UNIVERSITY OF CALIFORNIA,
IRVINE

Functional Analysis of Generalized Linear Models Under Nonlinear Constraints With Artificial Intelligence and Machine Learning Applications to the Sciences

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Management (Focusing on Mathematical Statistics)

by

K. P. Chowdhury

Dissertation Committee:
Associate Professor Weining Shen, Chair
Professor Knut Solna

2021

The parametric application of the methodology have already been published, "Functional analysis of generalized linear models under non-linear constraints with applications to identifying highly-cited papers" (© 2021 Elsevier and Journal of Informetrics). It is submitted for partial fulfillment of this dissertation with express permission of Elsevier under the author-use guidelines of Elsevier and Journal of Informetrics. Available here (last accessed: 02-01-2021), the personal use definition is reproduced here from the Elsevier website.

"Authors can use their articles, in full or in part, for a wide range of scholarly, non-commercial purposes as outlined below:

- Use by an author in the author's classroom teaching (including distribution of copies, paper or electronic)
- Distribution of copies (including through e-mail) to known research colleagues for their personal use (but not for Commercial Use)
- Inclusion in a thesis or dissertation (provided that this is not to be published commercially)
- Use in a subsequent compilation of the author's works
- Extending the Article to book-length form
- Preparation of other derivative works (but not for Commercial Use)
- Otherwise using or re-using portions or excerpts in other works

These rights apply for all Elsevier authors who publish their article as either a subscription article or an open access article. In all cases we require that all Elsevier authors always include a full acknowledgement and, if appropriate, a link to the final published version hosted on Science Direct."

DEDICATION

For Laura and Arabella.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to especially thank Weining and Knut for their capable, ethical and honest support and help in completing the dissertation and my doctorate. I would also like to acknowledge Laura Smith for her invaluable help in the organization of the various contents of the dissertation. I have enjoyed the many hours of discussions over these last few years regarding the applicability and usefulness of the mathematical findings with her. As such, the current version of the work would not have been possible without her capable inputs throughout this process.

Furthermore, I am also grateful to Journal of Informetrics, Elsevier and its entire editorial team as well as two anonymous reviewers whose comments helped strengthen the Parametric version of the paper considerably. As a result of their input the Parametric version is now published and may be accessed directly through the Journal of Informetrics website. I have also been given express permission by Elsevier regarding my ability to include this publication as part of my dissertation. In addition, to the materials submitted along with this dissertation from Elsevier regarding authorship rights you may find more information here. Under "Personal use" rights, as the sole author I have all relevant rights to use the paper as part of this Dissertation.

VITA

K. P. Chowdhury

RESEARCH INTERESTS

Bayesian Methodologies in Artificial Intelligence and Machine Learning Applications; Biostatistics; Bayesian Methodologies for Genetic Epidemiology; Generalized Linear Models; Model Evaluation,; Model Fit, Inference and Prediction (MIP); Methodologies for Categorical Data; Frequentist Estimation; Large-scale Non-convex Optimization; Graph Theory; Outlier Detection; Nonparametric Statistics; Dynamical Systems in the Sciences; Complex Networks; Mathematical Modeling; Algorithmic Game Theory; Variational Methods; Image Processing; Computational Mathematics.

EDUCATION

Ph.D. University of California, Irvine, CA. Expected Sep. 2021
Thesis: "Functional Analysis of Generalized Linear Models Under Nonlinear Constraints with Artificial Intelligence and Machine Learning Applications to the Sciences."

M.S. University of California, Irvine, CA. August 2018

M.S. **Applied Mathematics: Statistics**, California State University, Fullerton, CA. June 2014

B.A. **Business Economics and Pre-Medicine**, University of California, Riverside, CA. Graduated Magna Cum Laude; 4.0 major G.P.A. August 2005

PAPERS AND PUBLICATIONS

Refereed Journal Articles:

Chowdhury, K. P. (2021), Functional Analysis of Generalized Linear Models Under Nonlinear Constraints With Application to Identifying Highly-Cited Papers, *Journal of Informetrics, 15(1), 101112.*

Chowdhury, K. P. (2017), Supervised Machine Learning and Heuristic Algorithms for Outlier Detection in Irregular Spatiotemporal Datasets, *Journal of Environmental Informatics, 33(1), 1-16, March 2017. ISSN 1684-8799 (DOI: 10.3808/jei.2017003756).*

Chowdhury, K. P. et al. (2015), Regression of Algae Biomass over Variables with Disjoint Spatial Support, *Journal of Environmental Statistics, 7(5): 35-42.* Accepted Manuscript June, 2015.

