issue of low coverage. Specifically, if an event had a high (above -4) $\log P_c$ value by 3 days time to TCA (TTCA), we called it high-risk. If an event had a medium (between -7 and -4) $\log P_c$ value by 3 days TTCA, we called it medium-risk. If an event had a low (below -7) $\log P_c$ value by 3 days TTCA, we called it low-risk. We shall refer to the high-, medium-, and low-risk groups as Red, Yellow, and Green hereafter.

Incidentally, the fact that our model performed well when the data were separated into different risk groups supports the notion that the $\log P_c$ value behaves differently depending on the quantile it inhabits. In terms of the Beta regression model, this implies that the population-level trend is different for these different risk groups, which suggests that they ought to be modeled separately. For the simulation presented in this chapter, these definitions worked well and possess the additional advantage of aligning closely with thresholds presently in use operationally for categorizing conjunction event severity.

2.2.4.6 Checking Assumptions. A variety of assumptions are employed in the model. These include

- A linear model for $g(\mu) = \log(\mu/(1-\mu))$
- A linear model for $h(p) = \log(p/(1-p))$
- The dispersion parameter ϕ is constant across time.
- The logit of the mean values associated with the Beta distribution Y_c are parallel for each event. That is, the logit of the means for each event is

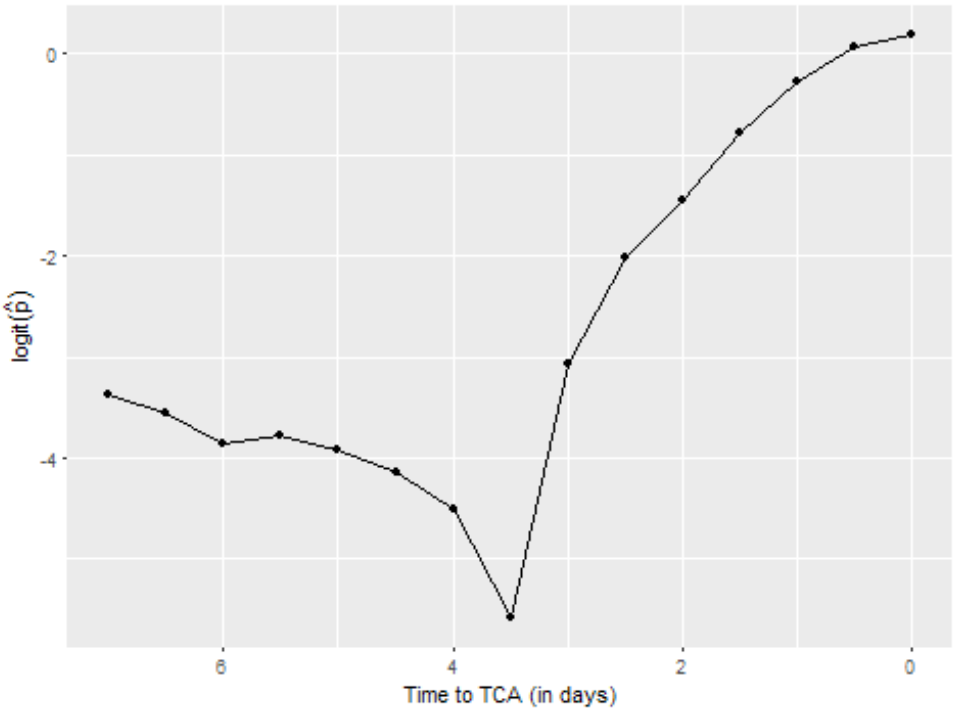

Figure 2.3. Plot of $\log(\hat{p}/(1-\hat{p}))$ vs. TTCA

separated by a random intercept.

- The random intercepts have a normal distribution.

We check these assumptions graphically below. Figure 2.3 plots $\log(\hat{p}/(1-\hat{p}))$ vs. TTCA to check the form of the linear model for p.

Figure 2.3 shows that, though a single second order polynomial may fit $logit(p)$, a piecewise function of two second order polynomials may be more appropriate. The graph suggests that separate models might be appropriate for the time intervals $(0, 3.5)$ and $(3.5, 7)$. Figure 2.4 checks the form of the linear model for μ. Figure 2.5 attempts to check the form of the linear model for ϕ by estimating ϕ via its profile likelihood.

Again, these graphs suggest that it may be appropriate to fit piecewise polynomials on the intervals $(0, 3.5)$ and $(3.5, 7)$. Finally, we try to visualize the form of the random effects in Figure 2.6 by plotting $\log(y/(1-y))$ vs. TTCA for 15 randomly

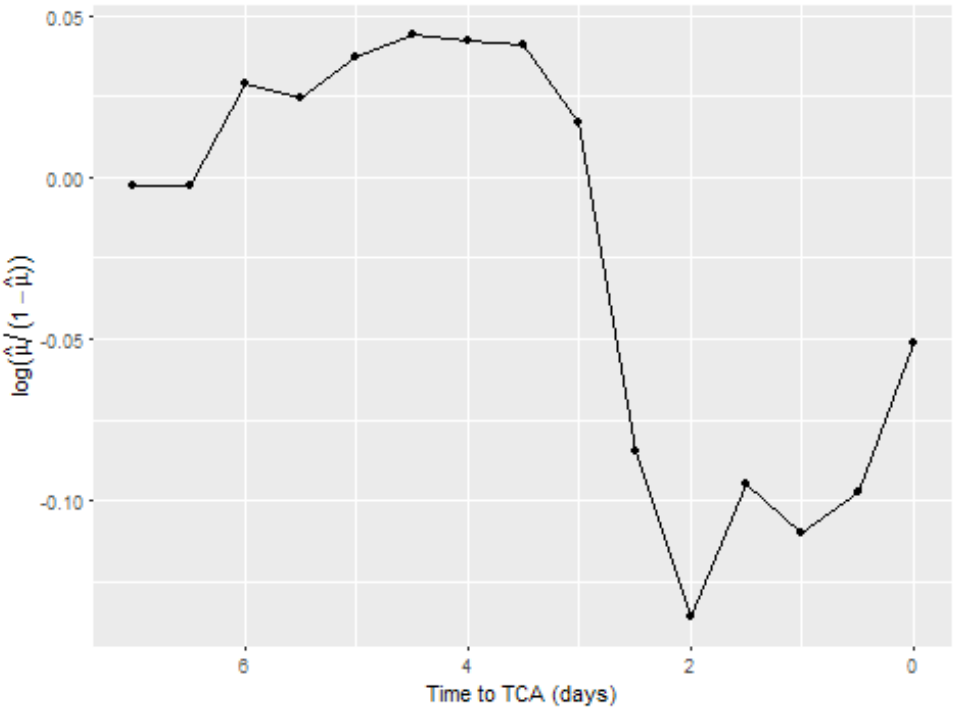

Figure 2.4. Plot of $\log(\hat{\mu}/(1-\hat{\mu}))$ vs. TTCA

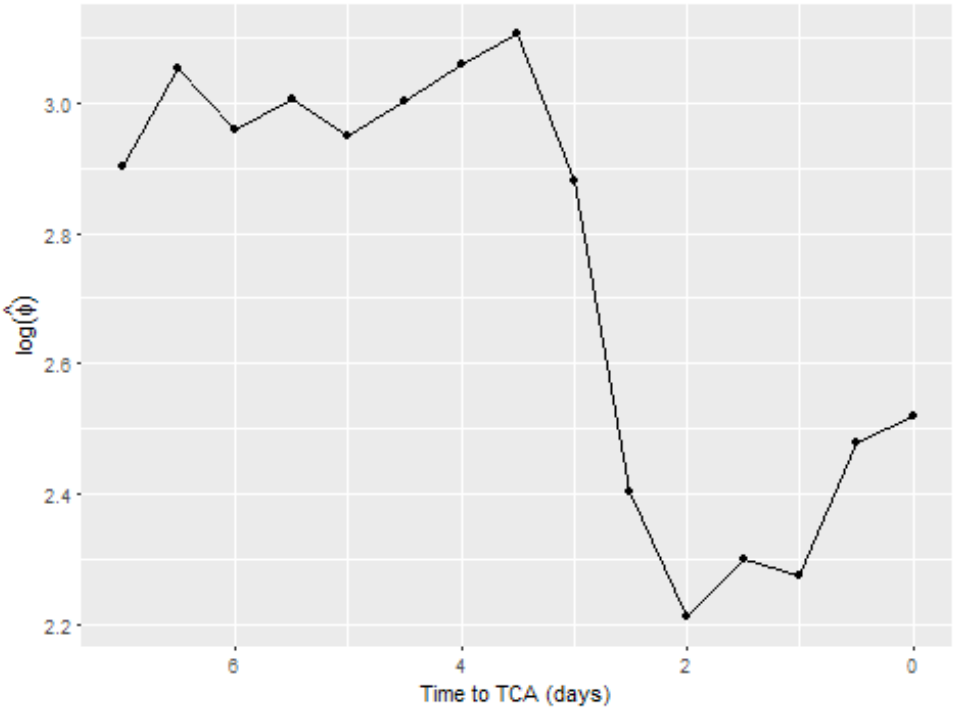

Figure 2.5. Plot of $\log(\hat{\phi})$ vs. TTCA

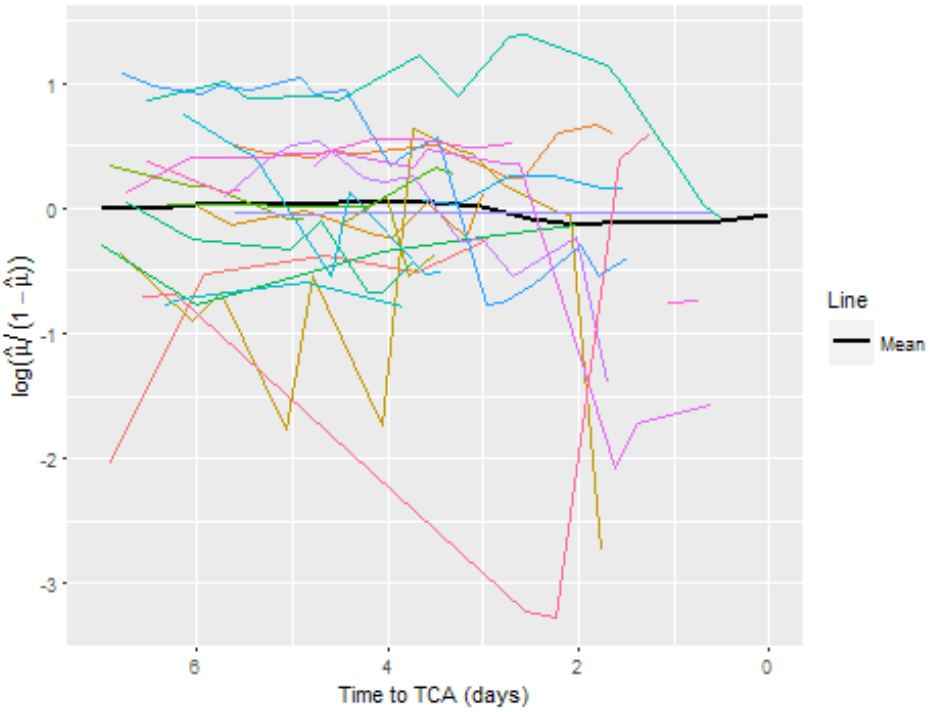

Figure 2.6. Spaghetti plot of $\log(y/(1-y))$ vs. TTCA

sampled events. Though a random intercept does *not* imply parallel trajectories over time, the trajectories are approximately parallel for small random intercepts.

The graph above depicts the overall mean (in black), $\log(\mu)/(1-\mu)$, as well as this value for 20 random events (color lines). Though there is significant heterogeneity in the paths of the events over time, it appears that a random intercept alone may be sufficient, as most paths seem to be parallel to the overall mean.

2.2.4.7 Threshold Model. As evidenced in the graphs above, it appears that at 3.5 days until TCA, there is a noticeable change in behavior in both the behavior of μ and p. As a result, modeling this change in behavior may be of interest. To model this behavior, we specify a trend on the interval (7, 3.5) days to TCA, and a different trend on the interval (3.5, 0) days to TCA.

We specify two such models. In the table below, "Threshold" is a model with

Table 2.2: Model Selection Output: Threshold Beta regression

Model	Mean Deviance	Penalty	Pen. Deviance
Beta Reg	-451	81.66	-369.3
Threshold	-492.9	85.31	-407.6
Threshold 2	-751.5	107.6	-653.9

this threshold modeling for μ only, while "Threshold 2" uses such a structure for p as well. It is clear from the results that a threshold may be appropriate for both parameters, as the DIC is lowest for the "Threshold 2" model.

The resulting model is given below. Notice that it has the same construction as the previous Beta regression model, but with different trends specified on the time intervals (7, 3.5) and (3.5, 0) days to TCA.

$$Y_{ij} \sim f(y_{ij}|\mu_{ij}, \phi, p_{ij}) \tag{2.29}$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = (\beta_{01} + \beta_{11}t_{ij} + \beta_{21}t_{ij}^2)I_{[0,3.5]}(t_{ij}) \tag{2.30}$$

$$+ (\beta_{02} + \beta_{12}t_{ij} + \beta_{22}t_{ij}^2)I_{(3.5,7]}(t_{ij}) + b_i \tag{2.31}$$

$$b_i \sim N(0, \tau_b) \tag{2.32}$$

$$\tau_b \sim Gamma(0.001, 0.001) \tag{2.33}$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = (\alpha_{01} + \alpha_{11}t_{ij} + \alpha_{21}t_{ij}^2)I_{[0,3.5]}(t_{ij}) \tag{2.34}$$

$$+ (\alpha_{02} + \alpha_{12}t_{ij} + \alpha_{22}t_{ij}^2)I_{(3.5,7]}(t_{ij}) \tag{2.35}$$

$$a_i \sim N(0, \tau_a) \tag{2.36}$$

$$\tau_a \sim Gamma(0.001, 0.001), \tag{2.37}$$

$$\phi \sim Gamma(0.001, 0.001) \tag{2.38}$$

$$\beta_k, \alpha_k \sim Normal(0, 1) \quad k = 0, 1, 2. \tag{2.39}$$

2.2.5 New Beta Regression Model

In light of the cubic nature of the means shown above, as well as the observed success of the Look-Up model, we introduce a new model, which we refer to as the New Beta Regression model. This model incorporates a cubic model into the model for μ_{ij}, as well as previous observations. Unlike the Look-Up method, we incorporate the values themselves rather than the quantiles.

$$Y_{ij} \sim f(y_{ij}|\mu_{ij}, \phi, p_{ij}) \tag{2.40}$$

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \beta_4 y_{(i-1)j} + b_i \tag{2.41}$$

$$b_i \sim N(0, \tau_b) \tag{2.42}$$

$$\tau_b \sim Gamma(0.001, 0.001) \tag{2.43}$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 y_{(i-1)j} + a_i \tag{2.44}$$

$$a_i \sim N(0, \tau_a) \tag{2.45}$$

$$\tau_a \sim Gamma(0.001, 0.001), \tag{2.46}$$

$$\phi \sim Gamma(0.001, 0.001) \tag{2.47}$$

$$\beta_k, \alpha_k \sim Normal(0, 1) \quad k = 0, 1, 2. \tag{2.48}$$

2.2.6 Bayesian Beta Cluster Regression

Though the Beta regression model accounts for more aspects of the data than the more naive models, results (presented in an upcoming section) suggest that this model may not be accurate enough to be operationally useful. In addition, there is an aspect of the data which this model cannot address, which may suggest a better model. In particular, operators often suspect that there are different categories of events: low risk, medium risk, and high risk. Furthermore, operators believe that these categories of events behave slightly differently, so that if one knew with a high degree of certainty which category of event was observed, we would be able to obtain

more accurate inference. However, these categories are relatively arbitrary. It may be of interest to know exactly how many significantly different categories of events there are. The Bayesian Beta Cluster Regression model is constructed to tackle this question, and hopefully provide more accurate inference. More technically, we are interested in how many clusters are in the data, what the clusters look like, and how likely we are to be able to identify which cluster is occurring by the time a decision is to be made.

2.2.6.1 Model. The model is conceptually straightforward. We now assume that there are K different means within the data, representing the means of K clusters. That is, each event is assumed to come from one of the K clusters. This can be expressed as a mixture model, so that

$$Y_{ij} \sim \sum_{k=1}^{K} \pi_k f(y_{ijk}|\mu_{ijk}, \phi_{ijk}, p_k) \tag{2.49}$$

where π_k is the probability Y_{ij} is from cluster k. In this general model, each cluster has its own mean μ_{ijk}, dispersion parameter ϕ_{ijk}, and probability of observing a P_c of zero p_k. Model selection will enable us to determine the size of K and which parameters do not vary across clusters.

To form a complete Bayesian specification, priors for all parameters are necessary. The same non-informative prior structure given above is used for the parameters of all clusters. In addition, we let

$$\pi_1, ..., \pi_K \sim Dirichlet(e_1, ..., e_K) \tag{2.50}$$

where the e_i are chosen to be non-informative, and thus all are set to one. Let S_{ik} be a random variable which indicates whether event i is part of cluster k or not, taking value 1 if this is true, and 0 if it is false. Then

$$Y_{ij}|S_{ik} \sim f(y_{ijk}|\mu_{ijk}, \phi_{ijk}, p_k), \tag{2.51}$$

where $f(\cdot)$ is a zero-inflated Bayesian Beta regression model, as given above.

Table 2.3: Model Selection Output: Beta cluster regression

Clusters	Mean Deviance	Penalty	Pen. Deviance
1	-728.4	103.7	-624.7
2	-1462	901	-561
3	-1841	1085	-756
4	-1960	1917	-43.64

2.2.6.2 Model Selection. Given below is a model selection table similar to that in section 2.3.2. The penalized deviance is given by number of clusters, k, for the above model.

As is evidenced above, the model with 3 clusters seems to fit better than the model with 1 cluster. Though it may seem odd that the model with 2 clusters does not also have a lower DIC than the model with 1 cluster, it should be noted that DIC is notoriously difficult to calculate for mixture distributions, and also has problems associated with it[16]. As a result, we also consider an ad-hoc method of selecting the number of components. We set k to the maximum number of possible clusters, and calculate the posterior probability of each component. Specifically, we set $k = 5$, and find that the probability of the first two components is 0.63, and 0.35, respectively. This yields some evidence that the number of clusters necessary is two.

From a practical standpoint, adding more clusters increases the computation time, making fewer clusters more desirable for implementation. In addition, as the number of clusters grows, one encounters other computational issues. For example, a greater number of clusters is generally accompanied by slower convergence, as well as label-switching[66][45], a problem in the posterior samples in which a parameter "switches clusters", due to the invariance of the likelihood to label-switching. This, combined with the above evidence, leads us to consider a model with only 2 components. Results from an initial fit of this model show that, if two clusters indeed do

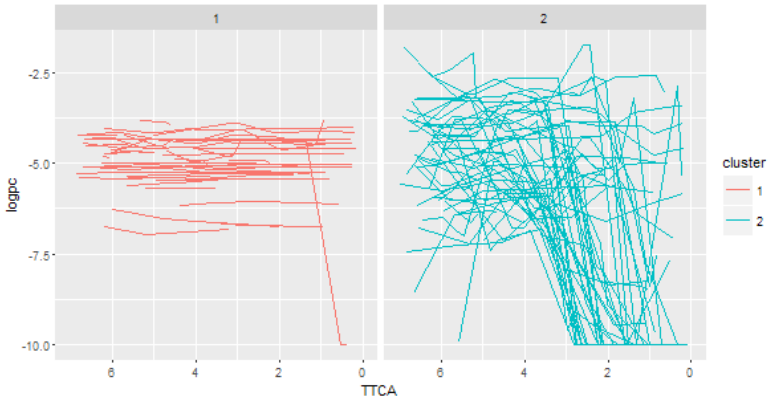

Figure 2.7. Clusters found from Beta clustering model

exist, one is a "high variability" cluster and one is a "low variability" cluster. Figure 2.7 shows these two clusters. Cluster 1 contains events which show little variability in $\log_{10} P_c$ over time. Interestingly, most of these values are centered around -5. In contrast, cluster 2 contains events which exhibit much more variability over time. Of course, the difficulty with such a model must identify the cluster appropriately to make valid predictions for a given event. Regardless of predictive performance, this model is useful because it gives us new insight into the data. For instance, if such a model could reliably classify events into "high" and "low" variability clusters, operators would have yet another way to temper their expectations about future $\log_{10} P_c$ behavior.

2.3 Measures of an Effective Model

In this section we discuss the properties on which we will compare our two models. We focus on model fit and decision-making performance.

2.3.1 Model Fit

The main concern in building predictive models is fitting the data well enough to predict new observations accurately. In order to quantify this, we check the bias, prediction errors, and upper bounds of the proposed models. Specifically, we would

like our models to be unbiased, so that the prediction errors are centered at zero. Secondly, we check to see if the prediction intervals are bigger or smaller for different times, predicted values, and times until prediction. Lastly, we check to see that our upper bounds have the correct coverage.

2.3.1.1 Prediction Coverage. Let t_{ij} be the TTCA of the j^{th} observation i^{th} event. Then our predictions and the associated confidence intervals are made at $t_{i(j+1)}$. Thus, we predict the distribution of $y_{i(j+1)}$. As noted before, we make predictions beginning with the second observation, $j = 2$. We construct 95% confidence intervals, and check whether the true value, $y_{i(j+1)}$, is contained in the interval. The average of the number of true values contained within these predicted intervals is our prediction coverage.

It should be noted that the time until prediction is *not* the same for all predictions. As noted before, the time at which new P_c values are received are random, producing irregular times between successive observations. The time between observations is usually at most 2 days. Though the time between observations varies, we calculate coverage irrespective of time, so that a prediction 2 days into the future contributes equally to the coverage as a prediction 0.5 days out. Though one would prefer to make predictions at the same number of days into the future for all events, this is not possible due to the nature of the data.

2.3.2 Decision-Making Efficacy

The models presented above are ultimately used to make decisions about whether to move a satellite or not. Below, we discuss some methods used in assessing the decision-making efficacy of these models.

2.3.2.1 Framework. In order to assess our models in the framework of making decisions about whether to continue active monitoring of an event, we implement a simple decision-making framework and study its properties in both models. Because the most weighty period for conjunction assessment operational decision-making occurs 2-3 days TTCA, we focus on this region of data. Specifically, we make predictions at 2 days TTCA and make a decision based on this prediction. Let $\hat{y}_2$ be an estimate of the $\log P_c$ predicted to occur at 2 days TTCA. We will make the decision that the $\log P_c$ values will remain above the threshold θ after 2 days TTCA if

$$\tilde{y}_2 > \theta \tag{2.52}$$

and will make the decision that the $\log P_c$ values will fall below the threshold θ otherwise. To couch this in the hypothesis testing framework, we write

$$H_0 : \tilde{y}_2 < \theta \quad vs. \quad H_1 : \tilde{y}_2 > \theta, \tag{2.53}$$

so that rejecting H_0 is synonymous with claiming the $\log P_c$ will remain high. In our simulations, we set $\theta = -5$ for the Red group and $\theta = -7$ for the Yellow group. Note that while -7 is the lower bound for being in the Yellow group at 3 days TTCA, -5 is below -4, the corresponding lower bound for the Red group. A lower threshold was chosen as these events are generally of much higher concern, thus one prefers an extra order magnitude of certainty before claiming the event is at a lower risk level. In order to explore this trade-off fully, we tried various quantiles of the distribution of $\tilde{y}_2$, which we describe below.