Chowdhury, K. P. et al. (2006), An Economic Analysis of the Impact of Pay-For-Performance Initiatives on Physicians, Patients and Insurance Providers, *Indiana Health Law Review, 3, 348-369.* Accepted Manuscript Jan., 2006.

Refereed Conference Journal Articles:

Chowdhury, K. P., Shen, Weining (2021), Bayesian Latent Adaptive Deep Neural Networks on Locally Compact Hausdorff Spaces With Applications to Support Vector Machines, *2021 World Meeting of the International Society for Bayesian Analysis*, June 28.

Chowdhury, K. P., Shen, Weining (2021), Nonparametric Functional Analysis of Generalized Linear Models Under Non-linear Constraints, *Symposium on Data Science and Statistics*, June 1^{st}.

Chowdhury, K. P., (2019), Flexible Functional Specification in Hierarchical Bayesian Estimation of Discrete Choices (Generalized Linear Models), *Joint Statistical Meetings, Denver, Colorado, August* 2^{nd}.

Chowdhury, K. P., (2019), Flexible Functional Specification in Hierarchical Bayesian Estimation of Categorical Data, *Symposium on Data Science and Statistics*, June 1^{st}.

Chowdhury, K. P., (2006), An Economic Analysis of the Impact of Pay-For-Performance Initiatives on Physicians, Patients and Insurance Providers, *The Center for Law and Health, Indiana University School of Law - Indianapolis*, June 30^{th}.

CHAIRED CONFERENCE SESSIONS

CS01- Data Science Shaping the Financial World (2021), *Symposium on Data Science and Statistics.*

Presenters

BERT as a Filter to Detect Pharmaceutical Innovations in News Articles, *Martha Czernuszenko, The University of Virginia.*

Dissecting the 2015 Chinese Stock Market Crash, *Min Shu, University of Wisconsin-Stout.*

A Hierarchical Bayesian Approach to Detecting Structural Changes in Bank Liquidity Premia, *Padma Ranjini Sharma, Federal Reserve Bank of Kansas City.*

SELECTED PRESENTATIONS

Invited Talks and Panels

Chowdhury, K. P. et al., Symposium on Pay-For-Performance, *The Center for Law and Health, Indiana University School of Law - Indianapolis, June* 2006.

CONFERENCES AND WORKSHOPS

Joint Statistical Meetings - Virtual Aug. 2021

2021 World Meeting of the International Society for Bayesian Analysis - Virtual	June 2021
Symposium on Data Science and Statistics - Virtual	June 2021
Joint Mathematics Meetings - Virtual	Jan. 2021
Joint Statistical Meetings - Denver, CO	Aug. 2019
Preparing to Teach Statistics Workshop, Colorado	July 2019
Symposium on Data Science and Statistics - Bellevue, WA	June 2019
Joint Mathematics Meetings - San Diego, CA	Jan. 2018
Symposium on Pay-For-Performance - Indianapolis, Indiana	June 2006

PROFESSIONAL MEMBERSHIPS

American Statistical Association (ASA)	Jan. 2013 – Present
American Mathematical Society (AMS)	Feb. 2014 – Present
American Finance Association (AFA)	Jan. 2018 – Present
American Economic Association (AEA)	Jan 2019 – Present
Institute of Mathematical Statistics (IMS)	Jan 2020 – Present
International Society for Bayesian Analysis (ISBA)	Jan 2021 – Present

ABSTRACT OF THE DISSERTATION

Functional Analysis of Generalized Linear Models Under Nonlinear Constraints With Artificial Intelligence and Machine Learning Applications to the Sciences