2.3.2.2 Type I and Type II Errors. As with most decision-making frameworks, our framework can admit Type I and Type II errors. The hypothesis in (22) is framed in terms of the event of a P_c value remaining high, as this is the event we are most concerned with. A Type I error here is the incorrect assertion of a high

P_c value (i.e. a false alarm), and a Type II error is a the more worrisome incorrect prediction of a low value (i.e. a missed detection). Thus, while we may find it acceptable to trigger an alarm when the $\log P_c$ value has actually dropped off, it is almost *never* acceptable to miss detecting a high $\log P_c$ value. Of course, we can make our system as powerful against this event as we want, with the trade off of triggering more false alarms. It's worth noting that a false alarm for a high value is the same thing as missed detection for a value which has dropped off. Ideally, we would like to have an alarm that detects high values *and* low values with a high degree of accuracy. However, since we are more concerned with high values, we seek to quantify how often, if ever, can we detect these low values while still maintaining the high accuracy needed for detecting the high values.

2.4 Results

2.4.1 Data

Recall that we split our data into two groups. The Red and Yellow groups are defined below.

- If an event had a high (above -4) $\log P_c$ value by 3 days time to TCA (TTCA), it is part of the Red group.
- If an event had a medium (between -7 and -4) $\log P_c$ value by 3 days TTCA, it is part of the Yellow group.

The dataset used for tuning (*i.e.*, setting the parameters for the informative prior distributions) and testing the model was taken from the NASA Conjunction Analysis and Risk Assessment historical Conjunction Data Message (CDM) database. For the Yellow group, five hundred events' worth of data from calendar year 2013 was used for model tuning (the "training" dataset), and the tuned model was evaluated against approximately 2000 events from 2014 (the "validation" dataset); so there was no overlap in terms of time-period or actual data between

the two datasets. For the Red group, 82 events were used for training and 70 were used for testing (this data set is far smaller, as these kinds of events are more rare). Data were taken from conjunctions against primaries in the orbital region defined by a perigee height between 500 and 750 km and an eccentricity less than 0.25. As described above, data flooring at a $\log_{10} P_c$ value of -10 was performed on the dataset. To qualify for use in tuning or evaluation, an event must have had at least two CDMs with a $\log_{10} P_c$ greater than -10.

2.4.2 Simulation Setup

To train our model, we perform a Bayesian analysis on the training data using non-informative priors. We determine the distribution parameters for the informative priors used in the test data by matching the first and second moments of the observed distributions to the hypothesized prior distributions. All MCMC inference is conducted in JAGS[62].

The simulation procedure for a given event is as follows. We attempt to make predictions for the next y value only after the second received CDM. We are interested in estimating the next $\log P_c$ value, which we predict by using the time at which it was observed. The predicted value is taken to be the mode of the posterior predictive distribution. In this context, it is important to use the posterior mode as opposed to the posterior mean, as the posterior predictive distribution is generally bimodal, with some mass at -10 and the remaining density between -10 and 0, inducing another peak. Thus, we choose the "most likely value" as opposed to the mean. We make predictions for five models: the Vertex model, the Beta Regression model, the New Beta Regression model, the Beta Clustering model, and the Look-Up model.

To further assess model fit, we also track a two-sided 95% credible interval for each prediction. We utilize the upper bound from the credible set for checking

coverage. This is also done for the Look-Up and LOCF methods, though here the interval is a confidence interval and is calculated using repeated cross-validation. In addition to coverage, we are also interested in how many of these upper bounds are low enough to be "useful". That is, we would like to know how many of these lower bounds are lower than the lower threshold of the Yellow and Red groups.

We present results for the Yellow group only. We found that the Red group had too small of training and testing sets to yield any kind of meaningful conclusions about predictive performance. Further simulation is required to determine just how different these two groups are, and if the results shown below hold for more high risk events like those in the Red group. We suspect this may be the case, as our construction of these two groups was somewhat arbitrary in the first case. As shown by the Beta clustering model, it is likely that when it comes to modeling, there is more of a delineation between "high variability" and "low variability" events than high-risk and low-risk.

2.4.3 Models

Here, we make a few remarks about the models chosen for simulation. We choose to compare seven of the proposed models: Last Observation Carried Forward (LOCF), the Look-Up model (LKUP), the Vertex model (Vertex), the Beta Regression model (BetaReg), the Beta Clustering model (BetaClust), and the New Beta Regression model (BetaNew). Not all models are shown here to reduce clutter in the graphs and to ease comparison. We do not include the Beta Threshold model because this gave results similar to the Beta Regression model. The functional data model is not considered because it too yielded results similar to the Beta Regression model. Though promising, this model did not have greater success because one of the assumptions of the model was not met. As mentioned earlier, the functional principal component scores are assumed to follow a normal distribution. In practice,

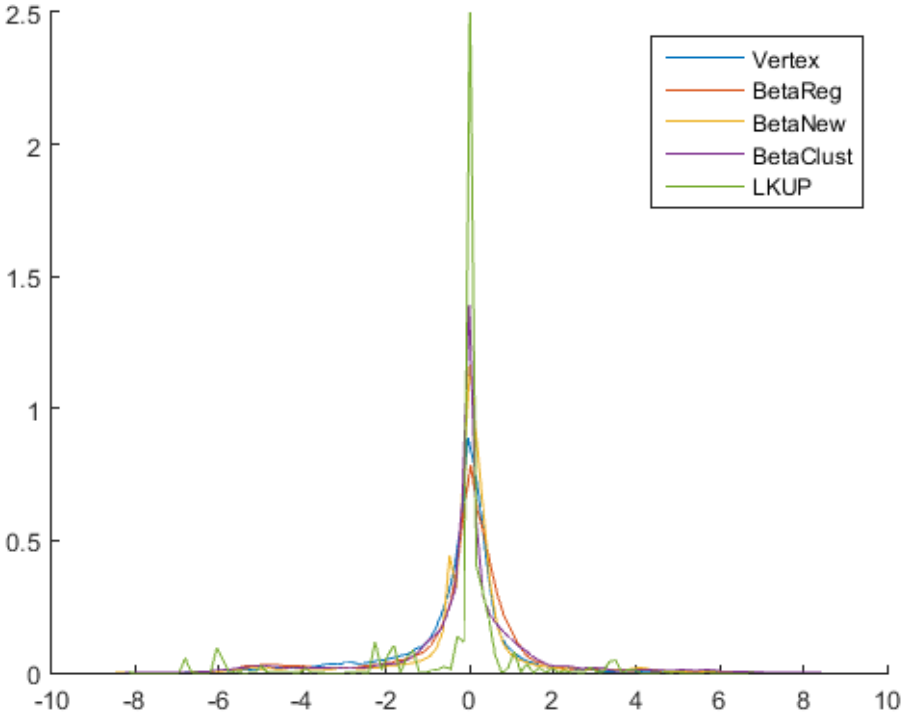

Figure 2.8. Density plot of estimated prediction errors for all models

we found these to have a bimodal distribution, suggesting clustering. This prompted us to investigate clustering models, such as the Beta clustering model.

Inference for all models is based on prediction at the next OCM. Originally, we investigated the potential use of basing inference on the estimated peak in the Vertex model. We found that this produced poor results, and that better inference resulted from simply making predictions at the next OCM.

2.4.4 Simulation Results

Our main goal of prediction is to make a decision by 2 days to TCA. As a result, for the Yellow group, we are interested in whether a $\log P_c$ which is observed after 2 days to TCA will be below -7 or not. Additionally, for this group, we may be interested in a "worst case" scenario, which would be observing a $\log P_c$ value above -4 at this time. We also present results for this prediction.

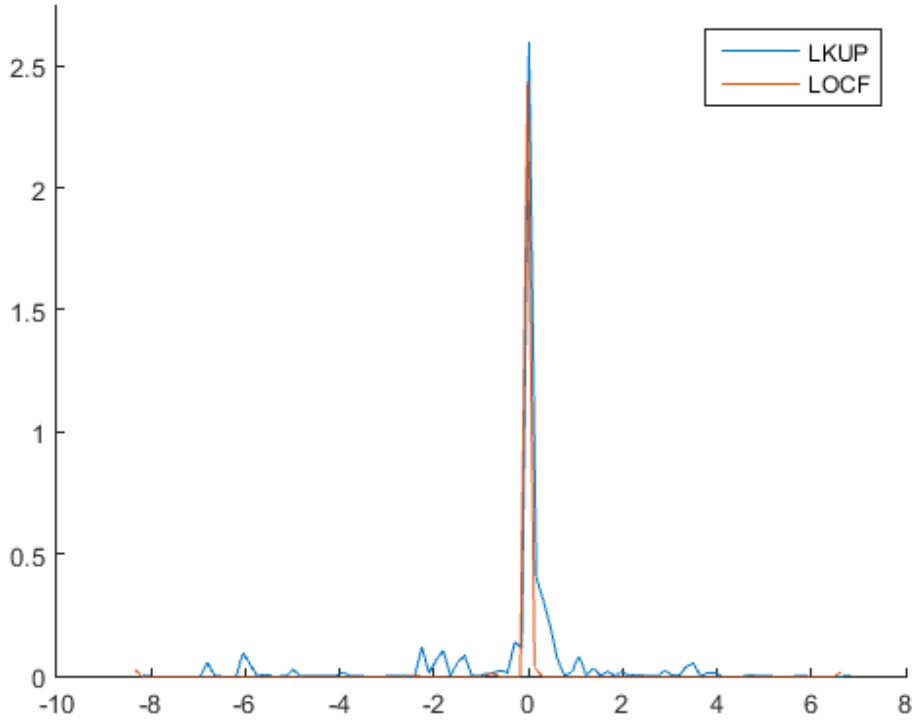

Figure 2.9. Density plot of estimated prediction errors for the Look-Up and LOCF methods

As the goal of our inference is prediction, we present prediction errors for the various models. Figure 2.8 shows the density of the prediction errors for all five models. The Look-Up model is more peaked than the other models, suggesting that is produces more prediction errors close to zero. Though this suggests the predictions are more accurate than the other models, a closer inspection suggests that the tails of the prediction errors are nearly as long as those produced by other models. Interestingly, the Beta clustering model is the second most peaked.

Figure 2.17 provides a look at the prediction errors of the Look-Up method vs. LOCF. Here we see that the LOCF method produces similar prediction errors to the Look-Up method. In addition, we see an interesting artifact of prediction error density generated by the Look-Up model: the tails are jagged, unlike the LOCF model, which has smooth tails. This suggests that many prediction errors are nearly identical, likely resulting from nearly identical predictions in similar situations. For

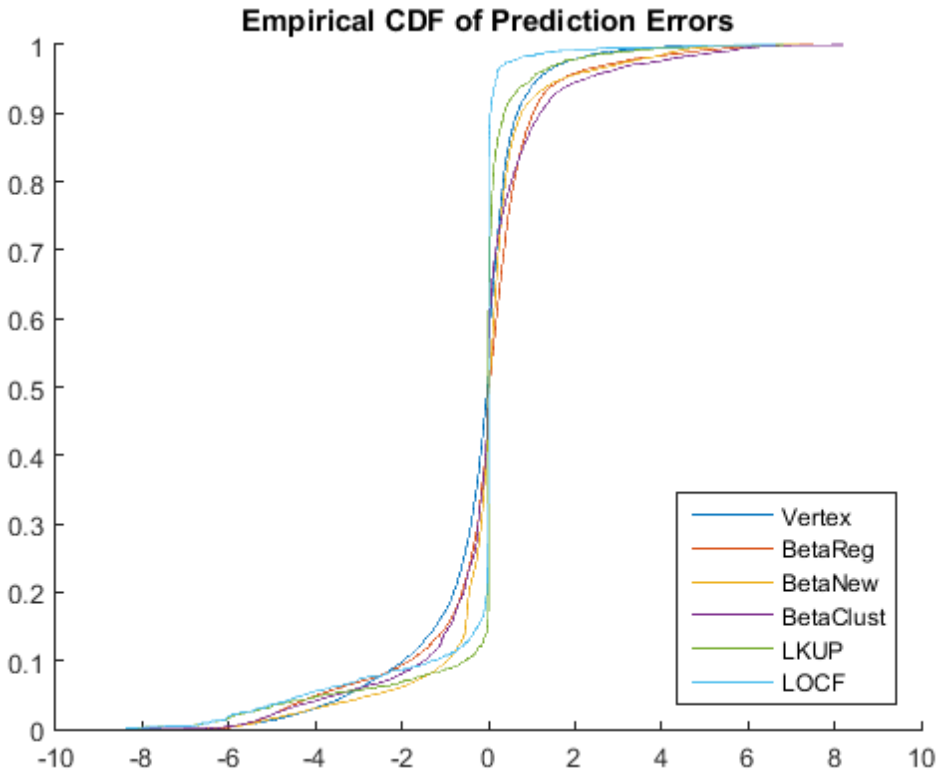

Figure 2.10. Empirical CDF of prediction errors for all models

instance, it may be that this model routinely generates a prediction of -7 when the next value is actually -10, resulting in many prediction errors around -3. This is likely due to the fact that the method makes predictions based on percentiles, and many percentiles have nearly identical values, resulting in nearly identical predictions.

To better understand the tail behavior of all of the models, we plot the empirical CDFs for the predictions errors produced by each model. The resulting plot in Figure 2.10 suggests that LOCF model has the shortest right-hand tail, resulting in smaller and fewer positive prediction errors. This is ideal, as a positive prediction error means that one predicted a low $\log P_c$ when in fact the next $\log P_c$ was higher, indicating an under-estimation of risk. Note that this implies that the practice of using the previous $\log_{10} P_c$ value for inference is more likely to overstate the risk than to understate it. We visualize exactly how much the risk is overstated later. The Beta Clustering model seems to generate the shortest left-hand tail.