By

K. P. Chowdhury

Doctor of Philosophy in Management (Focusing on Mathematical Statistics)

University of California, Irvine, 2021

Associate Professor Weining Shen, Chair
Professor Knut Solna

This thesis presents multiple fundamental mathematical contributions to Generalized Linear Models (GLMs) ubiquitous to the sciences. The methodologies considered are shown to overcome biased estimates for parameters of interest in the sciences through new mathematical results and their applications in both nonparametric and parametric settings. The results are shown to be uniformly better in comparison to existing widely used methods in the sciences. In extensive simulation studies the methodologies outperform existing Artificial Intelligence (AI) and Machine Learning (ML) methods in the sciences for all around better Model fits, Inference and Prediction (MIP) results without losing interpretability of the parameter estimates. This is because the mathematical construction and their accompanying mathematical foundations ensure that the estimation procedure strongly converges to the parameters of interest. In the first application, I present a parametric version of the methodology (© Elsevier and Journal of Informetrics) titled "Functional analysis of generalized linear models under non-linear constraints with applications to identifying highly-cited papers." In the second application, I extend this methodology in an entirely nonparametric setting which gives equivalent results to the parametric formulation under various circumstances, but may outperform it as well in others, especially if the underlying Data Generating Process (DGP)

is asymmetric. Furthermore, I show that the categorical data models on which the methodologies are applied can be extended to any GLM, continuous or otherwise, while maintaining model interpretability and convergence results. In addition, I present a new prediction performance diagnostic statistic, called Adjusted ROC Statistic (ARS), which allows us to compare whether the prediction performance of various models fitted are statistically different. The nonparametric methodology is then further extended to give a new formulation of the binary regression framework widely used in the sciences. Through extensive simulation studies I show that this version of the methodology is more robust than the previous versions discussed. This general framework is then extended to various AI and ML applications widely used in the sciences. The entirety of the work also has some important consequences for our continued discussion on "statistical significance" vs. "scientific significance." This includes the need for us to consider the strength of convergence of our methodology in addition to the subtle connections between Topological Spaces and Measure Spaces. Each of which are crucial to ensure almost sure convergence of the parameter estimates through the estimation algorithm presented termed, Latent Adaptive Hierarchical EM Like algorithm or LAHEML. As such, the results present a significantly expanded and more accurate toolset for Mathematicians, Statisticians, Scientists and Decision Makers at all levels for better model fit, inference and prediction outcomes.

Chapter 1

Introduction

This dissertation presents several extensions to the current Generalized Linear Model (GLM) through rigorous mathematical formulations of the underlying preliminaries. The extensions are applied through nonparametric and parametric applications to binary observed outcomes as a function of some measured explanatory variables. Such observed phenomena are ubiquitous to the sciences and hence, their multivariate extensions in ordered and unordered models remain important in modern Artificial Intelligence (AI) and Machine Learning (ML) settings. Accordingly, any improvement on Model Fit, Inference, and Prediction (MIP) results for binary outcomes remain relevant for numerous AI and ML applications in the sciences.

The contributions rely upon various insights on the limitations of the current GLM framework through pointwise discontinuity of the link function, which relates the observed outcomes to the mean of a particular model specification. In the presence of such discontinuity, I show that no estimation approach, whether Frequentist or Bayesian, can ensure that the underlying restrictions on the link condition is always satisfied for any particular estimation iteration. Thus, by using the likelihood principle, this limitation is shown to have severe

restrictions on the current regression framework, extending straightforwardly to various AI and ML applications.

In particular, I argue that the presence of such a limitation on the underlying GLM formulation implies that almost sure convergence cannot be asserted for GLMs, such as those in the exponential family including the Logistic or Probit regressions. As such, the findings provide a focused reason as to empirical variability in results observed across the sciences for MIP results. To overcome these subtle but pernicious limitations, I present a new latent variable framework rooted in Real and Functional Analysis that appears to be novel to the sciences for the parametric formulation in the Bayesian framework, and is entirely novel for the nonparametric application in either the Bayesian or Frequentist formulation. The nonparametric application is then further extended to give the most robust functional specification for any binary outcome model specification through this methodology, beyond the binary regression specification accepted in the sciences.