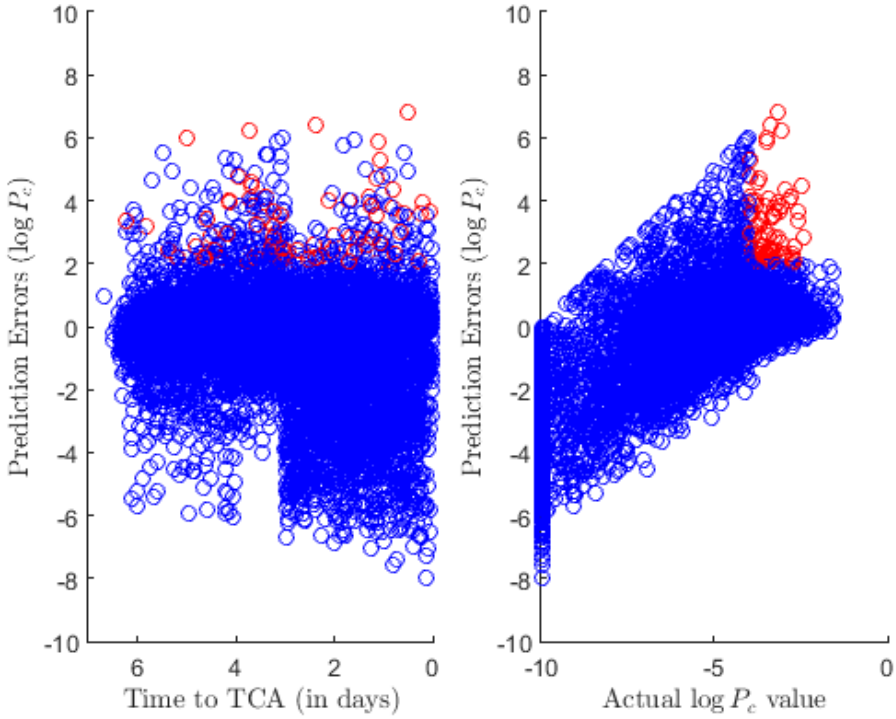

Figure 2.11. Vertex Model

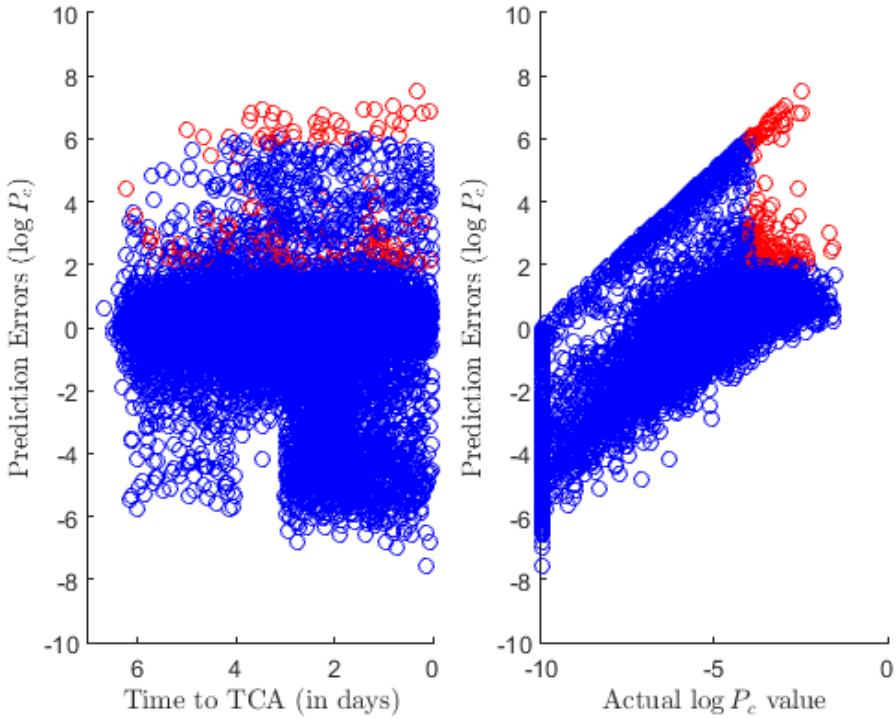

Figure 2.12. Beta Regression Model

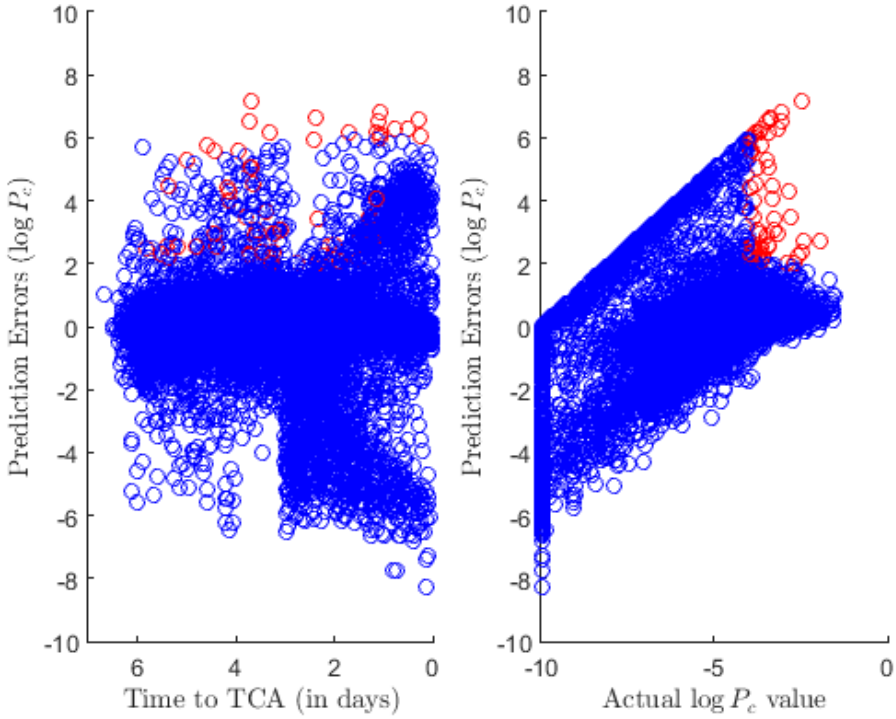

Figure 2.13. New Beta Regression Model

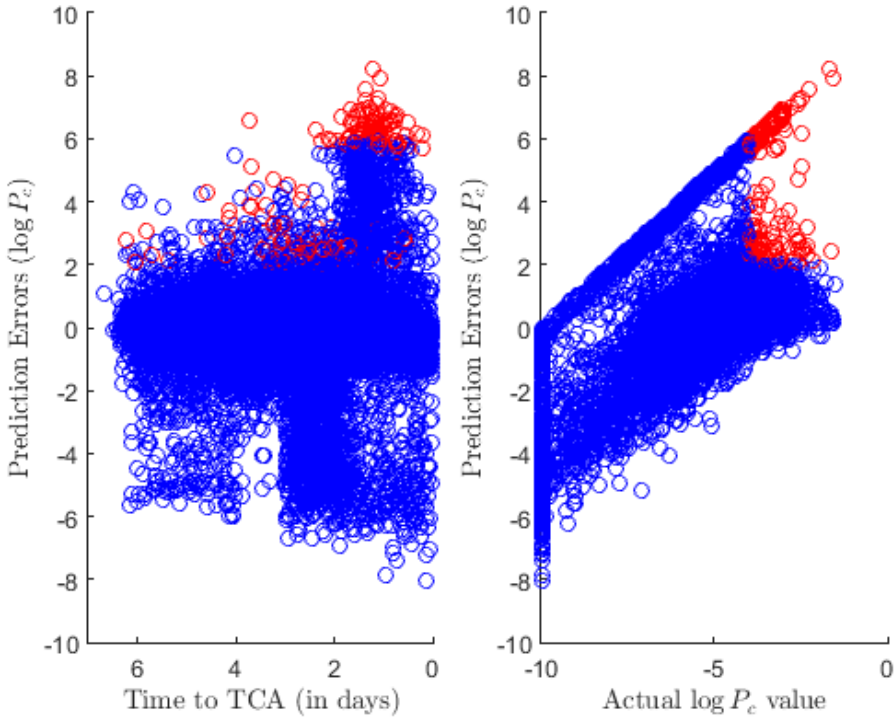

Figure 2.14. Beta Clustering Model

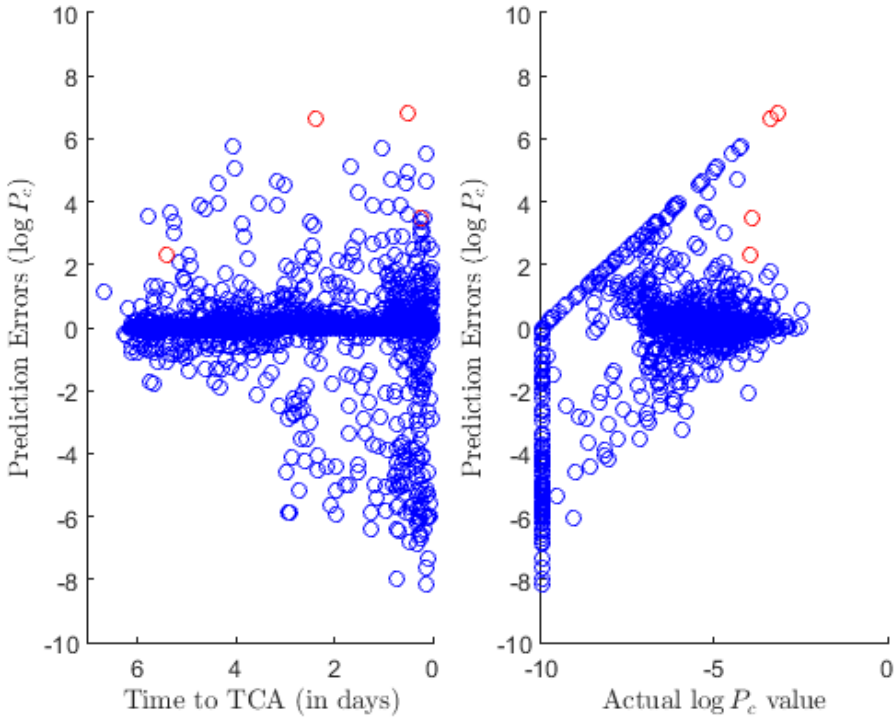

Figure 2.15. Look-Up Model

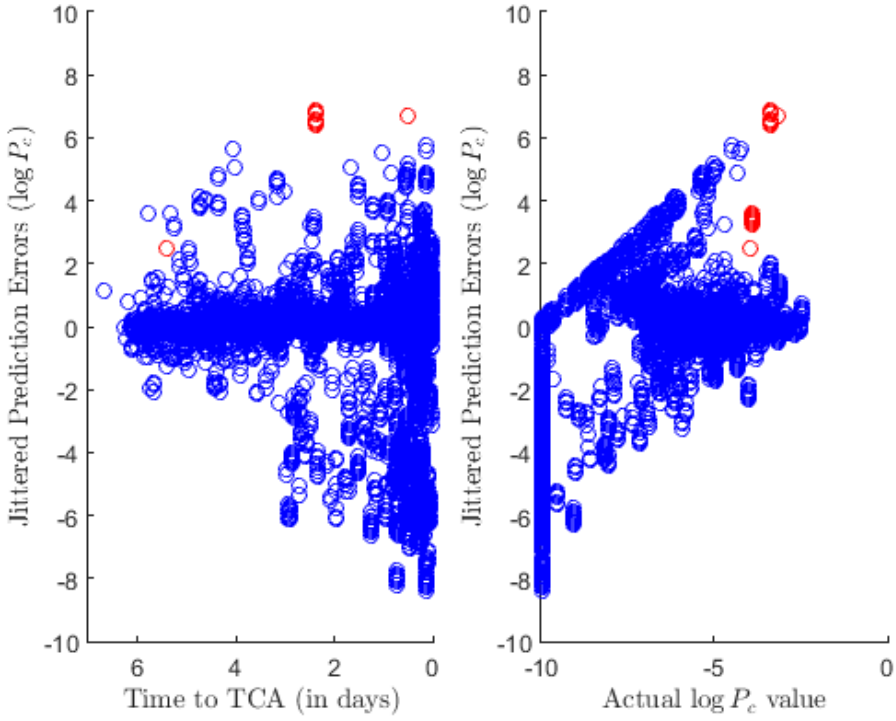

Figure 2.16. Look-Up Model (with jitter)

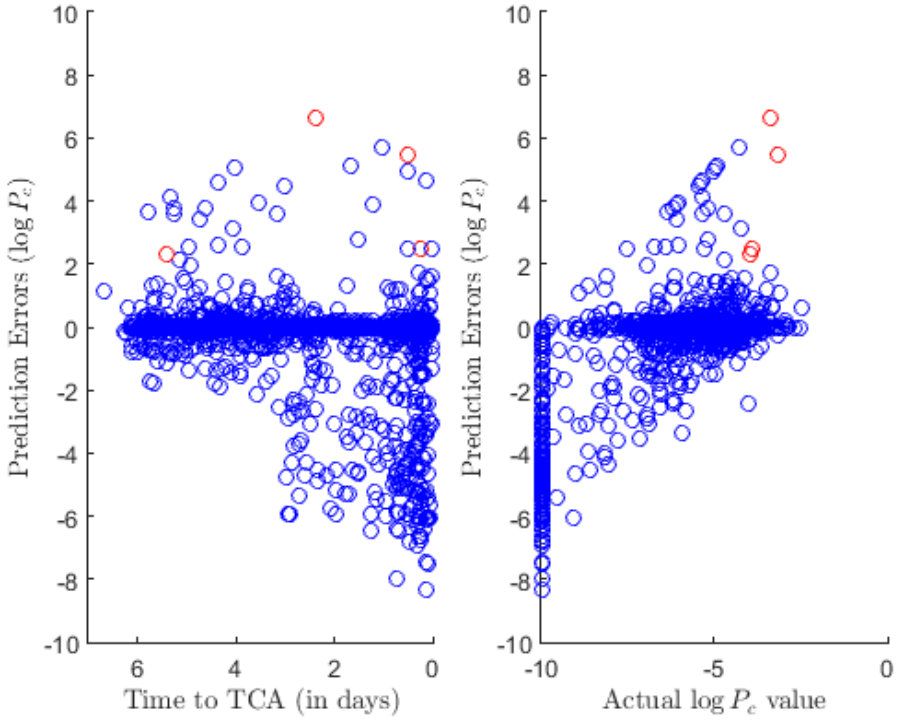

Figure 2.17. Last Observation Carried Forward

Figures 2.11-2.17 plot the prediction errors vs. time and vs. actual log P_c value for each of the models. All of the models have smaller prediction errors from 6 days to TCA to 4 days to TCA than later time points. Prediction errors in this time range are smallest for the Look-Up and LOCF methods, where they are virtually all zero, implying near-perfect prediction. However, prediction in this time range is not particularly of interest, as decisions are usually made at 2 days to TCA (or later). It is interesting to note that the LOCF method produces errors largest in the positive direction, suggesting that in this time frame one may see a significant drop-off, but one almost never sees a jump from, say, a $\log_{10} P_c$ of -10 to -4.

We focus on the time range of 2 days to TCA. Most of the models produce more and larger negative prediction errors than positive prediction errors in this time range, suggesting they err on the side of predicting "too high a risk". In general, this conservativeness is better operationally. This trend is **not** true for the

Beta Clustering model, which produces a far greater number of positive prediction errors. This is likely due to predicting a $\log P_c = -10$ far more often than it actually happens. Perhaps this could be ameliorated by exploring some tuning of the priors in the model, though this is likely more onerous than practical for decision-making. In addition, this model takes far longer to run the other models, which only adds to the time one would need to implement it.

To highlight underestimates of high-risk events, we color in red all observations where the prediction error was 2 or greater and the actual $\log P_c$ value is -4 or higher. Interestingly, we see that the Vertex and New Beta Regression models have few of these points, whereas the Beta Regression and Beta Clustering models have noticeably more. In addition, many of these points are within 2 days to TCA for the latter two models. Hence, for erring conservatively in the "worst case scenarios", the Beta Regression and Beta Clustering models fare poorly. The best models in this regard are the Look-Up and LOCF models, which have few red points. In fact, Figures 2.11 and 2.17 suggest that for these high values, these models almost always produce relatively small prediction errors.

Lastly, Figure 2.15 is slightly misleading, as many of the prediction errors overlap. Figure 2.16 jitters these prediction errors, so that one may see the phenomenon seen earlier in the densities, where many prediction errors are nearly identical. This same feature holds for the LOCF model as well. Though these models produce prediction errors tightly centered around zero, they still produce a fair number of large prediction errors (many are 4 or greater in magnitude). Generally, predictions which are within one or two orders of magnitude are considered useful, so these models are not necessarily guaranteed to be operationally useful, though they seem to be more promising than the others.

Figures 2.18-2.21 plot ROC curves for all of the models. These ROC curves present sensitivity and specificity for prediction if the final $\log P_c$ value will be above

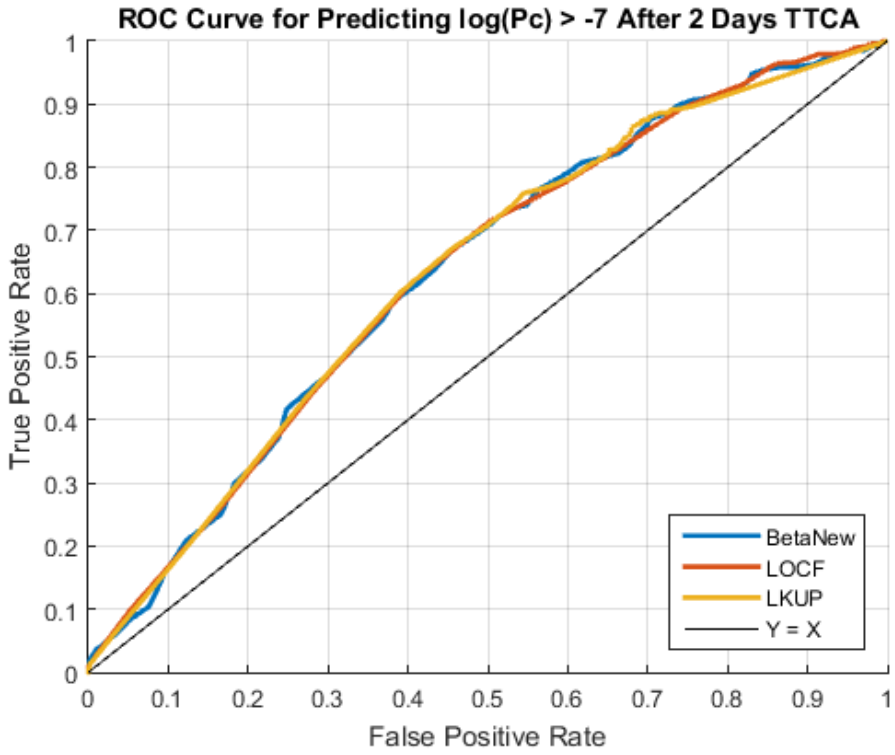

Figure 2.18. ROC curve for classifying final $\log_{10} P_c > -7$ (best models)

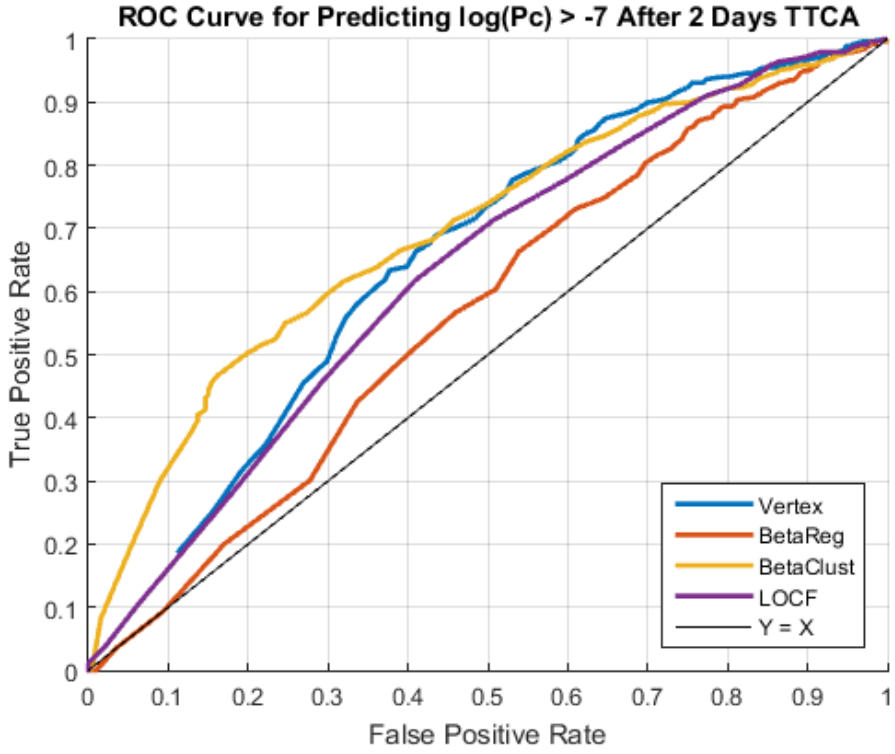

Figure 2.19. ROC curve for classifying final $\log_{10} P_c > -7$ (worst models)

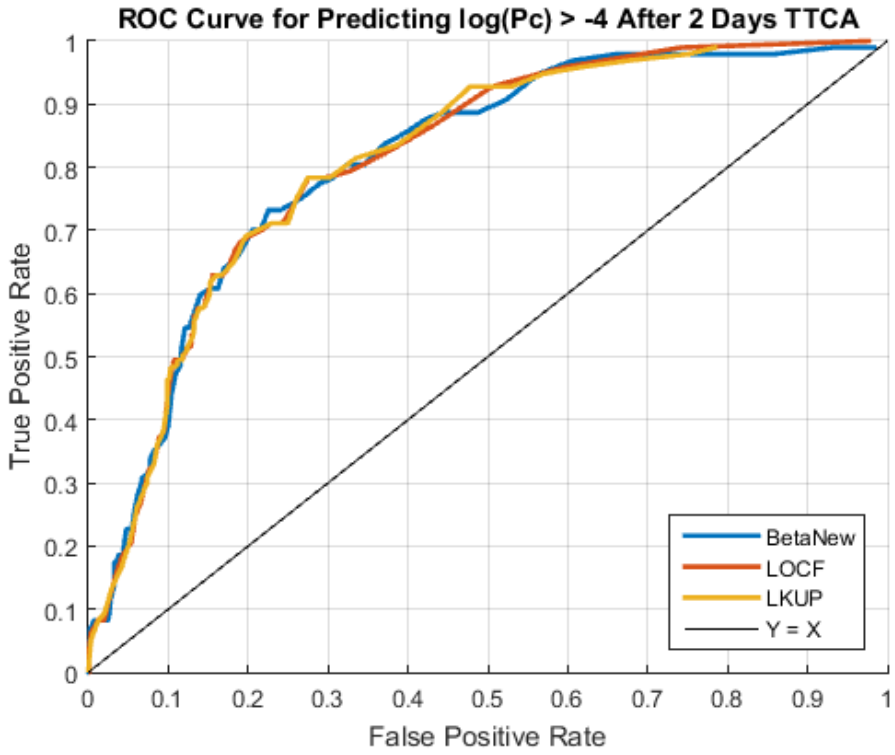

Figure 2.20. ROC curve for classifying final $\log_{10} P_c > -4$ (best models)

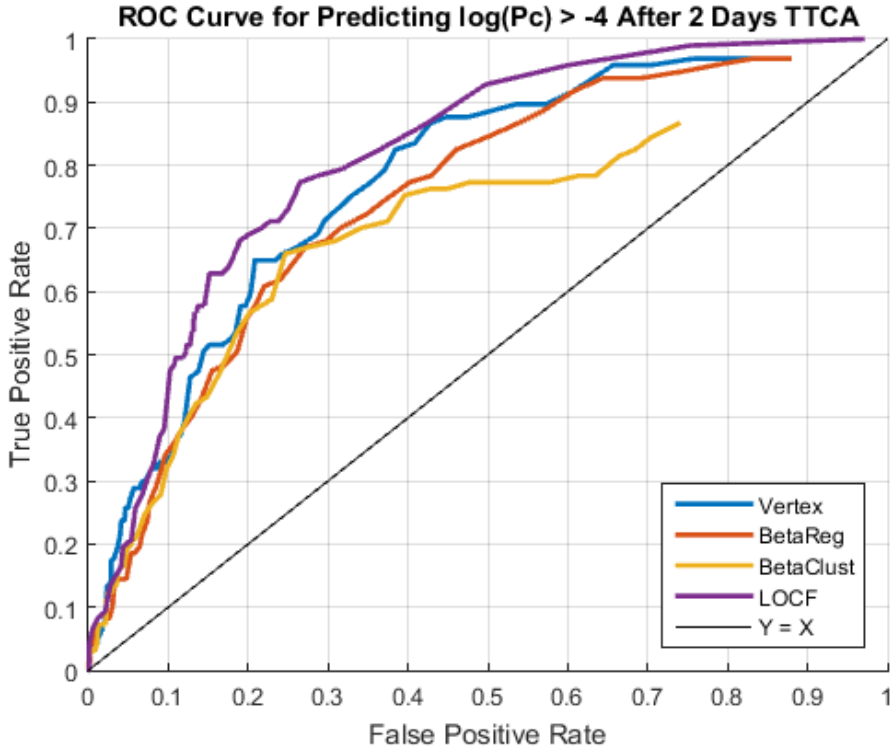

Figure 2.21. ROC curve for classifying final $\log_{10} P_c > -4$ (worst models)

-7 after 2 days to TCA. Recall that these simulation results are for the Yellow data set, so that one is primarily interested in knowing whether the $\log P_c$ will be above -7, and if it isn't one can "downgrade" the risk to low (the Green group). However, one may also be interested in knowing that this is still a medium-risk event and not a high-risk event, so we also consider prediction of values above -4.

Figures 2.18 and 2.20 plot the ROC curves for the Look-Up, LOCF, and New Beta regression models, and the results are nearly identical. For clarity, the results for the remaining models are graphed in Figures 2.19 and 2.21.

Figure 2.19 suggests that the Beta Clustering model is best for determining whether the final $\log_{10} P_c$ value will be above -7 if one can accept a true positive rate of 60% or lower, as it generates the fewest false positives. As we are concerned with being highly certain that the value will be above -7, this feature is not particularly useful. For a high true positive rate (90% and above), the models perform quite similarly, though the Beta regression model is somewhat inferior here. Note that, though the Look-Up method has a line in this region, no actual values occurred in this region, as it is simply connecting the point at (0.70, 0.88) with (1, 1). This feature of the Look-Up method is troubling, as it implies that one cannot easily implement a model with a true positive rate of 95% (or higher). This may be due to the way the confidence intervals were constructed. We consider this in future work.

Figures 2.18-2.21 showed the operating characteristics for the models when determining if the *last* $\log_{10} P_c$ value would be above -7 or -4. Figures 2.22-2.25 plot these operating characteristics for the models when the objective is to determine if the *next* $\log_{10} P_c$ value will be above -7 or -4 (given the next value occurs after 2 days to TCA). When the threshold is -7, many of these models are only slightly better than guessing, as evidenced by Figures 2.22 and 2.23. The Beta clustering model performs well here for lower true positive rates, though as mentioned above, this is not operationally useful. This model performs poorly when the threshold is

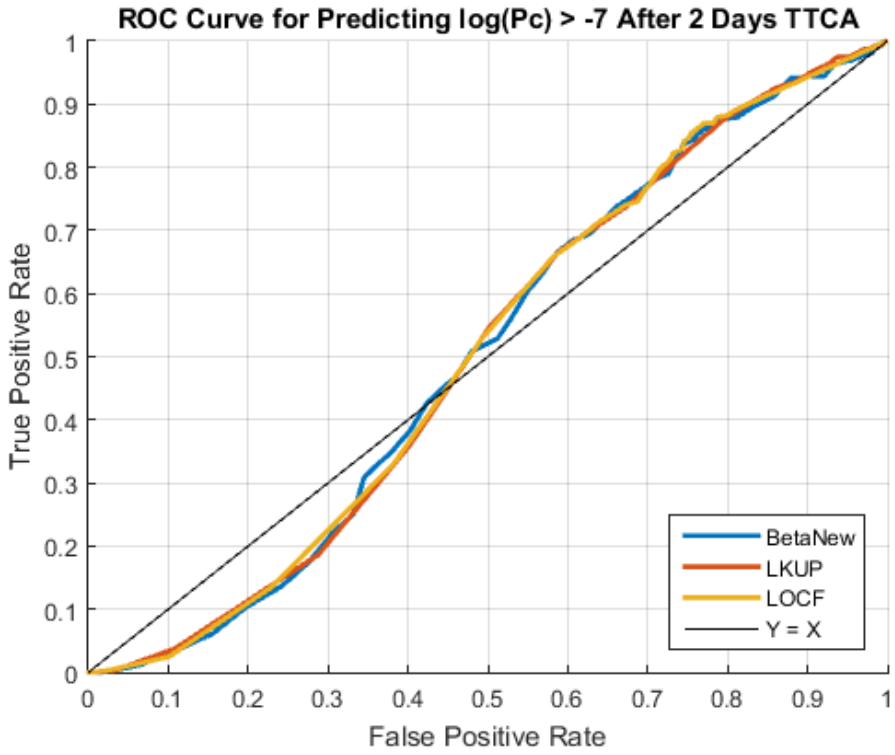

Figure 2.22. ROC curve for classifying next $\log_{10} P_c > -7$ (best models)

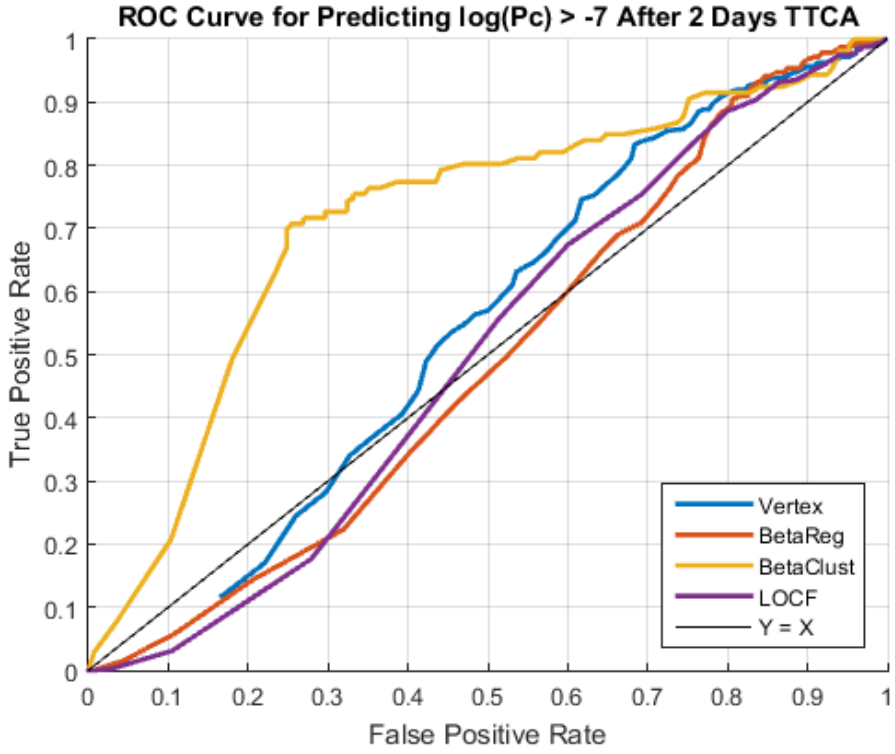

Figure 2.23. ROC curve for classifying next $\log_{10} P_c > -7$ (worst models)