Naturally, as the most basic formulation of a categorical outcome model, the methodology can readily be extended to more complicated AI and ML methods. Accordingly, I then use the robust parametric and nonparametric approaches to extend current AI and ML applications of Artificial Neural Networks (ANN), Regression Trees, Regression Forests as well as for Support Vector Machines (SVM). Finally, these updated model specifications are then used in real data applications across various disciplines such as Biostatistics and Physics. Thus, below I outline how the rest of the dissertations is structured.

1.1 Outline and Specific Contributions

The rest of the dissertation is structured as follows.

Chapter 2 provides the published version of the parametric methodology with minor variations to that published in Chowdhury (2021a). This paper lays the foundations to overcome pointwise discontinuity of the Logistic regression functional specification. It maintains much of the existing latent varible framework as in Albert and Chib (1993), but is shown to give superior results to it in various settings. The specific highlights are,

1. Robust functional form contains true parameters far more often than popular models.
2. Matches/outperforms widely used regression and Neural Network models.
3. Finds appropriate balance between Model Fit, Inference, and Prediction (MIPs).
4. Introduces new large-sample DGP test; can use to improve A.I. models.
5. For MIS field finds Popularity Parameter to be important for predicting citations.

Chapter 3 presents a continuation and extension of the methodology in Chapter 2 in a nonparametric setting. It shows that the parametric version of the methodology is nested within it and further makes the following contributions,

1. It provides rigrous mathematical foundations under which the link condition holding for each observation can be used to identify the true parameters for any particular model specification almost surely.
2. It presents a Latent Adaptive Hierarchical EM Like algorithm (LAHEML) that can be used in a completely nonparametric setting, without violating link function continuity.

3. The methodology is shown to be superior to the parametric version giving results superior to it especially if the underlying DGP is asymmetric.

4. Despite being nonparametric it does not lose interpretability of the parameter estimates.

5. It can either match or outperform all other models compared in regards to Inference and Prediction.

6. It can either match or outperform all other models compared in regards to Model Fit.

7. It can outperform ANN in prediction outcomes in both in-sample and out-of-sample data without losing interpretability of parameter estimates particularly when the data size is small. The results are especially relevant for Test Datasets (TeDs).

8. I present a new prediction comparison model evaluation statistic based on Chowdhury (2019) which is more general and call it Adjusted ROC Statistics (ARS). In addition, I further give its limiting distribution.

9. The methodology enables us to perform a large-sample asymptotic test to check whether any parametric distributional assumption, as a function of the estimated $\boldsymbol{\beta}'s$, hold! Accordingly, it is shown to be an extension of Li et al. (2018).

10. Thus, it allows us to check the adequacy of any model assumption on the underlying categorical outcomes and provides a ready test for statistical divergence.

Chapter 4 is the culmination of the previous two chapters which extends those results to beyond the binary regression assumptions. Its specific contributions are

1. Provides a robust functional specification for binary GLMs which is far more general than the existing GLM framework.

2. In doing so, it can identify the true underlying parameters almost surely even when the assumptions of the current binary regressions are violated.

3. It nests the model formulations in Chapter 2 and Chapter 3.

4. It can match or outperform all other models compared in regards to Inference and Prediction.

5. It can match or outperform all other models compared in regards to Model Fit when the datasets are more unbalanced or smaller.

6. It can match or outperform Neural Networks in prediction outcomes in both in-sample (TrD) and out-of-sample (TeD) data without losing interpretability of parameter estimates.

Chapter 5 presents more complicated model specifications for the methodologies specified in Chapter 2, Chapter 3 and Chapter 4. In particular, they are extended to many AI and ML applications widely recognized for their predictive power across the sciences such as ANN, Regression Trees, Regression Forests and SVM. In this chapter I further extend the methodologies to Broad Neural Networks and give an universal approximation theorem for such Natural Neural Networks (NNN or N^3).

Chapter 6 discusses future extensions and gives some further concluding thoughts.