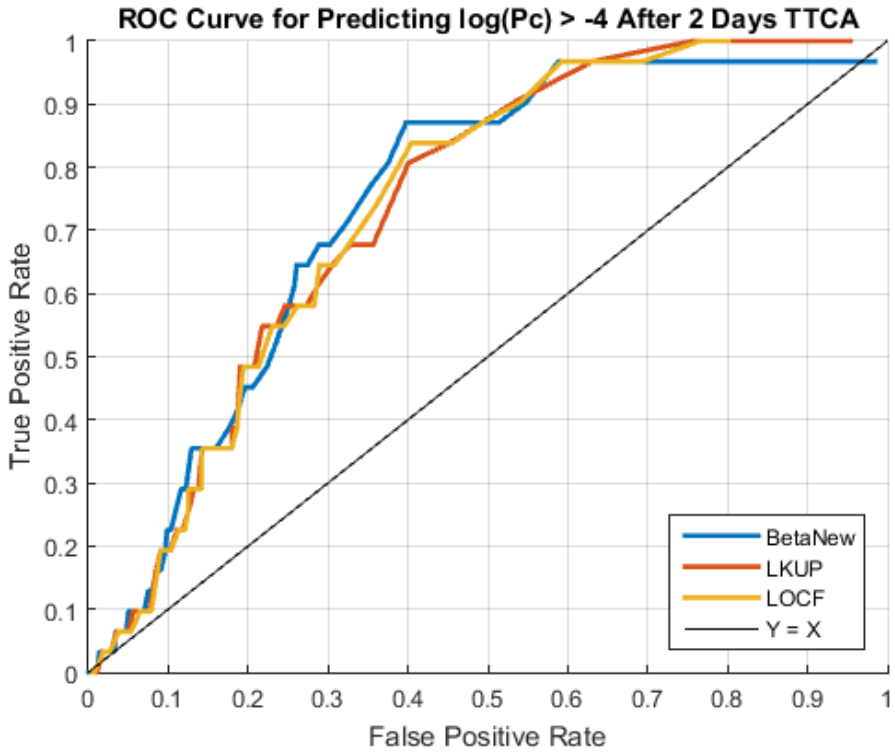

Figure 2.24. ROC curve for classifying next $\log_{10} P_c > -4$ (best models)

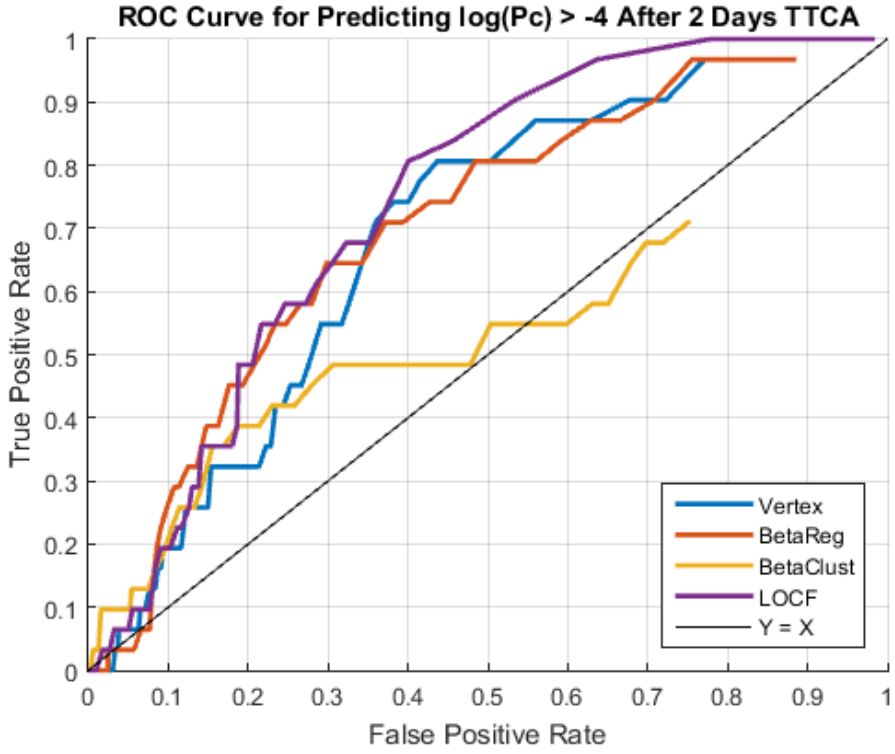

Figure 2.25. ROC curve for classifying next $\log_{10} P_c > -4$ (worst models)

-4. More investigation is needed to ascertain whether a model such as this could ever be operationally useful.

Overall, Figures 2.22-2.25 support the earlier conclusion that the three best models are the Look-Up, LOCF, and New Beta regression models. Not surprisingly, these models all share the same characteristic: they all utilize the most recent $\log_{10} P_c$ value. Though it's not particularly surprising that this value is useful, it *is* surprising that accounting for the trend over time doesn't seem to provide any noticeable improvement over simply using the previous value and ignoring the trend. This may be due to the fact that the trend is small and an event generally has only a few observations.

2.5 Conclusions and Future Work

We introduced a number of models for making predictions about future $\log P_c$ values. These models have various advantages and disadvantages, but all of the models produce unbiased predictions over time. Additionally, these predictions are within an order of magnitude at least 60% of the time, with the best models produce predictions within an order of magnitude 85% of the time. This is counter to the conventional wisdom that $\log P_c$ values cannot be predicted and that we can make no claims about future behavior. Indeed, we can make relatively strong claims about most of their future values, though this is more difficult as one approaches TCA. Still, these predictions are good enough to create decision-making frameworks that are better than guessing, and reliable enough to give us a high degree of confidence in saying whether a future CDM will contain high-risk values or not.

These models all reveal various features about $\log P_c$ values which had not been discussed before. For instance, the Beta Clustering model shows that it is likely that there are two kinds of events, those which have very low variability and those which have very high variability. This information may be used in turn to

diagnose what might be different about these events with low variability, and if perhaps the information for these events is simply more accurate. The Look-Up model suggests that the quantile of the $\log P_c$ value tells us a great deal about future values. Furthermore, it reveals that this holds true to a very high degree up to 4 days to TCA. And though the Vertex model is simple and does not utilize a longitudinal framework, it still has many nice properties. Fortunately, many simple models work well here, which eases computation time and interpretation.

It is worth noting that these models are simply an initial exploration into the problem of predicting $\log P_c$ values. Future work may hone these models to be more accurate, and ultimately to be more useful in the decision-making process. For instance, it seems clear from the ROC curves that some more exploration should be done in constructing confidence intervals for the Look-Up method. The Beta Clustering model suggests that perhaps identifying a particular cluster early will make prediction easier. More specifically, it may be worth exploring models which take into account not only the quantiles of the previous observations, but also the change in quantiles, as the low-variability cluster suggests the quantile does not change much over time. The New Beta Regression model suggests that a cubic trend is likely better for describing the mean of the $\log P_c$ values over time, as is supported by the discussion leading to the Beta Threshold model.

Due to the success of the Look-Up and LOCF models, it seems that future work should focus on non-parametric procedures. Though Beta regression is flexible, many of these features (such as the threshold feature) are more easily described through splines or quantiles. A sufficiently flexible non-parametric model which can account for quantile, time, and longitudinal observations may very well improve upon all of these models. Additionally, a non-parametric framework may more easily accommodate the identification of clusters. In addition to focusing on non-parametric models, future work may also consider other features in the data which might help

to explain the variability. This may include the positional error covariance matrix, tracking information, or miss distance. In initial exploration, we found that many of these covariates are noisy and may add more noise than the amount of variability in $\log P_c$ they describe. However, there may be a simple way to incorporate one or more of these features into a model so that prediction is improved.

CHAPTER THREE

Longitudinal Network Meta-Analysis

3.1 Background

Meta-analysis provides a framework for combining evidence from multiple data sources with the goal of improving inference. Meta-analyses are common in biostatistics, as researchers often attempt to combine the results from multiple clinical trials to assess the relative efficacy of two or more treatments. In addition to standard instructional texts now available (for example, Cooper[22], Higgins[41], or Egger[30]), the National Institute for Health and Care Excellence (NICE) attempted to standardize some of the practices in meta-analysis in a series of seven technical documents. The NICE documents include suggestions focused on network meta-analysis. Network meta-analysis attempts to compare treatments across trials, even when they were not observed head-to-head in a clinical trial.

The NICE documents consider network meta-analysis techniques for response at a single time point. In this chapter, we review some of the methods proposed for network meta-analysis of multiple time points. In addition to reviewing existing methods, we alter and propose some methods to extend existing longitudinal methods for meta-analysis to include a network of treatments. There are two main motivations for developing longitudinal models in this area. The main motivation is to provide a better framework for imputation when one does not have data for all trials at a given time point. Current methods of imputation are linear interpolation, last one carried forward (LOCF), or simply removing trials without the time point of interest. It is easy to think of examples where each one of these approaches is inappropriate. The second motivation for developing longitudinal methods for meta-analysis is to understand the trend over time, and possibly make prediction

at future time points. Understanding the trend of a treatment response over time is immensely useful for planning a trial, and ultimately could yield more powerful designs for future clinical trials.

Jones et al.[46] reviewed the literature for methods of meta-analyzing longitudinal data. Their conclusion was that practitioners were undecided on how to appropriately meta-analyze longitudinal studies, and subsequently Jones et. al proposed a few methods for doing so. Since then, various new methods have been proposed for the meta-analysis of longitudinal data, though few of these explicitly mention longitudinal data. In addition to the research which explicitly references longitudinal data, there also exists developments in the areas of meta-analysis for repeated measures, model-based meta-analysis, and multivariate meta-analysis. Thus, though models have been created and tested in each of these areas, there has been no exploration of the performance of these models against each other in various longitudinal settings. Furthermore, due to this lack of exploration and a cohesive framework, practitioners may still feel unclear in how to proceed with a network meta-analysis of longitudinal data. We seek to address these issues in this paper.

Ishak[42] proposed a general linear mixed model for the meta-analysis of longitudinal studies. This model did not support multiple treatment comparisons, but was one of the first proposed for explicitly meta-analyzing longitudinal data. What Ishak referred to as a general linear mixed model is also known as a multivariate mixed model. Many others proposed meta-analysis methods for nominal repeated measures are a special case of multivariate mixed models, for example Dakin[23], who presented a network meta-analysis for repeated measures. In addition to the special cases explored in meta-analysis of repeated measures, the multivariate mixed model is also the basis of inference for the meta-analysis of multiple outcomes (also referred to as multivariate responses). This area has received a great deal of attention in recent years, hinging on the development of suitable priors for covariance

matrices of multivariate outcomes known to be positively correlated. These developments could be incorporated in the meta-analysis of repeated measures, as discussed below.

Model-based meta-analysis is generally used to model dose-response over time, thus contributing to meta-analysis of longitudinal data. A common model for the dose-time relationship is the Emax model, as exemplified by Ahn and French[3]. Gross et. al[37] utilized this structure to compare various dose levels of two comparator drugs across studies, while accounting for various covariates and washout periods. Though the Emax model is common, one can find other examples of model-based meta-analyses in the literature, usually with a nonlinear dose effect over time, as in Mercier[56].

Methods proposed explicitly for the meta-analysis of longitudinal data tend to be less common, with notable exceptions. The Emax model proposed by Gross et al.[37] explicitly mentions longitudinal data. In addition, Ding and Fu[29] propose a model which simultaneously models the mean and the variance of the dose effect over time. Recently, Jansen[44] proposed a method using fractional polynomials to address the case of unknown treatment effect over time.

The purpose of this paper is two-fold. First and foremost, we present four models commonly found in the literature for network meta-analysis of longitudinal data. These models are Ding and Fu's BEST-ITP model, an Emax model parameterized by time (as opposed to dose level), a multivariate mixed model, and Jansen's fractional polynomial model. Where needed, we extend these methods to support mixed treatment comparisons, and offer other modeling suggestions. Second, we present a simulation comparing these four models with a univariate network meta-analysis, so that practitioners can see the impact of model misspecification, as well as when longitudinal models are advantageous. To our knowledge, such an exploration of the effects of model misspecification has never been done in this setting.

Furthermore, few[29] have shown the gains achieved when specifying a longitudinal model rather than a univariate model.

This chapter is structured as follows. Section 2 presents an overview of the four models mentioned above. Section 3 discusses the simulation structure and results. Section 4 offers conclusions and recommendations for model choice, as well as thoughts for future research.

3.2 Models

In what follows, let i denote the study index, j denote the treatment index, and k denote the time index.

3.2.1 Univariate Model

Before the longitudinal models, we present the network meta-analysis model for a single time point as presented in Dias[26]. For normally distributed data, the univariate network meta-analysis model can be written as

$$\begin{aligned} y_{ij} &\sim N(\theta_{ij}, se_{ij}) \\ \theta_{ij} &= \mu_i + d_{(bj)} \\ d_{(b1)} &= 0, \end{aligned}$$

where se_{ij} is the estimated standard error of response y_{ij}, b is the baseline treatment, i is the study index, and j is the treatment index. This model is known as the "fixed effects model", as the treatment effects $d_{(bj)}$ are fixed across studies. The model above assumes that each study has the same baseline treatment b, which has mean μ_i for that particular study. This is easily alleviated in practice by employing the network assumption,

$$d_{(bc)} = d_{(ac)} - d_{(ab)}. \tag{3.1}$$

Equation (3.1) is known as the consistency equation. For notational convenience, we assume that all studies have the same baseline treatment. We detail a general model with varying baseline treatments later. It is worth noting that a model which assumes the same baseline treatment for each study is relatively common, e.g. the case of each study comparing treatments to a placebo.

We can extend the previous model to include random effects. We have

$$\begin{aligned} \bar{y}_{ij} &\sim N(\theta_{ij}, se_{ij}^2) \\ \theta_{ij} &= \mu_i + \delta_{i(bj)} \\ \delta_{i(b1)} &= 0 \end{aligned}$$

where

$$\delta_{i(bj)} \sim N(d_{(bj)}, \sigma^2).$$

In a Bayesian model, we put priors on $d_{(bj)}$ and σ^2.

When one specifies random effects, they are correlated. Let J be the total number of treatments.

$$\begin{pmatrix} \delta_{i(b2)} \\ \vdots \\ \delta_{i(bJ)} \end{pmatrix} \sim MVN \left(\begin{pmatrix} d_{(b2)} \\ \vdots \\ d_{(bJ)} \end{pmatrix}, \Lambda \right) \tag{3.2}$$

$$\Lambda = \begin{pmatrix} \sigma^2 & \sigma^2/2 & ... & \sigma^2/2 \\ \sigma^2/2 & \sigma^2 & ... & \sigma^2/2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2/2 & \sigma^2/2 & ... & \sigma^2 \end{pmatrix} \tag{3.3}$$

A general model allowing for different baseline treatments in each study can be constructed by employing the consistency assumption. Recall that we have assumed

the network of treatments is consistent, so that $d_{(cj)} = d_{(bc)} - d_{(bj)}$. Then we can write the random effects in terms of any new baseline treatment c in a study

$$\begin{pmatrix} \delta_{i(c2)} \\ \vdots \\ \delta_{i(cJ)} \end{pmatrix} \sim MVN \left(\begin{pmatrix} d_{(bc)} \\ \vdots \\ d_{(bc)} \end{pmatrix} - \begin{pmatrix} d_{(b2)} \\ \vdots \\ d_{(bJ)} \end{pmatrix}, \Lambda \right)$$

$$\Lambda = \begin{pmatrix} \sigma^2 & \sigma^2/2 & ... & \sigma^2/2 \\ \sigma^2/2 & \sigma^2 & ... & \sigma^2/2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2/2 & \sigma^2/2 & ... & \sigma^2 \end{pmatrix}.$$

For the longitudinal models, we assume that treatment 1 is the baseline treatment, and that it is the same across trials. This is to ease notation. When this is not true, the modeler can utilize the consistency equation to rewrite the model as we have above.

3.2.2 BEST-ITP Model

Fu and Manner[34] developed an integrated two part (ITP) model for Bayesian adaptive design with delayed responses. Ding et. al[29] extended this model to network meta-analysis by adding a model for the overall residual variance. This model is the BEST-ITP model, and is given in equation (8) of Ding et. al. This model has the overall property that

$$\bar{Y}_{ijk} = \left(\phi_i + \theta_j + \frac{\epsilon_{ijk}}{\sqrt{n_{ijk}}} \right) \frac{1 - e^{p_j t_{ijk}}}{1 - e^{p_j d}}, \quad Var(\epsilon_{ijk}) = \sigma^2. \tag{3.4}$$

This model can be written as

$$\begin{aligned} \bar{Y}_{ijk} &\sim N(\mu_{ijk}, \sigma^2_{ijk}) \\ \frac{S^2_{ijk}(n_{ijk}-1)}{\sigma^2} \left(\frac{1 - e^{p_k t_{ijk}}}{1 - e^{p_j d}} \right)^2 &\sim \Gamma\{(n_{ijk}-1)/2, 2\}, \\ \mu_{ijk} &= \left(\phi_i + \theta_j + \frac{\epsilon_{ijk}}{\sqrt{n_{ijk}}} \right) \frac{1 - e^{p_j t_{ijk}}}{1 - e^{p_j d}} \end{aligned}$$

$$\sigma^2_{ijk} = \frac{\sigma^2}{n_{ijk}} \left(\frac{1 - e^{p_j t_{ijk}}}{1 - e^{p_j d}} \right)^2,$$

$$\theta_1 = 0,$$

where $p_j \in (-\infty, 0]$ for all j, and $d = \max(t_{ijk})$. Notice that as t_{ijk} increases, so too does the sample variance. This model assumes that the dose response plateaus, and has maximal magnitude when $t_{ijk} = d$. As Ding notes, one can make the model flexible by replacing

$$\frac{1 - e^{p_j t_{ijk}}}{1 - e^{p_j d}}$$

with

$$\left(\frac{1 - e^{p_j t_{ijk}}}{1 - e^{p_j d}} \right) \left(\frac{1 + e^{a_j + p_j d}}{1 + e^{a_j + p_j t_{ijk}}} \right),$$

where $a_j \in (-\infty, 0]$ for all j. This model still assumes a plateau of dose response, but supports an "S"-shape in the longitudinal trajectory. Of course, if one has only a few time points for a given study (2 or 3), the added parameter a_j can cause identifiability issues.

The rate parameters a_j and p_j may be problematic if one observes significant variability in the shape of a trajectory for a given treatment j from trial to trial. This is because these effects are fixed, and there has been no proposed method of implementing them as random. Specifying random effects in θ_j is straightforward, as one simply replaces θ_j with θ_{ij}, so that

$$\bar{Y}_{ijk} = \left(\phi_i + \theta_{ij} + \frac{\epsilon_{ijk}}{\sqrt{n_{ijk}}} \right) \frac{1 - e^{p_j t_{ijk}}}{1 - e^{p_j d}}, \quad Var(\epsilon_{ijk}) = \sigma^2,$$

where the θ_{ij} have the correlation structure given in (3.3). However, the extension to random effects for a_j and p_j is not so obvious, as one generally gives each a non-informative Uniform prior distribution, for example

$$a_j \sim U(-100, 0)$$

$$p_j \sim U(-100, 0),$$

for all j. One solution may be to specify a Normal prior on a suitable transformation $h(\cdot)$ of a_j and p_j, such as the natural log or square root of $-a_j$ or $-p_j$. Then one could proceed as usual to construct prior distributions for a_{ij} and/or p_{ij}. Of course, this creates more parameters to estimate and possibly new identifiability issues, so this should only be done if there is sufficient data and the data supports such added complexity. In addition, it may not be straightforward to specify a non-informative prior distribution for the variance of such random effects.

3.2.3 Emax Model

Consider an Emax model, which models dose response in the following way

$$Y = \frac{Emax \times t}{t + ED50} + \epsilon, \tag{3.5}$$

where $\epsilon \sim N(0, \sigma^2)$. Note that this model usually uses dose instead of time as the independent variable. In this setting, the $ED50$ represents the dose at which 50% of the effect is obtained. When dose is replaced with time, as above, the $ED50$ represents the time at which 50% of the effect is obtained. In either case, the $Emax$ can be interpreted as the maximum possible effect.

The $Emax$ and $ED50$ parameters may depend on both the study and the treatment. That is, for the i^{th} study, j^{th} treatment, and k^{th} time, one might have

$$Y_{ijk} = \frac{Emax_{ij} \times t_{ijk}}{t_{ijk} + ED50_{ij}} + \epsilon_{ijk}$$

One may specify a treatment effect for Emax only, ED50 only, or both.

In the literature, there are at least two ways one might specify an effect on these parameters. Mandema et. al[55] specify random and covariate effects to these parameters via the relationship

$$Emax_i = Emax \times (1 + \beta(x_{ij} - \tilde{x}_{..}) + \eta_i). \tag{3.6}$$

where $\tilde{x}_{\cdot\cdot}$ is the median of the covariate values, and η_i is a random effect.

Gross et. al[37] define covariate effects on the Emax parameter as

$$Emax = Emax^{'} \times \exp\left\{\beta(x_{ij} - \bar{x}_{\cdot\cdot})\right\}, \tag{3.7}$$

where $\bar{x}_{\cdot\cdot}$ is the mean of the covariate values and $Emax^{'}$ is the Emax unadjusted for covariates. This suggests that a random effect might enter via

$$Emax_i = Emax \times \exp\left\{\beta(x_{ij} - \bar{x}_{\cdot\cdot}) + \eta_i\right\}. \tag{3.8}$$

Let $Emax_0$ be the $Emax$ of the reference group and denote study and treatment effects by α_i and β_j, respectively. As demonstrated above, there are at least two ways we may specify study and treatment effects, the first being due to Mandema[55]

$$Emax_{ij} = Emax_0 \times (1 + \alpha_i + \beta_j), \tag{3.9}$$

and the second being exponential random effects, which have the form

$$Emax_{ij} = Emax_0 \times \exp\left\{\alpha_i + \beta_j\right\}. \tag{3.10}$$

A complete Bayesian specification can be completed by placing normal priors on the effects and uniform priors on $Emax_0$ and $ED50_0$.

3.2.4 Multivariate Mixed Model

In some cases, the modeler may not wish to specify a trend for time in the model, preferring to analyze the observations over time as a multivariate observation. The following development draws on the literature concerning meta-analysis for multiple outcomes.