Chapter 2

Parametric Application

2.1 Introduction

Binary models are central to scientific inquiry across many different fields including Informetrics, Informatics, Scientometrics and Bibliometrics as well as, Statistical, Biomedical, Social and Physical Sciences. It is particularly important for citation prediction (Wang, Wang, & Chen, 2019; Abrishami & Aliakbary, 2019). For example, Uddin and Khan (2016) used regression analyses to understand the impact of keywords on citation counts, and found that the author-defined keywords were statistically significant in explaining number of citations. Similarly, Sohrabi and Iraj (2017) used Logistic regression with repetitive keywords in article abstracts and keyword frequency per journal as independent variables, to show both were statistically significant in predicting citation counts. Therefore, it remains critical to understand model fit, inference and prediction MIP(s) performance of these binary regression models and their robustness for scientific inquiry.

To understand such Bernoulli outcome models, there are multiple statistical and econometric formulations, such as the "Binary Outcom" (BO) and "Latent Variable Outcome" (LVO) models. Unfortunately, the underlying assumptions of BO vs. LVO models are distinctly different, and it is often not clear in the literature as to which approach to take and how to reconcile any divergence in MIPs. A further complication is presented when the data are unbalanced (WLOG, more 0's than 1's in the Bernoulli outcome case), which is almost always of practical importance in applied settings. In addition, given the assumptions of the traditional logistic and Probit formulations, the link function (for example for the Logit BO model it is the log-odds function) is symmetric for both BO or LVO formulations. This implies that the probability of success or failure approaches 1 or 0, respectively at the same rate (see for example Agresty and Kateri 2011), an assumption frequently violated in real world data applications. Accordingly, it is well established that the parameter estimates in these models are susceptible to bias and inconsistency (e.g., Simonoff 1998, Abramson et al. 2000, Maity et al. 2018). This paper introduces a new functional analyses perspective

applied to Bernoulli outcomes in the familiar regression framework, that seeks to overcome these inconsistencies.

Broadly, any outcome variable can be modeled as a linear (LM) or as a non-linear (NLM) model or function of the explanatory variables. Here I broadly refer to these specifications as Generalized Linear Models (GLM), where we consider

$$E[\mathbf{Y}|\mathbf{X}] = \mathbf{c}(\mathbf{X})\beta + \epsilon. \tag{2.1}$$

Here the $n \times 1$ outcome variable Y is related to a $n \times (k+1)$ set of explanatory variables, $\mathbf{X} = (1, X_1, ..., X_k)$, through a continuous, bounded, real valued function $\mathbf{c}(\mathbf{X})$ of the same dimensions. The $(k+1) \times 1$ parameters of interest are $\beta = \{\beta_1, ..., \beta_{k+1}\}$. If $\mathbf{c}(\mathbf{X}) = I(\mathbf{X})$, where $I()$ is the identity link function, we have the well known LM widely used in the sciences. As is customary the expectation of the error term is also assumed to be 0.

I discuss both BO and LVO models in detail in section 3.2. However, for the present discussion it suffices to state that both models for the Bernoulli outcome case remains relevant for Artificial Intelligence (AI) and Machine Learning (ML) applications (Li et al. 2018). This is because they serve as the building blocks for various Multinomial extensions (e.g., Allenby and Rossi 1998, Murad et al. 2003). However, their usage appears to be field specific. For example, Hu et al. (2020) show the efficacy of the Logistic regression in identifying highly cited papers over four other classification techniques including c4.5, Support Vector Machine (SVM) and Artificial Neural Networks (ANN). In doing so, they highlight not only the importance of Journal Impact Factor (JIF) (e.g. Bai et al. 2019, Bornmann et al. 2014, Tsai 2014) and word embedding techniques (i.e. Zhang et al. 2018) in classifying potentially highly-cited papers, but also of Keyword popularity (KP) measures in both Marketing and Management Information Systems journals. Similarly, in Econometrics LVO models has been used to understand behavior of the average individual within a population