Ishak et. al[42] consider the following model. Suppose K measurements are taken over time on N units (e.g. studies in a meta-analysis); we denote by $\boldsymbol{y}_i$ the $K \times 1$ vector of observed values from the i^{th} trial and y_{ij} the measurement taken

at the j^{th} time. A general linear mixed model that can account for the correlations between the observations is given by

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\theta} + \boldsymbol{Z}_i\boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{X}_i$ is a $K \times p$ matrix of possibly time dependent covariates, $\boldsymbol{\theta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{Z}_i$ is a $K \times q$ matrix of covariates, $\boldsymbol{\delta}_i$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\epsilon}_i$ is a $K \times 1$ vector of residuals.

This model framework, known as a multivariate mixed model, can be extended to specify treatments within each study, so that one may perform a network meta-analysis. How this is done depends on what one specifies as the response variable. The literature contains examples of modeling the mean directly (e.g. mean change in baseline for each treatment), as well as of modeling the mean effect relative to a reference treatment (i.e. treatment effect). We treat only the direct modeling of the mean here, and refer readers to Ethfimiou et. al (2015) as a reference for modeling the treatment effect.

3.2.4.1 Modeling the Mean Directly Achana et. al[2] consider the following model for meta-analysis of multiple outcomes. Again, let y_{ijk} represent an observation (such as change from baseline) from subject i in treatment arm j at time k. Then

$$\begin{pmatrix} y_{ij1} \\ \vdots \\ y_{ijK} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \theta_{ij1} \\ \vdots \\ \theta_{ijK} \end{pmatrix}, \Sigma_{ij} = \begin{pmatrix} se_{ij1}^2 & \dots & r_{ik}^{1K} se_{ij1} se_{ijK} \\ & \ddots & \vdots \\ & & se_{ijK}^2 \end{pmatrix}\right)$$

$$\begin{pmatrix} \theta_{ij1} \\ \vdots \\ \theta_{ijK} \end{pmatrix} = \begin{pmatrix} \mu_{b1} + \phi_{i1} + \delta_{i(bj)1} \\ \vdots \\ \mu_{bK} + \phi_{iK} + \delta_{i(bj)K} \end{pmatrix}$$

$$\begin{pmatrix} \delta_{ib1} \\ \vdots \\ \delta_{ibK} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

where $\delta_{i(bj)k}$ is the random effect of the j^{th} treatment at time k relative to treatment b. Thus, b is the reference treatment. The vector of effects due to treatment is distributed $\boldsymbol{\phi} \sim MVN(0, \Gamma)$. Note that while one generally has y_{ijk} and se^2_{ijk} in a meta-analysis, one usually does *not* have access to r^{ik}_{ij}.

Because this model is multivariate, the random effects will be correlated across time in addition to across treatments. For instance, given study i and treatment j, one has

$$\begin{pmatrix} \delta_{i(bj)1} \\ \vdots \\ \delta_{i(bj)K} \end{pmatrix} \sim MVN \left(\begin{pmatrix} d_{(Aj)1} \\ \vdots \\ d_{(Aj)K} \end{pmatrix} - \begin{pmatrix} d_{(Ab)1} \\ \vdots \\ d_{(Ab)K} \end{pmatrix}, \Delta_{(bj)\cdot} \right)$$

$$\Delta_{(bj)} = \begin{pmatrix} \tau^2_{(bj)1} & \dots & \rho^{1K}_{bj} \tau_{(bj)1} \tau_{(bj)K} \\ & \ddots & \vdots \\ & & \tau^2_{(bj)K} \end{pmatrix},$$

where ρ^{ik}_{bj} captures the across-time correlation. For identifiability, one generally assumes that a common between-study variance among treatment arms, so that $\sigma^2_{(bj)} = \sigma^2$. Additional, one assumes that the correlation is constant across treatments, so that $\rho^{ik}_{bj} = \rho^{ik}$. As a result, one has

$$\begin{pmatrix} \delta_{i(bj)1} \\ \vdots \\ \delta_{i(bj)K} \end{pmatrix} \sim MVN \left(\begin{pmatrix} d_{(Aj)1} \\ \vdots \\ d_{(Aj)K} \end{pmatrix} - \begin{pmatrix} d_{(Ab)1} \\ \vdots \\ d_{(Ab)K} \end{pmatrix}, \Delta \right)$$

$$\Delta = \begin{pmatrix} \tau^2_1 & \dots & \rho^{1M} \tau_1 \tau_K \\ & \ddots & \vdots \\ & & \tau^2_K \end{pmatrix}$$

Because the random effects are with respect to the same reference treatment within each study, they must also be correlated across treatments, as in equation (3.3). Similarly, one has

$$\begin{pmatrix} \begin{pmatrix} \delta_{i(bj_1)1} \\ \vdots \\ \delta_{i(bj_1)K} \end{pmatrix} \\ \begin{pmatrix} \delta_{i(bj_2)1} \\ \vdots \\ \delta_{i(bj_2)K} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \delta_{i(bj_P)1} \\ \vdots \\ \delta_{i(bj_P)K} \end{pmatrix} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \begin{pmatrix} d_{(bj_1)1} \\ \vdots \\ d_{(bj_1)K} \end{pmatrix} \\ \begin{pmatrix} d_{(bj_2)1} \\ \vdots \\ d_{(bj_2)K} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} d_{i(bj_P)1} \\ \vdots \\ d_{i(bj_P)K} \end{pmatrix} \end{pmatrix}, \Pi \right) \tag{3.11}$$

$$\Pi = \begin{pmatrix} \Delta & \frac{1}{2}\Delta & ... & \frac{1}{2}\Delta \\ \frac{1}{2}\Delta & \Delta & ... & \frac{1}{2}\Delta \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\Delta & \frac{1}{2}\Delta & ... & \Delta \end{pmatrix}, \tag{3.12}$$

so that the vector of random treatment effects is equally correlated among all combinations of treatments.

3.2.4.2 Structured Covariance Matrix As noted above, one generally has more than two time points to consider in longitudinal data. As a result, the covariance matrices for both the random effect and error vectors contain multiple correlation coefficients. To make these matrices identifiable, one can impose a structure on the covariances to reduce the number of parameters. Ishak et. al[42] used this approach in a meta-analysis of longitudinal data which compares a single treatment across studies. They proposed using an AR(1) covariance structure, as it only

requires the estimation of one correlation coefficient for each matrix. The resulting covariance matrices for the error and random effect vectors are

$$\Sigma_{ij} = \begin{pmatrix} s_{ij1}^2 & \dots & r^K s_{ij1} s_{ijK} \\ & \ddots & \vdots \\ & & s_{ijK}^2 \end{pmatrix}$$

$$\Delta = \begin{pmatrix} \tau_1^2 & \dots & \rho^K \tau_1 \tau_K \\ & \ddots & \vdots \\ & & \tau_K^2 \end{pmatrix}.$$

This kind of correlation structure is tenable if the time between observations is equally spaced. One may consider other simple correlation structures, such as a one- or two-banded Toeplitz correlation structure[42].

If the time between successive observations is *not* equally spaced, then one may consider defining a correlation coefficient which is dependent on time. These covariance structures are commonly found in spatial statistics. These correlation structures include exponential, Gaussian, and spherical correlations, to name a few. To simplify the analysis and reduce the number of parameters, one might assume that $r^K = \rho^K$ for all K, so that the correlation of the errors across time is the same as the correlation of the treatment effects across time. One may assume this same correlation for the study effects, so that $\Gamma = S_\phi R S_\phi'$. Though this is a strong assumption, without patient-level data it is not possible to model these correlations. Thus, they are not identifiable for aggregate-level data, and a particularly strong assumption must be made when implementing this kind of model.

3.2.4.3 Unstructured Covariance Matrix Wei and Higgins[80] discuss a prior scheme for unstructured covariance matrices. These priors are intended for multiple outcomes, but can be adapted for longitudinal data, as in both cases correlation is typically assumed to be positive.

3.2.5 Fractional Polynomials Model

The nature of the underlying trend for the mean is often not known. Jansen et. al[44] proposed using fractional polynomials (Royston and Altman[68]) to address this issue.

A first-order fractional polynomial is obtained by describing the outcome of interest as a function of transformed time t in a linear model:

$$\theta_t = \beta_0 + \beta_1 t^p. \tag{3.13}$$

The power p is chosen from the following set: -2, -1, -0.5, 0, 0.5, 1, 2, 3 with $t^0 = \log t$.

A second-order fractional polynomial is defined as:

$$\theta_t = \beta_0 + \beta_1 t^p + \beta_2 t^{p_2}. \tag{3.14}$$

If $p_1 = p_2 = p$, the model becomes a "repeated powers" model:

$$\theta_t = \beta_0 + \beta_1 t^p + \beta_2 t^p \log t. \tag{3.15}$$

3.2.5.1 Random effects. As in the other models, one may model the treatment effects as fixed or random. Below we give the model which assumes the treatment effects are random, and the fixed effects version follows as discussed above. Note that there are M treatment effects in the model, and that they do not necessarily all have to be either fixed or random. That is, one may have both fixed and random treatment effects. This is discussed further in Jansen[44]. The model proposed by Jansen is

$$\theta_{ijt} = \begin{cases} \beta_{0ij} + \sum_{m=1}^{M} \beta_{mij} t^{p_m} \text{ with } t^0 = \log(t) & \text{if } p_1 \neq ... \neq p_M \\ \beta_{0ij} + \beta_{1ij} t^{p_1} + \sum_{m=2}^{M} \beta_{mij} t^{p_1} (\log t)^{m-1} & \text{if } M > 1, p_1 = ... = p_M \end{cases}$$

$$\begin{pmatrix} \beta_{0ij} \\ \vdots \\ \beta_{Mij} \end{pmatrix} = \begin{pmatrix} \mu_{0ib} \\ \vdots \\ \mu_{Mib} \end{pmatrix} + \begin{pmatrix} \delta_{0i(bj)} \\ \vdots \\ \delta_{Mi(bj)} \end{pmatrix} \tag{3.16}$$

$$\begin{pmatrix} \delta_{0ib} \\ \vdots \\ \delta_{Mib} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.17}$$

$$\begin{pmatrix} \delta_{0i(bj)} \\ \vdots \\ \delta_{Mi(bj)} \end{pmatrix} \sim MVN \left(\begin{pmatrix} d_{0(Aj)} \\ \vdots \\ d_{M(Aj)} \end{pmatrix} - \begin{pmatrix} d_{0(Ab)} \\ \vdots \\ d_{M(Ab)} \end{pmatrix}, \Sigma \right) \tag{3.18}$$

$$\begin{pmatrix} d_{0(AA)} \\ \vdots \\ d_{M(AA)} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.19}$$

One can assume heterogeneity for any number of the d_{iAk}. For instance, assuming heterogeneity for d_{0Ak} only implies a random intercept model, implying between-study variance of effect estimates remains constant over time.

Because the random effects modify regression coefficients, it is likely that they are correlated with each other. If one specifies a second order model with correlation among the regression coefficients, then

$$\begin{pmatrix} \delta_{0ibk} \\ \delta_{1ibk} \\ \delta_{2ibk} \end{pmatrix} \sim MVN \left(\begin{pmatrix} d_{0Ak} \\ d_{1Ak} \\ d_{2Ak} \end{pmatrix} - \begin{pmatrix} d_{0Ab} \\ d_{1Ab} \\ d_{2Ab} \end{pmatrix}, \Delta = \begin{pmatrix} \sigma_0^2 & \rho_{01}\sigma_0\sigma_1 & \rho_{02}\sigma_0\sigma_2 \\ . & \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ . & . & \sigma_2^2 \end{pmatrix} \right)$$

Typically when one specifies heterogeneity for multiple treatment arms (> 2 treatments) within the same study, one assumes that the heterogeneity is the same for all $d_{mi(bk_1)}$, $d_{mi(bk_2)}$, ..., $d_{mi(bk_P)}$. For example,

$$\begin{pmatrix} \delta_{0i(bj_1)} \\ \vdots \\ \delta_{0i(bj_P)} \end{pmatrix} \sim MVN \left(\begin{pmatrix} d_{0(bj_1)} \\ \vdots \\ d_{0(bj_P)} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2/2 & ... & \sigma^2/2 \\ \sigma^2/2 & \sigma^2 & ... & \sigma^2/2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2/2 & \sigma^2/2 & ... & \sigma^2 \end{pmatrix} \right)$$

If we extend this notion of equal heterogeneity of random effects across studies, then we have

$$\begin{pmatrix} \begin{pmatrix} \delta_{0i(bj_1)} \\ \vdots \\ \delta_{Mi(bj_1)} \end{pmatrix} \\ \begin{pmatrix} \delta_{0i(bj_2)} \\ \vdots \\ \delta_{Mi(bj_2)} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} \delta_{0i(bj_P)} \\ \vdots \\ \delta_{Mi(bj_P)} \end{pmatrix} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \begin{pmatrix} d_{0i(bj_1)} \\ \vdots \\ d_{Mi(bj_1)} \end{pmatrix} \\ \begin{pmatrix} d_{0i(bj_2)} \\ \vdots \\ d_{Mi(bj_2)} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} d_{0i(bj_P)} \\ \vdots \\ d_{Mi(bj_P)M} \end{pmatrix} \end{pmatrix}, \Pi \right)$$

$$\Pi = \begin{pmatrix} \Delta & \frac{1}{2}\Delta & ... & \frac{1}{2}\Delta \\ \frac{1}{2}\Delta & \Delta & ... & \frac{1}{2}\Delta \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\Delta & \frac{1}{2}\Delta & ... & \Delta \end{pmatrix},$$

similar to equation (3.12).

3.2.6 Model Comparison

Table 3.1 provides a comparison of some of the defining characteristics among the models.

3.3 Comparison to Univariate Model

It is important to understand how these models compare to the standard network meta-analysis at a single time point. This is useful for understanding model assumptions and for exploring these assumptions visually.

Table 3.1: Comparison of models

Feature	BEST-ITP	Emax	Multi	Frac Poly
Assumes model for σ^2	Yes	No	No	No
Assumes response plateaus	Yes	Yes	No	No
Assumes monotonic response over time	Yes	Yes	No	No
Assumes univariate meta-analysis model at each time point	No	No	Yes	Yes
Treats missing time points as NA	No	No	Yes	No

3.3.1 BEST-ITP

The BEST-ITP model is a straightforward extension of the univariate model. The main innovation in the BEST-ITP model is the joint model for the variance. If we assume that the standard errors are fixed (as in the univariate model), we have

$$y_{ijk} = \left(\phi_i + \theta_{(bj)} + \epsilon_{ijk}\right) \frac{1 - e^{p_j t_{ijk}}}{1 - e^{p_j d}}$$

$$\epsilon_{ijk} \sim N(0, se_{ijk}),$$

which is a slight variant on the fixed effects univariate network meta-analysis model. Note that for the K^{th} time point, this is is the same as the univariate model **only** if $t_{ijk} = d$ or $p_j = p_k$ for all j, k.

Suppose that one observes the response at a single time point t, as in a univariate meta-analysis. We can simplify our model to

$$y_{ij} = \left(\phi_i + \theta_{(bj)} + \epsilon_{ij}\right) \frac{1 - e^{p_j t}}{1 - e^{p_j d}}$$

$$\epsilon_{ij} \sim N(0, se_{ij}).$$

Let $\kappa_j = \frac{1-e^{p_j t}}{1-e^{p_j d}}$. Then

$$y_{ij} = \phi_i \kappa_j + \theta^*_{(bj)} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, se_{ij}\kappa_j),$$

where $\theta^*_{(bj)} = \theta_{(bj)}\kappa_j$. From this equation, it is clear that the BEST-ITP model does not reduce to the univariate model when κ_j depends on j. It's worth noting

that when $t_{ijk} \geq d$, it *does* reduce to the univariate meta-analysis model, suggesting some guidance on choosing d. Furthermore, when κ_j is small (i.e. the responses have nearly plateaued), the model is approximately a univariate meta-analysis model at the given time point.

The fact that the BEST-ITP model does not reduce to a univariate model at each time point makes it more difficult to directly interpret differences in effects between the models at a given time point. On the other hand, it implies that when the model is true, the univariate model should **not** fit the data correctly.

The above development does imply at least one model checking technique. We can write the model as

$$\begin{aligned} y_{ijk} &= \phi_i \kappa_{jk} + \theta_{(bj)} \kappa_{jk} + \epsilon_{ijk} \kappa_{jk} \\ &= \phi^*_{ijk} + \theta^*_{jk} + \epsilon^*_{ijk} \end{aligned}$$

where $\kappa_{ijk} = \frac{1-e^{p_j t_{ijk}}}{1-e^{p_j d}}$. A plot of ϕ^*_{ijk} vs. θ^*_{ijk} should yield a straight line with a common slope of $\theta_{(bj)}/\phi_i$ for each study i and treatment j. Estimates for these effects can be easily obtained via least squares, which are invariant to heterogeneity of variance, as is the case here.

3.3.2 Emax

The Emax model reduces to the univariate model under certain conditions. In particular, if the $ED50$ is assumed to not vary among studies and treatments, *and* the effects are assumed to be multiplicative, then the Emax model reduces to a univariate meta-analysis at each time point. Under these conditions, we have

$$\begin{aligned} y_{ijk} &= \frac{Emax_{ij} \times t_{ijk}}{t_{ijk} + ED50} + \epsilon_{ijk} \\ &= \frac{Emax_0 \times (1 + \phi_i + \theta_j) \times t_{ijk}}{t_{ijk} + ED50} + \epsilon_{ijk}, \end{aligned}$$

which for a particular time point t_{ijk} we can write as

$$\begin{aligned} y_{ij} &= \frac{Emax_0 \times (1 + \phi_i + \theta_j) \times t}{t + ED50} + \epsilon_{ij} \\ &= (1 + \phi_i + \theta_j)\kappa + \epsilon_{ij} \\ &= \kappa + \phi_i^* + \theta_j^* + \epsilon_{ij}, \end{aligned}$$

where $\kappa = (Emax_0 \times t)/(t + ED50)$, $\phi_i^* = \phi_i\kappa$, and $\theta_j^* = \theta_j\kappa$. We can write this in the usual univariate meta-analysis form by letting $\mu_i = \kappa + \phi_i^*$. In this case, the effects of the longitudinal model are simply fixed or random intercepts, implying parallel trajectories along an Emax curve. When the effects are exponential, this is approximately true, as $\exp(\phi_i + \theta_j) \approx 1 + \phi_i + \theta_j$, given by the first order Taylor series expansion of $f(x) = \exp(x)$.

When the $ED50$ can not be assumed to be the same across studies and treatments, the Emax model does not reduce to the univariate model.

3.3.3 Mulviariate

The multivariate model reduces to the univariate model by definition, and allows for more general correlation structures.

3.3.4 Fractional Polynomials

The fractional polynomials model also reduces to the univariate model. For ease of notation, suppose one has a fractional polynomials model where $p_m \neq 0$ for all m and $p_1 \neq ... \neq p_M$. Then the model is

$$y_{ijk} \sim \beta_{0ij} + \sum_{m=1}^{M} \beta_{mij} t^{p_m} + \epsilon_{ijk}$$

$$\beta_{mij} = \mu_{mib} + d_{m(bj)} \text{ for } m = 1, ..., M,$$

when the treatment effects are fixed. Rearranging, the model becomes

$$y_{ijk} \sim \beta_{0ij} + \sum_{m=1}^{M} \mu_{mib} t^{p_m} + \sum_{m=1}^{M} d_{m(bj)} t^{p_m} + \epsilon_{ijk},$$

highlighting the study-specific and treatment-specific effects. For a given time point $t = t_{ijk}$, we can write

$$\begin{aligned} y_{ij} &\sim \beta_{0ij} + \sum_{m=1}^{M} \mu_{mib} t^{p_m} + \sum_{m=1}^{M} d_{m(bj)} t^{p_m} + \epsilon_{ij} \\ &= \beta_{0ij} + \mu^*_{ib} + d^*_{m(bj)} + \epsilon_{ij}, \end{aligned}$$

where $\mu^*_{ib} = \sum_{m=1}^{M} \mu_{mib} t^{p_m}$ and $d^*_{m(bj)} = \sum_{m=1}^{M} d_{m(bj)} t^{p_m}$. As before, we can write this in the usual univariate model with $\mu_i = \beta_{0ij} + \mu^*_{ib}$. Note that this holds for the more general case of when one or more of the powers p_m is zero or repeated.

3.4 Exploratory Analysis

Exploratory data analysis is an essential part of building a successful model. The models mentioned above can support many variants, making it difficult to choose both the form of the model as well as the class of the model. Here, we offer some tips on exploring the data to reveal patterns which can help one in making these decisions.

3.4.1 Variance Plots

Longitudinal studies report some form of an effect estimate at each time point, usually accompanied by some measure of variability of the estimate. This measure of variability is often the standard error, which is frequently treated as known in the subsequent meta-analysis. While this is convenient, this may not be realistic. For instance, many studies do not directly report a standard error of the effect estimate, instead providing only a graph of the effects over time with error bars. To include such studies in a meta-analysis, researchers often measure these error bars by hand. Thus, the data used in meta-analysis often contain estimates of variance estimates, suggesting the practice of treating these quantities as known as being inappropriate. As discussed later, the veracity of such estimates are often dubious.

When one chooses to consider the variance as a random variable, one must further choose how to model the variance over time. The BEST-ITP model automatically accounts for unknown variances, but this is not true for the other models presented above. To assess the model for the standard error, we recommend two plots. In order to discern the trend over time, we recommend a spaghetti plot of the sums of squares (SS) vs. time. This can reveal the trend of increasing variability, as is assumed in the BEST-ITP model.

Another plot which can be useful is a plot of SS vs. effect size. This is particularly useful in diagnosing the appropriateness of the BEST-ITP model, as this model implies a squared relationship between the effect size and the sums of squares when the sample size is constant. These two plots, along with a spaghetti plot of effects vs. time, are given in Figure 3.1. The model which generated this data is the BEST-ITP model, as one can discern by the increasing sums of squares over time and the parabolic relationship between effect and sums of squares.

3.4.2 Effect Plots

Here, we describe some useful plots for discerning the trend of the main effect over time.