(see Greene 2003 for a summary), for calculating propensity scores for causal interpretation and program evaluation (see Imbens and Rubin 2015 for an excellent summary), as well as to understand the degree of heterogeneity through finite and infinite mixture distributions (Andrews et al. 2002). In Psychology (e.g., Talukder 2008, Hofmans 2017); Experimental Economics (e.g., Edelman et al. 2017, Hallsworth et al. 2017); Biomedical Sciences, (e.g., Zhang et al. 2017, Davison et al. 2017, Mandal 2017) and in the Physical Sciences (e.g., Hatlab et al. 2018, Beita-Antero et al. 2018) there is a rich history of both formulations. Evidently the methodology used is context and field specific, with inferences drawn based on established, field specific criteria[1]. Furthermore, even in the presence of AI methods such as ANN and ML methods such as SVM, since the Logistic regression can give better model fits, prediction and inference, its application and improvements remain highly relevant for any classification exercise.

Therefore, this contribution seeks to reconcile some of these incongruities in traditional widely used Bernoulli outcome models with specific focus on the Logistic regression. Its contribution is four fold. First, I present a new functional specification which ensures that traditional i.i.d. regression model assumptions hold for each $y_i \in \mathbf{Y}$ (y_i is 1×1) and $\mathbf{x_i} \in \mathbf{X}$ ($\mathbf{x_i}$ is $(k+1) \times 1$) and which corrects for much of these induced biases in regression parameter estimates in widely used existing models. To ensure comparability to existing methods, I further ensure that the new specification is isomorphic to existing models if the data actually support them. Second, to aid in model comparison between the existing and proposed models, I introduce an asymptotic test for congruence of parameter estimates of the proposed and existing models. I then present estimation algorithms for the Logistic regression formulation of the new model in both frequentist and Bayesian frameworks[2].

[1]This is a result of the applicability of the specifications above being relevant to field specific questions. For example, in Business and Economics one may ask, whether consumers receive more utility from products they buy, where as in the Biomedical Sciences we may be concerned with whether a particular drug is more effective than current alternatives.

[2]I stress however, that as the new formulation becomes a constrained optimization problem it can be time sensitive for large datasets in the frequentist case.

As such a new Bayesian Hierarchical estimation methodology is used for simulation and Scientometric applications. Accordingly, I show the proposed methodology applied in the Logistic case, requires roughly two-thirds the number of Markov Chain Monte Carlo (MCMC) iterations for convergence as opposed to existing LVO Bayesian models. It does so while giving better MIP results compared to existing models (whether existing BO or LVO models are compared), including AI methods such as Artificial Neural Networks (ANN), in myriads of circumstances. The results are shown to be robust in general, but are especially relevant when the assumptions of more traditional models are violated. Finally, I reintroduce an ROC based predictive statistic ROC-Statistic (RS) (Chowdhury, K. P., 2019) to show the interplay and importance of MIPs of the new methodology in Informetrics, Informatics, Statistics and Applied Mathematics in general.

Thus, the remainder of this article is organized as follows. I first discuss the preliminaries and set up of existing Binary Outcome and Latent Variable Outcome models in section 2.2. I then discuss under what circumstances they are equivalent. Then I expand on the proposed methodology and give the proofs for existence and uniqueness of the parameter estimates for the new functional specifications in Section 2.3 deferring all technical proofs to the appendix. To specifically apply this model, I consider the Logistic regression specification and then give estimation procedures in the Bayesian framework in Section 2.4 (the frequentist algorithm is deferred to Appendix .1.2). To ensure comparability of models, I then present an asymptotic test that allows us to compare model fit congruence between existing and proposed models in Section 2.5. This is followed by extensive numerical simulations in Section 2.6 and an application to classification of highly-cited papers in Section 2.7. This is done using the Logistic formulation for the proposed model under varying data generating processes (DGPs), sample sizes and unbalancedness specifications. I then discuss the importance and broad applicability of the methodology for MIP results in Section 2.8 and finally end with some concluding thoughts in Section 2.9.

2.2 Preliminaries

Any Bernoulli outcome regression model with accompanying covariates $(\mathbf{Y}, \mathbf{X})$ (defined as in the introduction in section 2.1) can be modeled using a Binary Outcome (BO) or Latent Variable (LV) model specification. However, the assumptions under each are meaningfully different. The Binary Outcome Model has the following assumptions 1-3.