3.4.2.1 Random Effects Model Discerning the functional form of the main effect can be difficult, as variability in the main effect can come from study effects, treatment effects, and pure error. The first step in choosing a model is often looking at a spaghetti plot of the main effects over time, which can be more illustrative when one colors the individual curves by treatment (as shown above). This can often be enough to narrow the choices down to one or two classes of models. However, further inspection can be useful in inspecting the sources of variability in the data, which have implicit relationships in each of the models. To explore these relationships, we

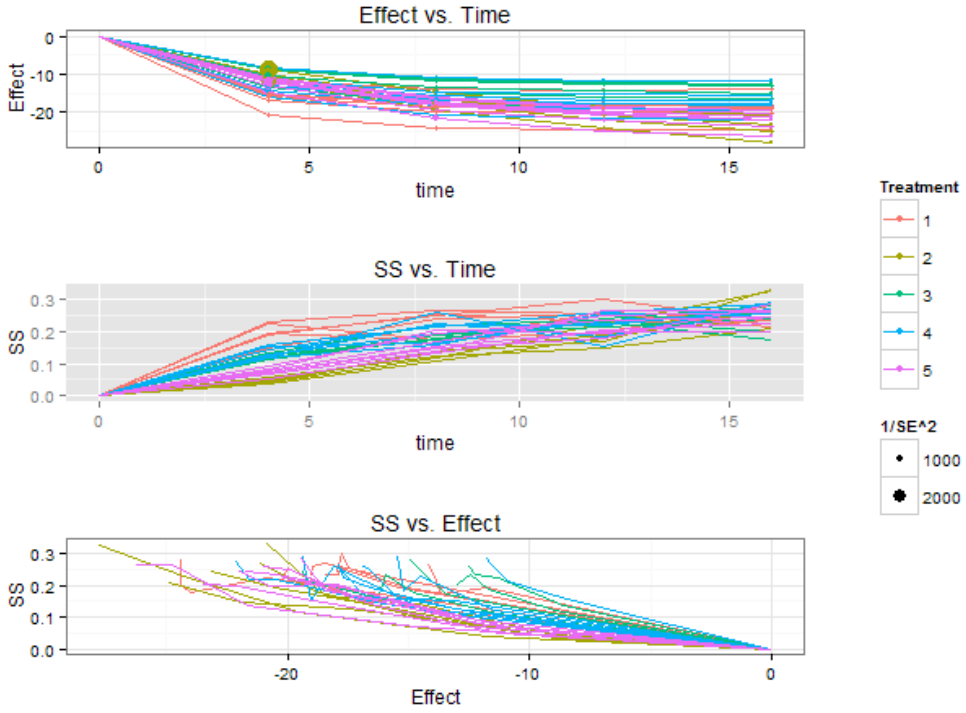

Figure 3.1. Example of some diagnostic plots with generated BEST-ITP data

suggest fitting a random effects model at each time point k:

$$\begin{aligned} \bar{y}_{ijk} &= \mu + \alpha_{ik} + \beta_{jk} + \epsilon_{ijk} \\ \alpha_{ik} &\sim N(0, \sigma^2_{(\alpha)k}) \\ \beta_{jk} &\sim N(0, \sigma^2_{(\beta)k}) \end{aligned} \tag{3.20}$$

This model is related to three of the four above classes of models. When the treatment effects are considered fixed, the multivariate mixed model reduces to (3.20) at each time point. The same is true for the fractional polynomial model when one conditions on a time point. When the treatment effects are considered random, the residuals from fitting (3.20) will be correlated across time and/or within treatments.

The BEST-ITP model reduces to (3.20) when $t = d$. The BEST-ITP model approaches the univariate model as $t \to d$. As a result, one should observe decreasing variability in the residuals as $t \to d$ after fitting (3.20). Additionally, if

the BEST-ITP model is the true underlying model, the residuals from (3.20) will exhibit a decreasing trend over time. This is because the differences among the p_j will cluster the residuals according to treatment, and this clustering will dissipate as ($f(p_j, t_{ijk}, d) \to 1$ (as $t \to d$).

Finally, the Emax model is a nonlinear model which does not reduces to (3.20) only under certain conditions. Thus, this graphical check can not be used to diagnose this model. However, this model is particular to many therapeutic areas, so that the researcher generally knows beforehand to expect a model of this kind.

3.4.2.2 Plots After fitting the random effects model, there are several plots which can be beneficial in choosing an appropriate model. Spaghetti plots of $\hat{\alpha}_{ik}$ vs. k and $\hat{\beta}_{jk}$ vs. k can help one visualize the functional form of the model. One can also check the relationship between the effects by plotting $\hat{\alpha}_{ik}$ vs. $\hat{\beta}_{jk}$. If the data is sparse, one may plot $\hat{\sigma}^2_{(\alpha)k}$ vs k, $\hat{\sigma}^2_{(\beta)k}$ vs. k, and $\hat{\sigma^2}_{(\alpha)k}$ vs. $\hat{\sigma}^2_{(\beta)k}$.

Figure 3.2 shows spaghetti plots for the estimated study and treatment effects. As is clear from the bottom plot, the study and treatment effects seem to have a linear relationship. This is an assumption in the fractional polynomials, the model which generated this data.

In addition to spaghetti plots of the study and treatment effects, residual plots can also be useful. Inspecting the spaghetti plot of residuals vs. time can illustrate a trend in residual variance over time. Recall that for the BEST-ITP model, this variance should decrease over time, as there is extra variability at earlier time points not accounted for by the random effects model. The multivariate mixed model and fractional polynomials model, when true, should exhibit constant variance in the residuals over time after fitting (3.20). In addition to a spaghetti plot of residuals, inspecting a scatterplot matrix of the residuals over the binned time points can illustrate correlation not captured by (3.20).

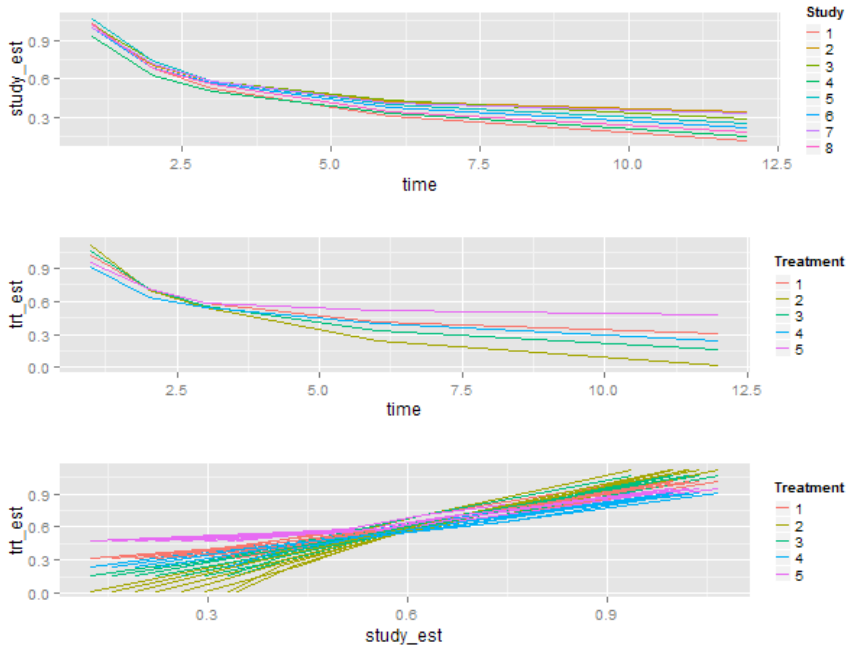

Figure 3.2. Plots of study and treatment effects from data generated using the Fractional Polynomials model

The scatterplot matrix in Figure 3.3 shows decreasing correlation over time. This can of plot can help one choose a correlation structure in a multivariate mixed model, although there is often not enough time points to construct a meaningful variogram. In this case, the data were generated from a multivariate mixed model with an exponential correlation function.

While it is not necessary that every study and treatment be represented at each time point in this graphical check, it is important to have as many of possible of each in order to adequately assess the variability at each time point. If the data contains many time points but is sparse at some time points, one may consider binning adjacent time points where appropriate.

Though we have offered several graphical checks to help choose an appropriate model, this decision will often be dictated by the needs and focuses of the researcher. For example, if a researcher is interested in the underlying shape of the effect, he or she may choose to implement various fractional polynomial models. On the other

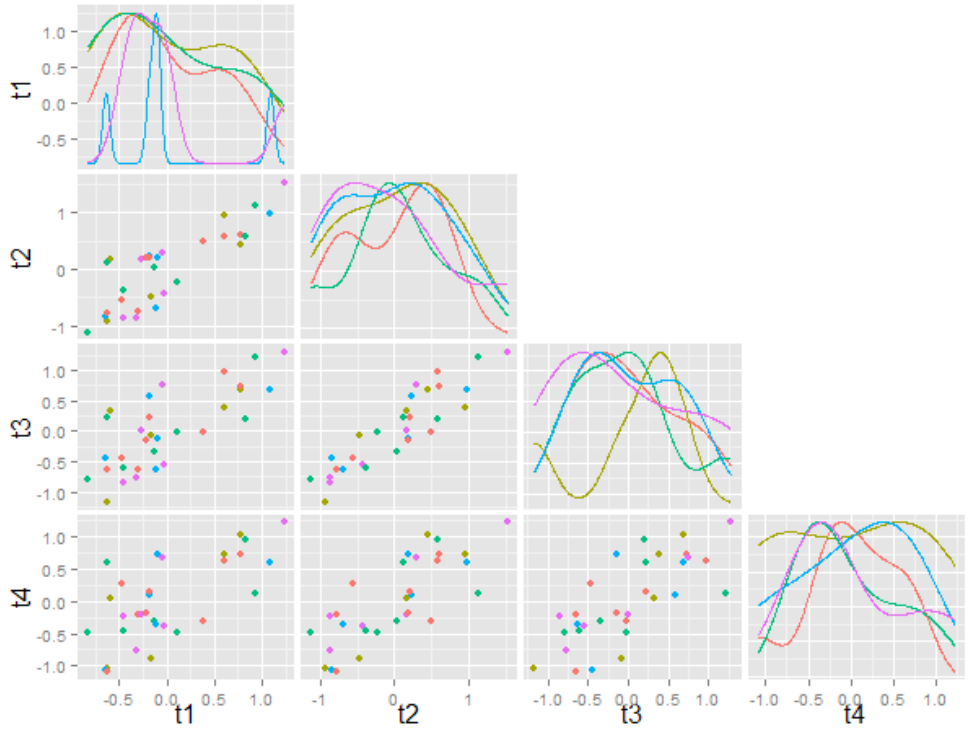

Figure 3.3. Correlation scatterplot matrix for data generated using the Multivariate model

hand, if the researcher is interested in making predictions and willing to assume a plateau effect of the drug, then he or she may choose the BEST-ITP model. The models presented above are fairly flexible, and can be made to handle most common situations.

3.5 Model Fit and Model Comparison

Checking model fit is similar to the univariate case, where deviance residuals, leverages, and the DIC can all be used to diagnose the overall fit of the model to the data. These techniques are discussed in Dias[26]. When the standard error is assumed fixed, the likelihood is a function of the μ_{ijk} only. Hence, the residual deviance can be written as

$$D_{res} = \sum_i \sum_j \sum_k -2\left\{\log(f(\bar{y}_{ijk}|\hat{\mu}_{ijk}, \hat{se}_{ijk})) - \log(f(\bar{y}_{ijk}|\hat{\mu}^S_{ijk}, \hat{se}_{ijk}))\right\}$$

where $\hat{\mu}_{ijk}$ is the estimate of μ_{ijk}, $\hat{\mu}^S_{ijk}$ is the estimate of μ_{ijk} under the saturated model, and $\hat{se}_{ijk}$ is the estimated standard error, assumed to be known. In the models we consider, $f(\cdot)$ is the pdf of a normal distribution.

When one specifies a model for the sample variance as well as the sample mean, as in the BEST-ITP model, the deviance is a function of an underlying precision τ as well as the μ_{ijk}. It is well known that the sample mean and variance are independent statistics when they are calculated from normally distributed data. As a result, we have

$$\begin{aligned} D_{res} = \sum_i \sum_j \sum_k -2 \big\{ & \log(g(\bar{y}_{ijk}|\hat{\mu}_{ijk}, \hat{\tau}) h(S^2_{ijk}|(n_{ijk}-1)/2, 2(n_{ijk}-1)\hat{\tau})) \\ & - \log(g(\bar{y}_{ijk}|\hat{\mu}^S_{ijk}, \hat{\tau}^S) h(S^2_{ijk}|(n_{ijk}-1)/2, 2(n_{ijk}-1)\hat{\tau}^S)) \big\}, \end{aligned}$$

where $g(\cdot|\mu, \tau)$ is a Normal distribution with mean μ and precision τ, $h(\cdot|\alpha, \beta)$ is a Gamma distribution shape parameter α, and scale parameter β.

Conveniently, because the sample mean and sample variance are independent, we can compute deviance residuals for $\bar{y}_{ijk}$ and S^2_{ijk} separately, and sum them to obtain the total residual deviance. This is apparent when one takes the $\log(\cdot)$ of the joint distribution for $\hat{y}_{ijk}$ and S^2_{ijk}:

$$\begin{aligned} D_{res} = \sum_i \sum_j \sum_k -2 \big\{ & \log(g(\bar{y}_{ijk}|\hat{\mu}_{ijk}, \hat{\tau})) \\ & + \log(h(S^2_{ijk}|(n_{ijk}-1)/2, 2(n_{ijk}-1)\hat{\tau})) \\ & - \log(g(\bar{y}_{ijk}|\hat{\mu}^S_{ijk}, \hat{\tau}^S)) \\ & + \log(h(S^2_{ijk}|(n_{ijk}-1)/2, 2(n_{ijk}-1)\hat{\tau}^S)) \big\}. \end{aligned}$$

Rearranging, we have

$$\begin{aligned} D_{res} = & \sum_i \sum_j \sum_k -2 \left\{ \log(g(\bar{y}_{ijk}|\hat{\mu}_{ijk}, \hat{\tau})) - \log(g(\bar{y}_{ijk}|\hat{\mu}^S_{ijk}, \hat{\tau}^S)) \right\} \\ & + \sum_i \sum_j \sum_k -2 \left\{ \log(h(S^2_{ijk}|\alpha_{ijk}, \hat{\beta}_{ijk}) - \log(h(S^2_{ijk}|\alpha_{ijk}, \hat{\beta}^S_{ijk}) \right\}, \end{aligned}$$

where $\alpha_{ijk} = (n_{ijk} - 1)/2$, $\hat{\beta}_{ijk} = 2(n_{ijk} - 1)\hat{\tau}$, and $\hat{\beta}^S_{ijk} = 2(n_{ijk} - 1)\hat{\tau}^S$. Thus, we can independently assess the lack of fit in the mean and in the precision.

Define dev^m_{ijk} as the deviance residual computed for $\bar{y}_{ijk}$ and dev^v_{ijk} as the deviance residual computed for S^2_{ijk}. That is,

$$D_{res} = \sum_i \sum_j \sum_k dev^m_{ijk} + \sum_i \sum_j \sum_k dev^v_{ijk}.$$

Using this notation, we can also write separate leverages for $\bar{y}_{ijk}$ and S^2_{ijk}. We have

$$leverage^m_{ijk} = \bar{dev}^m_{ijk} - \tilde{dev}^m_{ijk}$$

and

$$leverage^v_{ijk} = \bar{dev}^v_{ijk} - \tilde{dev}^v_{ijk}$$

where $\tilde{dev}^m_{ijk}$ is the posterior mean of the deviance due to $\bar{y}_{ijk}$ and $\tilde{dev}^v_{ijk}$ is the posterior mean of the deviance due to S^2_{ijk}

3.6 Practical Concerns

Here, we outline a few implications of the practical choices one makes in implementing these models. First, we discuss the issue of "simultaneous vs. separate" models, also discussed in Dias[27]. To our knowledge, the implications of implementing a meta-analysis using separate models for the baseline treatment and comparison treatments has not been explored. Though Dias[27] suggests how to implement separate models, there is no theoretical justification for why this technique should work, and furthermore no discussion of the assumptions one makes when employing this technique. We feel that this issue is important and often overlooked. In addition, some longitudinal models can *not* be implemented using separate models, presenting a challenge unique to longitudinal models.

The second issue we discuss is that of "adjusting the standard error for correlation", as mentioned in Jansen[44]. At the heart of this practice is a misconception

about how correlation affects the standard error of the mean response. This issue is important because correlation between time points is often not reported, making correct specification of a longitudinal model more difficult. More importantly, handling this correlation correctly affects the estimate of the standard errors, ultimately impacting inference on efficacy of treatments. This is a particularly important issue when one is modeling change from baseline.

3.6.1 "Simultaneous vs. Separate" Models

Consider again the univariate model

$$\begin{aligned} y_{ij} &\sim N(\theta_{ij}, se_{ij}) \\ \theta_{ij} &= \mu_i + d_{(bj)} \\ d_{(b1)} &= 0, \end{aligned}$$

where as before, se_{ij} is the estimated standard error, assumed to be fixed. Certainly this assumption already impacts inference, as if it is not true the resulting inference will mis-state the magnitude of the $d_{(bj)}$, as it will ignore the variability in the se_{ij}. The BEST-ITP model provides one possible framework for modeling the standard errors, based on the assumption of an underlying error variance common to all trials. Of course, one could argue that the estimated se_{ij} are generally representative of the true variability, and that as the goal of inference is estimating the $d_{(bj)}$, assuming them to be fixed provides one with more power by employing a reasonable assumption.

For the same reason, the NICE documents[27] suggests employing what is known as "separate" models for the baseline treatment and the other treatments. This is an approach which is intended for the usual situation of most clinical trials having the same baseline treatment (e.g. placebo). In this case, one models the

placebo treatments separately from the differences $d_{(bj)}$. Specifically, one models

$$\mu_i \sim N(m, \sigma_m^2),$$

where the μ_i are the common baseline treatment. Note that not all baseline treatments are necessarily included in the μ_i modeled above, only those which represent the main reference treatment, such as a placebo. For example, if most trials had a placebo treatment and study i did not, one would not include its reference treatment in the baseline model, as one is simply trying to model the treatment from which all treatment differences are measured.

The goal of modeling the baseline treatment separately is to ensure that "the information in the baseline model does not propagate to the relative treatment effects model."[27] That is, the intent is to keep the variability in estimating the baseline treatment effect in each study μ_i from impacting the estimation of the $d_{(bj)}$. This practice relies on the assumption that the difference between the treatment and placebo effect is not affected by the variability of the placebo effect. That is, one assumes that study variability is actually a location shift of the responses, so that the difference between the responses stays the same. This assumption may be written as

$$\theta_{ij} = \bar{y}_{i1} + d_{(bj)}$$
$$\bar{y}_{i1} \sim N(m, \sigma_m^2)$$

so that one effectively replaces μ_i with $\bar{y}_{i1}$.

The difference in inferences between simultaneous and separate models is in the variability of $\bar{d}_{(bj)}$. To see this, consider a simple example. Suppose that the variance of response is the same across trials for all treatments, that $\sigma^2 = \sigma_{ij}^2$ for all i, j. Let n_1 be the number of studies with the baseline treatment and n_j be the number of studies with treatment j. The maximum likelihood estimate of $d_{(bj)}$ is

given by

$$\bar{d}_{(bj)} = \bar{y}_{\cdot j} - \bar{y}_{\cdot 1},$$

where $\cdot$ denotes summing over the index it replaces. In the simultaneous model, $\bar{y}_{\cdot 1}$ estimates m, and therefore has variability associated with it. Then

$$\begin{aligned} Var(\bar{d}_{(bj)}) &= Var(\bar{y}_{\cdot j}) + Var(\bar{y}_{\cdot 1}) \\ &= \frac{\sigma^2}{n_j} + \frac{\sigma^2}{n_1}. \end{aligned}$$

If one assumes that $n_j \leq n_1$, as is usually the case in meta-analysis, this variance is highest when $n_j = n_1$, treatment j and the placebo have the sum number of arms across trials. Note that as $n_j/n_1 \to 0$, the added variance due to the estimate $\bar{y}_{\cdot 1}$ becomes increasingly negligible.

When one implements separate models, $\bar{y}_{\cdot 1}$ is treated as fixed. Then

$$\begin{aligned} Var(\bar{d}_{(bj)}) &= Var(\bar{y}_{\cdot j}) \\ &= \frac{\sigma^2}{n_j}, \end{aligned}$$

so that this model underestimates the variance of $\bar{d}_{(bj)}$ when the simultaneous model is true. Future work is necessary on the impact of mis-specifying between simultaneous and separate models, and the practical benefits of each.

Choosing between simultaneous and separate models is a matter of taste for a univariate analysis, and each are straightforward to implement. This is not so for longitudinal models, and some may not yield a straightforward analysis under separate models, and some are much more difficult to implement under separate models. We give an example of each below.

As an example of the difficulty is simply constructing separate models in the longitudinal setting, consider an Emax model where the $ED50$ is allowed to vary by study and treatment. In this case, the difference between treatment j and the

baseline b in study i at time point k is

$$Y_{ijk} - Y_{ibk} = \left(\frac{Emax \times t_{ijk}}{t_{ijk} + ED50_{ij}} + \epsilon_{ijk} \right) - \left(\frac{Emax \times t_{ijk}}{t_{ijk} + ED50_{ib}} + \epsilon_{ijk} \right).$$

It's clear that when one combines these models, one will not end up with an Emax model, and one that is substantially more complicated than the simultaneous model.

Practical implementation is an issue for the multivariate model. This model specifies random vectors of length K which are correlated across time and with each other. One can use conditional identities to implement a simultaneous model of dimension K in WinBUGS or JAGS. When one wishes to specify separate models, the differences from baseline are correlated, and must be modeled as a single vector. Thus, if a study has two treatments in addition to baseline, one must specify a response of dimension $2K$. In addition to this greater computational cost, WinBUGS and JAGS do not allow multivariate models of varying dimensions to be specified, making this model difficult to implement.

3.6.2 "Adjusting the Standard Error for Correlation"

Repeated measures are common in clinical trials. One generally has at least two time points, the time of baseline and the time of the endpoint of interest. One often observes positive correlation between these measurements, as healthy patients tend to respond more positively to treatments, and sicker patients tend to respond less frequently to treatments. In this case the correlation, ρ, is the *within-patient* correlation. That is, a patients response at multiple endpoints are correlated with each other. This correlation is generally assumed to be positive.

This correlation is often not recorded, and causes problems when one wants to compute the variance of a change from baseline estimate. This issue is discussed in more detail in the NICE documents[26]. Here, i is the study index, and k is the treatment index.

> However, in practice many studies fail to report an adequate measure of the uncertainty for the before-after difference in outcome and instead report the mean and variance, $y_{ik}^{(b)}$ and $V_{ik}^{(b)}$, (or other measure of uncertainty) at baseline (before), and at follow-up times (after), $y_{ik}^{(a)}$ and $V_{ik}^{(a)}$, separately. While the mean change from baseline can be easily calculated as
>
> $$y_{ik}^{\Delta} = y_{ik}^{(b)} - y_{ik}^{(a)}$$
>
> To calculate V_{ik}^{Δ} for such trials, information on the within-patient correlation ρ is required since
>
> $$V_{ik}^{\Delta} = V_{ik}^{(b)} + V_{ik}^{(a)} - 2\rho\sqrt{V_{ik}^{(b)}V_{ik}^{(a)}}$$
>
> Information on the correlation ρ is seldom available. It may be possible to obtain information from a review of similar trials using the same outcome measures, or else a reasonable value for ρ, often 0.5 (which is considered conservative) or 0.7, can be used alongside sensitivity analyses.

We can see that simply adding the variances, would lead to an overestimate of the variance of the mean change from baseline if one ignores the positive within-patient correlation. However, this is a common way to construct an estimate for the variance in meta-analysis, so that one often over-estimates the standard errors.

In his paper introducing fractional polynomial methods for longitudinal meta-analysis, Jansen[44] notes that there is within-trial correlation in the change from baseline (CFB):

> The CFB in pain at each time point are correlated over time. Unfortunately, this within-trial correlation was not reported for the included studies. As such, we performed sensitivity analyses assuming different values for the correlation: 0, 0.5, 0.9....Estimates for σ_{ijk}^2 (variance of CFB) were obtained from the reported standard errors as presented in Table I and adjusted by dividing these variance estimates by $1 - \rho^2$.

This "within-trial" correlation is the same as the aforementioned "within-patient" correlation. Though Jansen is correct in saying that one generally underestimates the variance of correlated observations, this occurs only when one calculates

an overall variance by pooling observations across time, as in regression. However, the estimated variances in meta-analysis are conditional at a time point t_{ijk}. Thus, it is not the case that positive correlation will result in underestimating variability in this setting, as this is **not** how these variances are calculated. As noted above, the variances at a time point t are generally estimated using $V_{ik}^{\Delta} = V_{ik}^{(b)} + V_{ik}^{(a)}$, which overestimates the actual variance if the correlation is positive.

In fact, the CFB measurements *are* correlated, but the impact on estimation of variance is slightly more subtle. To see exactly what the correlation for the CFB might be, first consider a mean response that is distributed as

$$\bar{\boldsymbol{Y}} \sim N\left(\boldsymbol{\mu}, \Omega = \frac{1}{N}\Sigma\right),$$

so that the mean effect is still correlated across time. The mean change from baseline is a linear transformation of the mean effect. For example, consider the case where $n = 3$. Let

$$\boldsymbol{A} = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then the mean change from baseline $C\bar{F}B$ is

$$C\bar{F}B = \boldsymbol{A}'\bar{\boldsymbol{Y}},$$

so that

$$C\bar{F}B \sim N(\boldsymbol{A}'\boldsymbol{\mu}, \boldsymbol{A}'\Omega\boldsymbol{A}).$$

If

$$\Omega = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix},$$

one can show that

$$\boldsymbol{A}'\Omega\boldsymbol{A} = \begin{bmatrix} 2-2\rho & 1-\rho \\ 1-\rho & 2-2\rho \end{bmatrix}.$$
$$= (2-2\rho)\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

Thus, regardless of the strength of correlation, if the correlation is the same among all pairs of time points, the correlation among the time points of the CFB measurements is 0.5. This case is identical to that of observing > 2 arms in a clinical trial, and thus knowing that the treatment effects relative to baseline are correlated. In this case, one assumes that the treatments effects are equally correlated with each other, resulting in a correlation of 0.5.

Notice that while this correlation does *not* impact the estimates of the marginal distributions of $C\bar{F}B_j$, it *does* impact the conditional distributions $C\bar{F}B_j|C\bar{F}B_{j'\neq j}$. One can write this distribution as

$$C\bar{F}B_j|C\bar{F}B_{j'\neq j} \sim N\left((\theta_j-\theta_i) + \frac{1}{n-1}\sum_{j=1}^{n-1}\left[C\bar{F}B_j - (\theta_j-\theta_1)\right], \frac{n}{2(n-1)}\sigma^2\right),$$

a relation which is commonly used for specifying random effects for multi-arm trials[26]. When $n > 2$, the variance of the conditional distribution is at most $3\sigma^2/4$, suggesting again that using the marginal variances would overestimate the variance of the conditional distribution.

One might have a different structure for Ω. For instance, suppose Ω has an AR(1) structure, so that

$$\Omega = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}.$$

Then

$$\boldsymbol{A}'\Omega\boldsymbol{A} = \begin{bmatrix} 2-2\rho & 1-\rho^2 \\ 1-\rho^2 & 2-2\rho^2 \end{bmatrix}.$$

Interestingly, one can see that a different adjustment must be made for the variance at time point three than time point two.

Thus, the general estimates for $var(CFB)_j$ typically *overestimate* the variance at any given time point.

3.7 Simulation

The title of a talk given by Richard Riley concerning multivariate meta-analysis is quite apt in the discussion of the pursuance of longitudinal meta-analysis methods: "Multivariate meta-analysis – is it worth the extra effort?"[67] Indeed, though multivariate methods potentially offer more insight into data, they present several challenges to the practitioner, including added assumptions, estimation difficulties, and missing correlations[43]. These difficulties all apply to longitudinal meta-analysis, with the added difficulty of model specification. This is the result of specifying a model for the trend of the mean over time. Even when the practitioner adequately handles all of these difficulties, there is no guarantee that the resulting inference will better than had one simply used univariate methods. Jackson[43] notes that the statistical properties of the individual parameter estimates are often only marginally improved.

There is reason to believe that longitudinal methods promise more than just marginal improvement in many applications. For instance, longitudinal methods are likely to provide more robust inference when an actual trend exists, reducing bias. This is because a univariate analysis only considers a single observation from each study, and thus will treat outliers equally to all other observations. Because a meta-analysis typically involves 20 or less studies, this can create considerable bias.

This is especially true when one uses a naive interpolation method or simply leaves out studies without data at the time point of interest.

As with multivariate meta-analysis, longitudinal meta-analysis offers inferences not available at all in the univariate setting. It allows the practitioner to make inferences about how the differences in treatments change over time, which may ultimately prove useful in planning a new trial. For instance, in the case where treatment efficacy is known to plateau, one could plan a trial so as to stop soon after the plateau is reached, so as not to take unnecessary measurements. In addition, one may use this information in conjunction with data about the drug of interest to select a sample size. When dosing information is available, this could lead to more powerful adaptive designs, as one is more informed about how the efficacy according to dose is likely to change over time.

We consider this question of utility in a simulation. Jansen[44] presents longitudinal data from 17 studies of treatments for osteoarthritis. There are six treatments represented in the studies, with observations at varying times. The number of follow-up times varies from 2 to 13, with the soonest follow-up time being 1 week and the longest being 52 weeks. The response variable in mean change from baseline in visual analogue scale (VAS) score in pain. This data is plotted in Figure 3.4, where each color denotes a specific treatment. From the plot, one can see that the treatment effect seems to plateau around 8 weeks, suggesting the BEST-ITP and Emax models may be appropriate. In addition, it appears that the most effective treatment (that which lowers the mean VAS score the most), is likely 3HYGF20. Though the VAS score is measured on a scale from 0 to 100, the mean change in VAS score can be considered to be normally distributed as a result of the central limit theorem.

Jansen[44] models these data with the fractional polynomials model with $p_1 = 0.5$ and $p_2 = 1$. The best model chosen via the DIC has one random treatment

effect, which models β_{2ij}. This model is shown in Figure 3.10. The model does not have an intercept, as measures of change from baseline must be 0 at $t = 0$. In total, the model has two fixed study effects terms, μ_{1i} and μ_{2i}, one fixed treatment effect term, $d_{2(bj)}$, and one random treatment effect term, $\delta_{1i(bj)}$. As the response variable $\bar{y}_{ijk}$ must lie in $[-100, 100]$, a non-informative prior for standard deviation terms is $U(0, 100)$.

We compare all of our models to the univariate model, shown in Figure 3.6. This model is implementing by utilizing only the observations taken at the time points of interest. Thus, we do not attempt to impute, and leave the effects of various imputation methods for future research. The BEST-ITP model used in simulation is given in Figure 3.7. This model does not have a model for the variance, as the standard errors are all assumed fixed in this simulation. The Emax model used in simulation is given in Figure 3.8, and is of the form discussed earlier which reduces to a univariate meta-analysis at each time point. Finally, our multivariate model is given in Figure 3.9. As we only select two time points for comparison, this model is bivariate, utilizing the observations from each time point. Note that we use a variation of the model similar to Model 3 given in Achana[2]. This is to enable inference for all treatments at both time points, as this is not possible with the more general model presented above.

To test the utility of longitudinal models on this data set, we simulate data similar to the Jansen data, and make inferences using all of the other models. The simulated data is generated using the fitted Fractional Polynomials model proposed by Jansen. After all models are fit to the simulated data, we calculate $E(\widehat{CFB_{3HYGF20}}) - E(\widehat{CFB_{3HYGF20}})$ at times $t = 3$ and $t = 8$. These time points were chosen because they may represent clinical points of interest. Additionally, a relatively high number of studies reported outcomes at these times (8 and 7 studies, with a total of 19 and 15 responses), so as to give the univariate model the

best chance at yielding accurate inference. Other time points with a higher number of studies reporting were not as clinically meaningful, and thus not considered: 1 week, 2 weeks, and 12 weeks. The procedure for our simulation is outlined below.

Due to the complexity of the models, it is difficult to construct models with the

Algorithm 3 Longitudinal Meta-Analysis Simulation

procedure SIMULATION
  Fit $\{y_{ijk}\}$ using Fractional Polynomials model
  Store estimates $\hat{\mu}_{1i}$, $\hat{\mu}_{2i}$, $\hat{d}_{1(bj)}$, $\hat{d}_{2(bj)}$, and $\hat{\sigma}_{\delta}$
  for $r = 1 : 1000$ **do**
    Generate for all i, j, $\tilde{\delta}^r_{i(bj)} \sim N(\hat{d}_{2(bj)}, \hat{\sigma}_{\delta})$
    Generate for all i, j, k, $\tilde{\epsilon}^r_{ijk} \sim N(0, se_{ijk})$
    Compute $\tilde{\beta}^r_{2ijk} = \hat{\mu}_{2i} + \tilde{\delta}^r_{i(bj)}$
    Compute $\tilde{\theta}^r_{ijk} = \hat{\beta}_{1ijk} + \tilde{\beta}^r_{2ijk}$
    Compute $\tilde{y}^r_{ijk} = \tilde{\theta}^r_{ijk} + \tilde{\epsilon}^r_{ijk}$
    for M in models **do**
      Compute $E(\widehat{CFB_{3HYGF20}}) - E(\widehat{CFB_{3HYGF20}})$ at $t = 3, 8$
    end for
  end for
end procedure

same number of parameters to enable direct comparison. We attempt some level of consistency in the following way. The proposed Fractional Polynomials model has one random treatment effect term $\delta_{i(bj)}$, and one fixed treatment effect term $d_{2(bj)}$. We include one random and one fixed treatment effect in the BEST-ITP and Emax models. For these models, the fixed treatment effect terms are p_j and α_j, respectively. At each time point ($t = 3$ and $t = 8$), we perform a univariate analysis with a single random treatment effect. For direct comparison, we implement the multivariate model with a random treatment effect at each time point. Though not exact, these restrictions allow some level of consistent complexity among the models. Note that this reflects the common scenario faced by the statistician of having to

choose among model types of similar complexity which are *not* necessarily subsets of a more general model.

Figure 3.5 plots selected quantiles of $E(\widehat{CFB_{3HYGF20}}) - E(\widehat{CFB_{3HYGF20}})$ for all of the models at the two time points. The results have a few interesting features. First, the univariate model is biased, and has worse coverage than the longitudinal models. In fact, the coverage for the univariate model is 84% at $t = 3$, and 63% at $t = 8$. This is a result of the small sample size, leading to far fewer treatment comparisons than the longitudinal models can utilize. As mentioned earlier, if a time point has a few outlying observations, these have far greater influence in the univariate model, due to the smaller sample size. In addition, the consistency assumption propagates the influence of an outlier, compounding the problem. To put this problem in context, the univariate utilizes 19 and 15 responses at the two time points to perform inference, whereas the longitudinal models utilize 150 observations. The only exception is the multivariate model, which uses all of the observations a univariate model would use at either time point, 34.

Another interesting feature in the results is the tight credible sets yielded by the BEST-ITP and Emax models. These should not be taken to suggest that these models are superior, as we know that neither one is the true model (in this case, fractional polynomials). We can see from the simulated medians of $E(\widehat{CFB_{3HYGF20}}) - E(\widehat{CFB_{3HYGF20}})$ that these two models are somewhat biased, as the medians do not intersect with the true value (except for the Emax model at $t = 8$). This is a result of employing the wrong model. Furthermore, though it is encouraging that they both perform well, their coverage is perhaps *too* good, as they both exhibit 100% coverage for their 95% credible sets. Thus, one has coverage above the nominal rate with smaller credible sets. However, one should not expect this in general. These credible sets are narrow because these models are less flexible than the fractional polynomials and multivariate model, and thus make similar

inferences at every simulation iteration. Recall that these models both assume a plateau, so that if this is not true, these models will perform incredibly poorly. In the Jansen data, the data seems to plateau, although there is some evidence of an increasing trend over time after 10 weeks. Thus, the Emax and BEST-ITP models fit "just well enough", and produced reliable credible sets. This would not be true if the true model were much different than these two models, or if inference were perhaps made a different time point, where the trend of the true model is much different than the trend estimated by these two models.

Lastly, the fractional polynomials and multivariate models produced credible sets which all had coverage at about 95%. Though these credible sets are wider than the BEST-ITP and Emax credible sets, as discussed above, this is an artifact of the relative inflexibility of these two models and that they are not too different from the underlying model. In addition, the fractional polynomials and multivariate models are unbiased.

This simulation shows that using a longitudinal model provides more robust inference than simply using data at a single time point. Furthermore, even if one chooses the wrong model, one can still obtain better inference than simply using the univariate model. Further simulation is needed to explore the conditions for exactly when this is true. As discussed above, though model mis-specification did not produce egregiously incorrect inference here, it certainly could if the model is much different than the true model. Model specification is difficult, particularly in this setting, as many of the models are new or not often utilized. Hopefully the model diagnostic techniques suggested above can lead one to specifying more correct models, and avoiding incorrect inference.

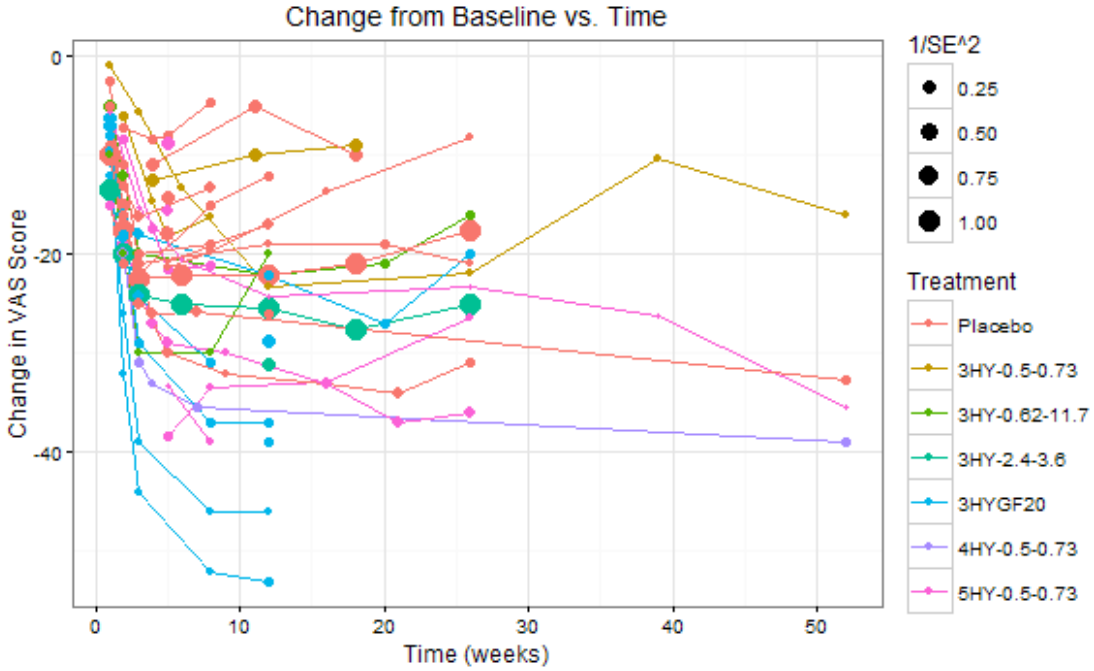

Figure 3.4. Spaghetti plot of Jansen data

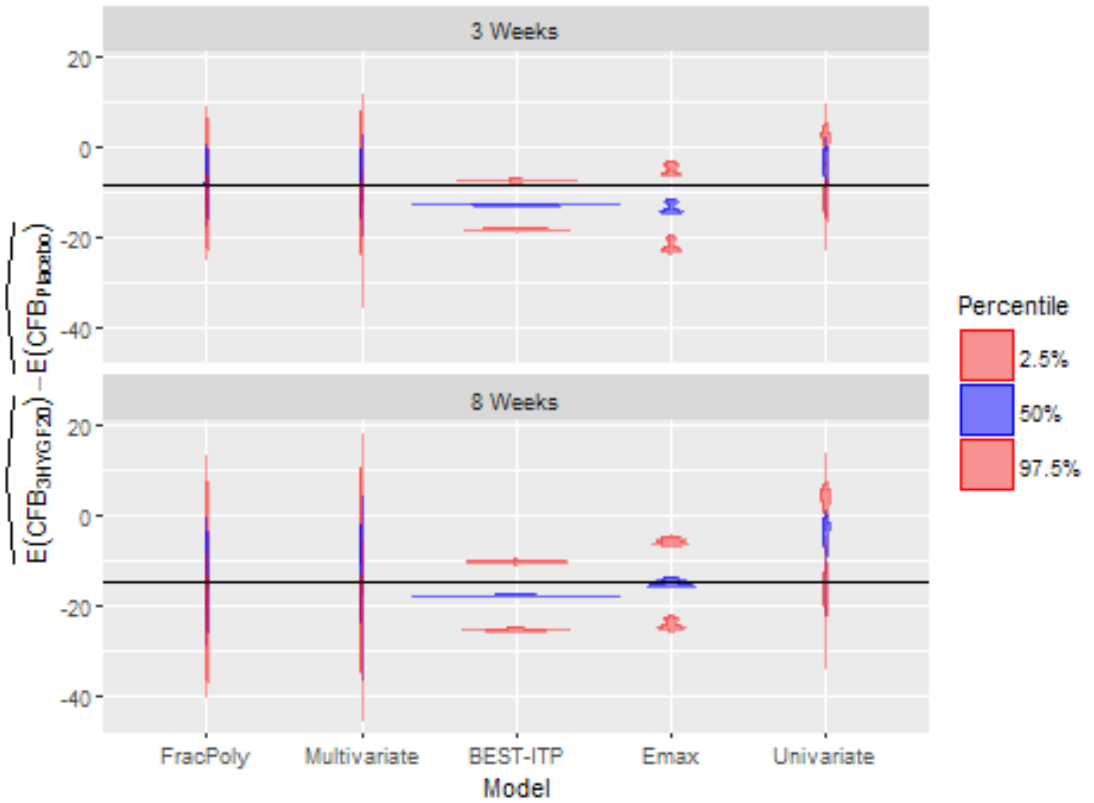

Figure 3.5. Violin plots of selected percentiles collected from 1,000 simulations

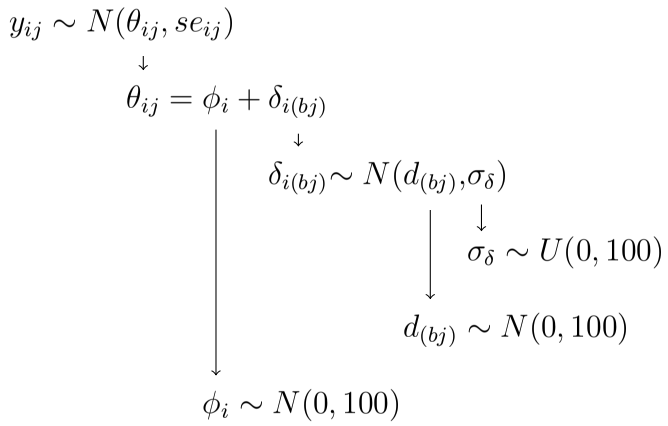

Figure 3.6. Univariate meta-analysis model used in simulation

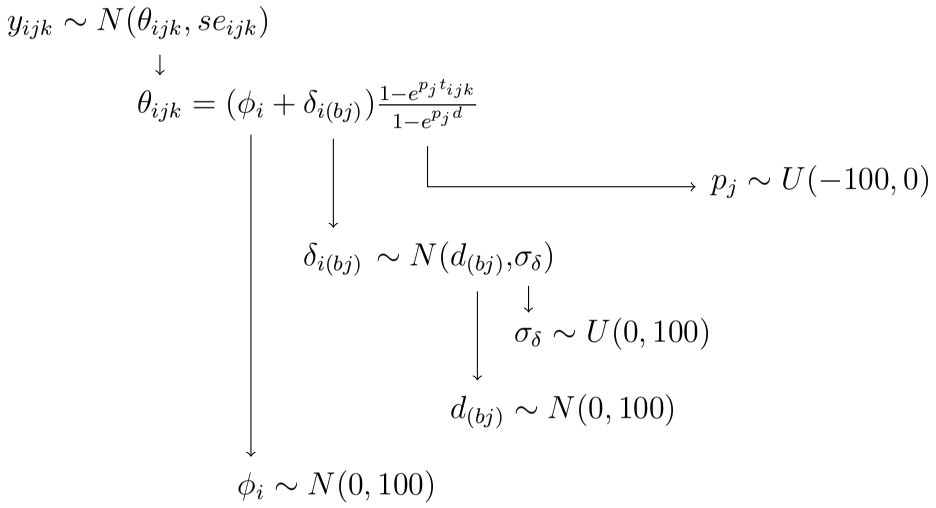

Figure 3.7. BEST-ITP meta-analysis model used in simulation

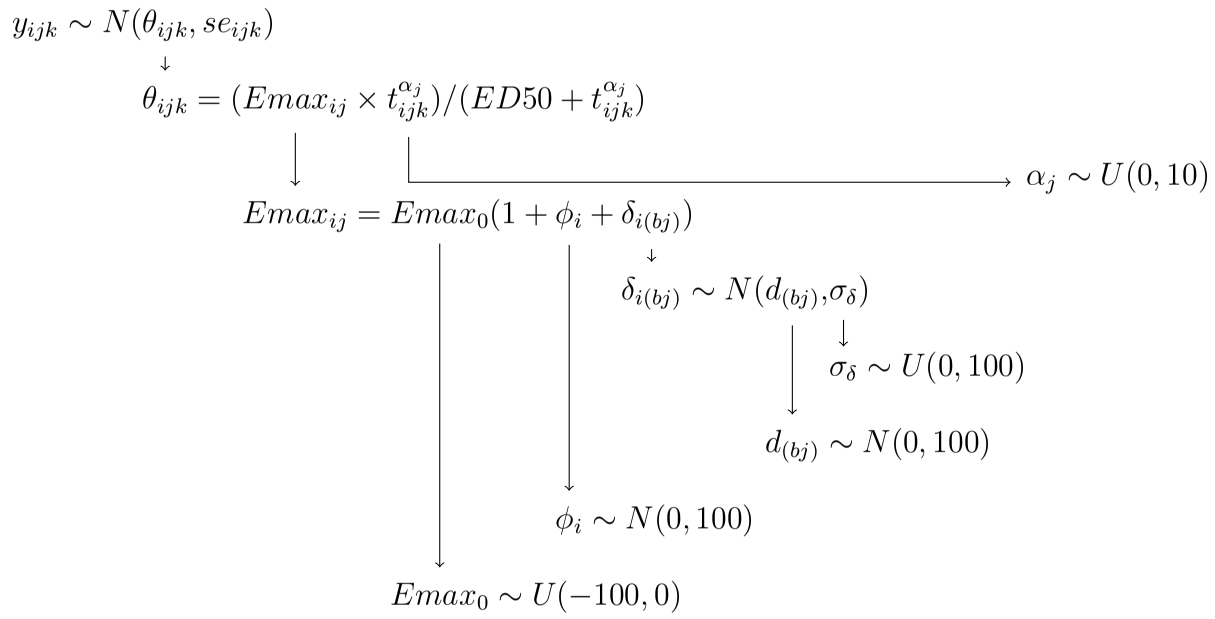

Figure 3.8. Emax meta-analysis model used in simulation

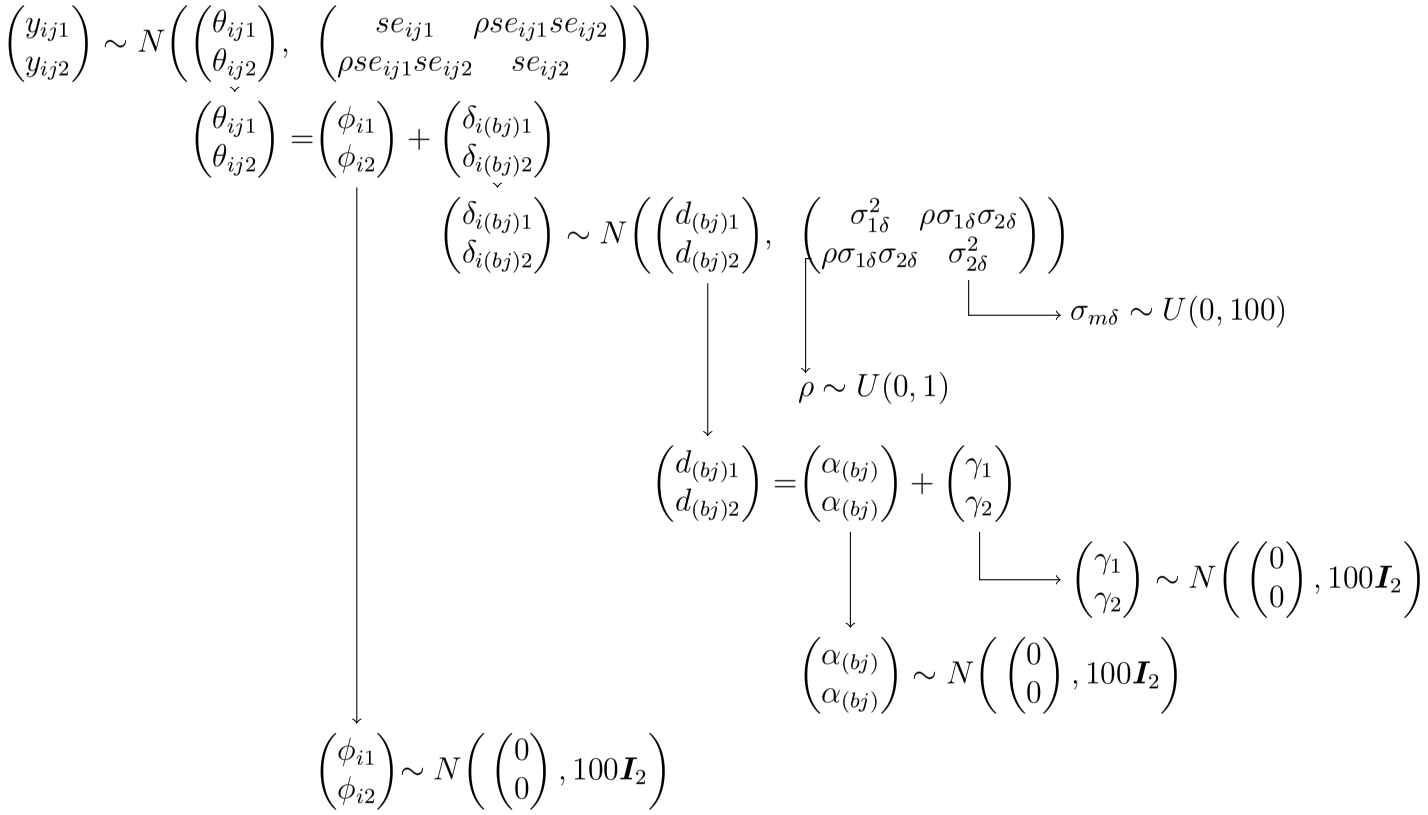

Figure 3.9. Multivariate meta-analysis model used in simulation

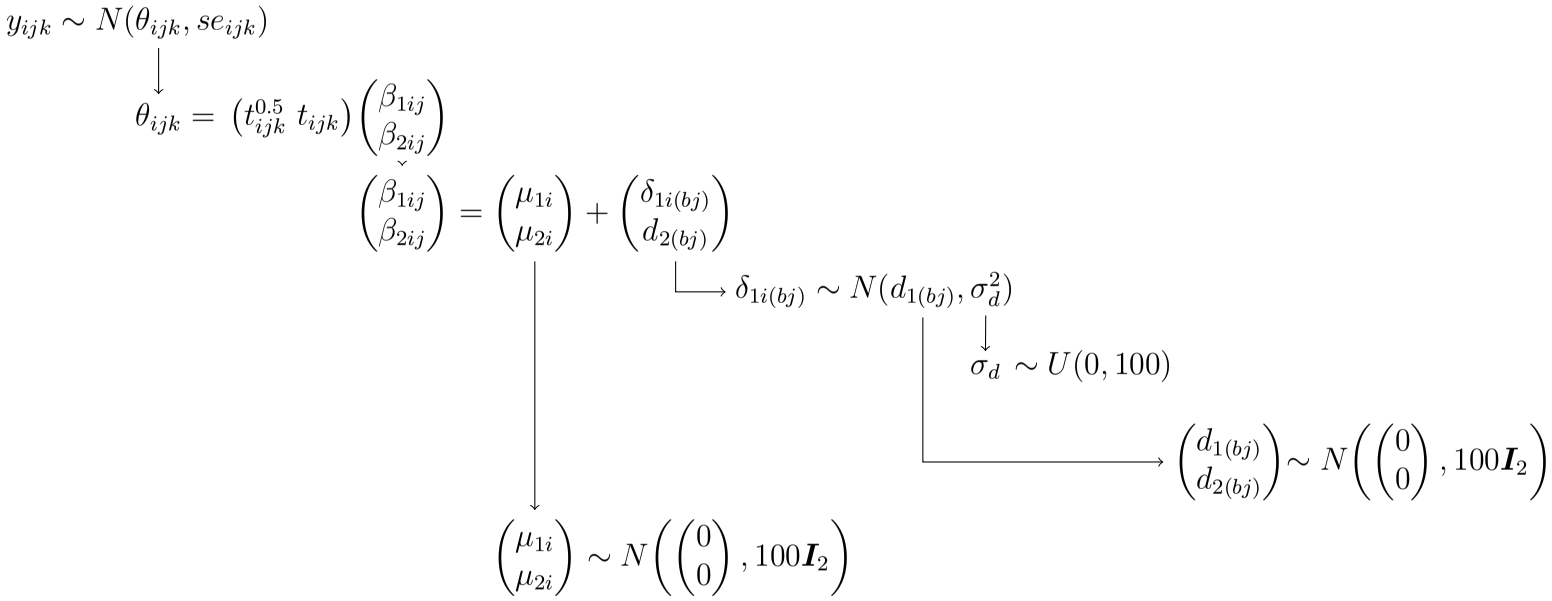

Figure 3.10. Fractional Polynomials meta-analysis model used in simulation

3.8 Appendix

3.8.1 Squared Relationship of Effect and SS in the BEST-ITP Model

The error inside model specifies a unique trend on both the mean and the variance over time. The exponential term which controls the percentage of total effect seen in the mean is squared in the model for the standard error term, suggesting a quadratic relationship. To see this, recall that in the error inside model, we have

$$\bar{Y}_{ijk} = \mu_{ijk}\left(\frac{1-e^{p_j t_{ijk}}}{1-e^{p_j d_i}}\right) \quad \text{and} \tag{3.21}$$

$$\sigma^2_{ijk} = \frac{\sigma^2}{n_{ijk}}\left(\frac{1-e^{p_j t_{ijk}}}{1-e^{p_j d_i}}\right)^2. \tag{3.22}$$

We can rewrite equation (3.22) as

$$S^2_{ijk} = \sigma^2\left(\frac{1-e^{p_j t_{ijk}}}{1-e^{p_j d_i}}\right)^2. \tag{3.23}$$

Rearranging (3.21) and substituting for the exponential term in (3.24), we have

$$\begin{aligned} S^2_{ijk} &= \sigma^2\left(\frac{\bar{Y}_{ijk}}{\mu_{ijk}}\right)^2 \\ &= \xi_{ijk}\bar{Y}^2_{ijk}, \end{aligned}$$

where $\xi_{ijk} = \sigma^2/\mu^2_{ijk}$. Note that μ_{ijk} only depends on k through n_{ijk}, so that if one has an equal number of patients across time, the above relation reduces to

$$S^2_{ijk} = \xi_{ij}\bar{Y}^2_{ijk}. \tag{3.24}$$

This implies that for each treatment group within a study, one should observe a quadratic relationship between the sums of squares and the mean. Furthermore, equation (3.24) does not have an intercept or linear term, making visual and analytic inspection of the error inside assumption relatively easy. Note that this relationship is true if the number of patients is the same across time for each treatment in each study, and that visual and analytic checks will be approximate if the sample size changes only slightly over time.

CHAPTER FOUR

Conclusions

In this dissertation we considered a variety of Bayesian longitudinal models in novel applications. The first application was to the problem of trending in observed probabilities of collision of two satellites orbiting Earth. This application presented data with a number of modeling challenges, including a bounded response variable, irregularly spaced observations, and few reliable covariates. We presented a number of models for handling such data, including some innovative Bayesian Beta mixed models. Ultimately, we found that the simpler Look-Up method worked better than most other methods, although the New Beta regression method had similar properties. Ultimately, this points to two truths. First, that the variability in the any data set dictates how predictive a model can be, no matter how clever the model is. Second, that often the simpler model is the more useful and enlightening model. In this case, we find that simply knowing the percentile of the previous value and how the percentiles change over time is enough to parse most of the variability in the data.

This problem elucidates a few new directions for future research. Statistically, one interesting problem is that of using previous values in longitudinal studies when the observations are irregularly spaced. This can be handled when the observations are regularly spaced using state space models, but extension to irregular spacing is not straightforward. The usual approach is implementation of a random effect. However, in applications such as ours, the last observed value carries the most weight about future values, and trajectories are often erratic. As for trending in probabilities of collision, we mentioned earlier that future research should focus on non-parametric methods. In addition, we believe that future research should focus on the 10-20% of

events which are not well predicted by the models presented. It may be that there is a way to know when one has "unusual data", as suggested by the Beta clustering model.

The second application of Bayesian longitudinal models presented was network meta-analysis. Here, we took steps to collect the existing research and to offer up some new models which may prove to be useful in this field. We feel that simply providing the general framework and simple diagnostic tools represents a major step forward in this area, as these are rare in the literature. Our development opens the door for many new avenues of research: sample size determination, effects of model mis-specification, measures of longitudinal inconsistency, etc.

BIBLIOGRAPHY

[1] Celestrak: Iridium 33/cosmos 2251 collision. http://celestrak.com/events/collision. Accessed: 2016-06-28.

[2] Felix A Achana, Nicola J Cooper, Sylwia Bujkiewicz, Stephanie J Hubbard, Denise Kendrick, David R Jones, and Alex J Sutton. Network meta-analysis of multiple outcome measures accounting for borrowing of information across outcomes. *BMC medical research methodology*, 14(1):92, 2014.

[3] Jae Eun Ahn and Jonathan L French. Longitudinal aggregate data model-based meta-analysis with nonmem: approaches to handling within treatment arm correlation. *Journal of pharmacokinetics and pharmacodynamics*, 37(2):179–201, 2010.

[4] Maruthi R Akella and Kyle T Alfriend. Probability of collision between space objects. *Journal of Guidance, Control, and Dynamics*, 23(5):769–772, 2000.

[5] Salvatore Alfano. Relating position uncertainty to maximum conjunction probability©. 2005.

[6] J Antoch and Paul Janssen. Nonparametric regression m-quantiles. *Statistics & probability letters*, 8(4):355–362, 1989.

[7] William N. Barker. Astrodynamics concepts and terminology, omitron, inc.

[8] Pallab K Bhattacharya and Ashis K Gangopadhyay. Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, pages 1400–1415, 1990.

[9] Wagner Hugo Bonat, Paulo Justiniano Ribeiro Jr, and Walmes Marques Zeviani. Likelihood analysis for a class of beta mixed models. *Journal of Applied Statistics*, 42(2):252–266, 2015.

[10] Howard D Bondell, Brian J Reich, and Huixia Wang. Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010.

[11] Simone Borra and Agostino Di Ciaccio. Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods. *Computational statistics & data analysis*, 54(12):2976–2989, 2010.

[12] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

[13] J Russell Carpenter, F Landis Markley, and Dara Gold. Sequential probability ratio test for collision avoidance maneuver decisions. *The Journal of the Astronautical Sciences*, 59(1-2):267–280, 2012.

[14] J Russell Carpenter, F Landis Markley, and Dara Gold. Wald sequential probability ratio test for analysis of orbital conjunction data. In *AIAA Guidance, Navigation, and Control (GNC) Conference*, 2013.

[15] J Russell Carpenter, FL Markley, KT Alfriend, C Wright, and J Arcido. Sequential probability ratio test for collision avoidance maneuver decisions based on a bank of norm-inequality-constrained epoch-state filters. In *AAS/AIAA Astrodynamics Specialist Conference, Girdwood, AK*, 2011.

[16] Gilles Celeux, Florence Forbes, Christian P Robert, D Michael Titterington, et al. Deviance information criteria for missing data models. *Bayesian analysis*, 1(4):651–673, 2006.

[17] Edilberto Cepeda and Dani Gamerman. Bayesian methodology for modeling parameters in the two parameter exponential family. *Revista Estadística*, 57(168-169):93–105, 2005.

[18] E Cepeda-Cuervo. Modeling variability in generalized linear models. *Mathematics Institute, Universidade Federal do Rio de Janeiro*, 2001.

[19] F Kenneth Chan. *Spacecraft collision probability*. Aerospace Press El Segundo, CA, 2008.

[20] Ken Chan. Improved analytical expressions for computing spacecraft collision probabilities. *Advances in the Astronautical Sciences*, 114:1197–1216, 2003.

[21] Probal Chaudhuri et al. Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of statistics*, 19(2):760–777, 1991.

[22] Harris Cooper, Larry V Hedges, and Jeffrey C Valentine. *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation, 2009.

[23] Helen A Dakin, Nicky J Welton, AE Ades, Sarah Collins, Michelle Orme, and Steven Kelly. Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension. *Statistics in medicine*, 30(20):2511–2535, 2011.

[24] Herbert A David. Tables related to the normal distribution: A short history. *The American Statistician*, 59(4):309–311, 2005.

[25] Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265. ACM, 1986.

[26] Sofia Dias, Alex J Sutton, AE Ades, and Nicky J Welton. Evidence synthesis for decision making 2 a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5):607–617, 2013.

[27] Sofia Dias, Nicky J Welton, Alex J Sutton, and AE Ades. Evidence synthesis for decision making 5 the baseline natural history model. *Medical Decision Making*, 33(5):657–670, 2013.

[28] Peter Diggle. *Analysis of longitudinal data*. Oxford University Press, 2002.

[29] Ying Ding and Haoda Fu. Bayesian indirect and mixed treatment comparisons across longitudinal time points. *Statistics in medicine*, 32(15):2613–2628, 2013.

[30] Matthias Egger, George Davey-Smith, and Douglas Altman. *Systematic reviews in health care: meta-analysis in context*. John Wiley & Sons, 2008.

[31] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

[32] Jorge I Figueroa-Zúñiga, Reinaldo B Arellano-Valle, and Silvia LP Ferrari. Mixed beta regression: a bayesian perspective. *Computational Statistics & Data Analysis*, 61:137–147, 2013.

[33] Ryan C Frigm, Joshua A Levi, and Dimitrios C Mantziaras. Assessment, planning, and execution considerations for conjunction risk assessment and mitigation operations. In *Proceedings of SpaceOps 2010 Conference: Delivering on the Dream, Huntsville, Alabama*, pages 25–30, 2010.

[34] Haoda Fu and David Manner. Bayesian adaptive dose-finding studies with delayed responses. *Journal of biopharmaceutical statistics*, 20(5):1055–1070, 2010.

[35] Alan E Gelfand, Adrian FM Smith, and Tai-Ming Lee. Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, 87(418):523–532, 1992.

[36] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[37] Jorge Luiz Gross, James Rogers, Daniel Polhamus, William Gillespie, Christian Friedrich, Yan Gong, Brigitta Ursula Monz, Sanjay Patel, Alexander Staab, and Silke Retlich. A novel model-based meta-analysis to indirectly estimate the comparative efficacy of two medications: an example using dpp-4 inhibitors, sitagliptin and linagliptin, in treatment of type 2 diabetes mellitus. *BMJ open*, 3(3):e001844, 2013.

[38] Xuming He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997.

[39] MD Hejduk, D Plakalovic, ME Hametz, LK Newman, JC Ollivierre, BA Beaver, and RC Thompson. Launch cola operations: Examination of data products, procedures, and thresholds. In *Flight Dynamics (FD) Task Order 21 Technical Memorandum*, 2014.

[40] Wallace Hendricks and Roger Koenker. Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American statistical Association*, 87(417):58–68, 1992.

[41] Julian PT Higgins, Sally Green, et al. *Cochrane handbook for systematic reviews of interventions*, volume 5. Wiley Online Library, 2008.

[42] K Jack Ishak, Robert W Platt, Lawrence Joseph, James A Hanley, and J Jaime Caro. Meta-analysis of longitudinal studies. *Clinical Trials*, 4(5):525–539, 2007.

[43] Dan Jackson, Richard Riley, and Ian R White. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.

[44] JP Jansen, MC Vieira, and S Cope. Network meta-analysis of longitudinal data using fractional polynomials. *Statistics in medicine*, 2015.

[45] Ajay Jasra, CC Holmes, and DA Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

[46] Ashley P Jones, Richard D Riley, Paula R Williamson, and Anne Whitehead. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clinical Trials*, 6(1):16–27, 2009.

[47] A Kelly and W Watson. Collision avoidance: Coordination of predicted conjunctions between nasa satellites and satellites of other countries. In *Advanced Maui Optical and Space Surveillance Technologies Conference*, volume 1, page 78, 2014.

[48] Michael G Kenward and Geert Molenberghs. Last observation carried forward: a crystal ball? *Journal of biopharmaceutical statistics*, 19(5):872–888, 2009.

[49] Donald J Kessler and Burton G Cour-Palais. Collision frequency of artificial satellites: The creation of a debris belt. *Journal of Geophysical Research: Space Physics*, 83(A6):2637–2646, 1978.

[50] Robert Kieschnick and Bruce D McCullough. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical modelling*, 3(3):193–213, 2003.

[51] Roger Koenker. A note on l-estimates for linear models. *Statistics & probability letters*, 2(6):323–325, 1984.

[52] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[53] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[54] Kenneth Lange. *Numerical analysis for statisticians*. Springer Science & Business Media, 2010.

[55] JW Mandema, DH Salinger, SW Baumgartner, and MA Gibbs. A dose–response meta-analysis for quantifying relative efficacy of biologics in rheumatoid arthritis. *Clinical Pharmacology & Therapeutics*, 90(6):828–835, 2011.

[56] François Mercier, Laurent Claret, Klaas Prins, and René Bruno. A model-based meta-analysis to compare efficacy and tolerability of tramadol and tapentadol for the treatment of chronic non-malignant pain. *Pain and therapy*, 3(1):31–44, 2014.

[57] John A Nelder and RJ Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

[58] Tereza Neocleous and Stephen Portnoy. On monotonicity of regression quantile functions. *Statistics & Probability Letters*, 78(10):1226–1229, 2008.

[59] Philip Paolino. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9(4):325–346, 2001.

[60] Russell P Patera. General method for calculating satellite collision probability. *Journal of Guidance, Control, and Dynamics*, 24(4):716–722, 2001.

[61] Franco Peracchi. On estimating conditional quantiles and distribution functions. *Computational statistics & data analysis*, 38(4):433–447, 2002.

[62] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, page 125. Vienna, 2003.

[63] Stuart J Pocock. *Clinical trials: a practical approach*. John Wiley & Sons, 2013.

[64] Joel R Primack. Debris and future space activities. In *University of California at Santa Cruz. Presented at the Conference on Future Security in Space, at New Place (Southampton, England) May*, volume 28, page 29, 2002.

[65] Nalini Ravishanker and Dipak K Dey. *A first course in linear model theory*. CRC Press, 2001.

[66] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

[67] Richard D Riley. Multivariate meta-analysis–is it worth the extra effort?

[68] Patrick Royston and Douglas G Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied statistics*, pages 429–467, 1994.

[69] Chandan Saha and Michael P Jones. Bias in the last observation carried forward method under informative dropout. *Journal of Statistical Planning and Inference*, 139(2):246–255, 2009.

[70] M Samanta. Non-parametric estimation of conditional quantiles. *Statistics & Probability Letters*, 7(5):407–412, 1989.

[71] WF Sheppard. New tables of the probability integral. *Biometrika*, 2(2):174–190, 1903.

[72] Alexandre B Simas, Wagner Barreto-Souza, and Andréa V Rocha. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 54(2):348–366, 2010.

[73] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[74] Robert A Stine. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031, 1985.

[75] Robert Stiratelli, Nan Laird, and James H Ware. Random-effects models for serial observations with binary response. *Biometrics*, pages 961–971, 1984.

[76] Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

[77] Winfried Stute. Conditional empirical processes. *The Annals of Statistics*, pages 638–647, 1986.

[78] Young K Truong. Asymptotic properties of kernel estimators based on local medians. *The Annals of Statistics*, pages 606–617, 1989.

[79] Jay Verkuilen and Michael Smithson. Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, 37(1):82–113, 2012.

[80] Yinghui Wei and Julian Higgins. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in medicine*, 32(17):2911–2934, 2013.

[81] Yichao Wu and Yufeng Liu. Stepwise multiple quantile regression estimation using non-crossing constraints. *Statistics and its Interface*, 2:299–310, 2009.

[82] Xiaoli Xu and Yongqing Xiong. A method for calculating collision probability between space objects. *arXiv preprint arXiv:1311.7216*, 2013.

[83] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.