Assumption 2.1. *$y_i \in \mathbf{Y}$ are independent and Bernoulli distributed such that*

$$y_i = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1-p. \end{cases} \tag{2.2}$$

Assumption 2.2. *The systematic component of the explanatory variables are considered fixed.*

Assumption 2.3. *There is a link function $g(\mu) = \mathbf{Y}$ that relates the mean of an observation to the systematic component.*

In contrast the LV models have assumptions 4-5.

Assumption 2.4. *$y_i \in Y$ are independent and identically distributed observed across some threshold $m \in \mathbf{R}$ such that there exists another random variable $\mathbf{Y}^*$ with $y_i \in \mathbf{Y}$ and $y_i^* \in \mathbf{Y}^*$ satisfying,*

$$y_i = \begin{cases} 1 \; if \; y_i^* > m \\ 0 \; if \; y_i^* \leq m. \end{cases} \tag{2.3}$$

Assumption 2.5. *The systematic component of the explanatory variables are considered fixed.*

Clearly[3] the underlying measure spaces on which the models are defined are distinct. Yet under both models the decision to be made by the modeler amounts to choosing a distribution on the error term and deciding which variables to use for the systematic components. Under the LV model a further distributional specification on $\mathbf{Y}^*$ must be made. If it is assumed to be Bernoulli, the two models are indeed identical for the appropriate success probabilities. Furthermore, it is not difficult to see that though no specific assumption is made on the link function for LV models, a specification on the error, also known as the probability of success, leads to a deterministic functional specification of the link function in Assumption 2.3 above.

For example, let $\mathbf{P}(\mathbf{x_i})$ be the probability of success for the linear Logistic BO model (Luce 1959) for the i^{th} observation. Then,

$$\mathbf{P}(\mathbf{x_i}) = \frac{exp[\lambda(\mathbf{x_i}, \beta)]}{(1 + exp[\lambda(\mathbf{x_i}, \beta)])}, \tag{2.4}$$

where $\lambda(\mathbf{x_i}) = \beta'\mathbf{x_i}$, in LM and $\beta = \{\beta_1, ..., \beta_{k+1}\}$. Then,

$$ln\{\mathbf{P}(\mathbf{x_i})/(1 - \mathbf{P}(\mathbf{x_i}))\} = \lambda(\mathbf{x_i}, \beta). \tag{2.5}$$

The quantity on the left in (2.5) is the familiar log-odds ratio for the Logit and is the link function assumed to hold in expectation (average) in Assumptions 2.1–2.3 for i.i.d. $y_i \in Y$ (for the Probit model Thurstone 1927, $P(\mathbf{X})$ has the Standard Normal formulation). Now consider the LV model for the Logistic regression, and restrict the threshold m to be 0, and let the probability of success be the Logistic distribution. Then,

$$y_i^* = \beta'\mathbf{x_i} + \epsilon_i^* \iff \mathbf{P}(\mathbf{x_i}) = Pr(y_i^* > m) = F(-\epsilon_i^* < \beta'\mathbf{x_i}) = F(\beta'\mathbf{x_i}),\ m = 0, \tag{2.6}$$

[3]All of these assumptions above are in addition to the Full Rank and Non-Micronumerosity assumptions for the explanatory variables that accompany these traditional models.

where F, is the cdf of $-\epsilon_i^*$ (and ϵ_i^* by symmetry). It is well known that if $m = 0$, a Logistic distribution assumption on the error gives us the same link function specification under both BO and LVO models (Cameron and Trivedi 2010).

Evidently, though no assumption is made on the link function specification in LV models, by construction of GLM the link condition Assumption 2.3 of BO model is identical to LV model at least in this simple linear regression model. Furthermore, by independence, this link condition should hold for every observation and not just in expectation as in the traditional BO model framework. Below I extend this insight to Generalized Linear Models (GLMs), under any specification of the error term in BO or LV models such that the link condition holds for every observation.

2.3 Methodology

In this section I lay the groundwork for the viability of the model specification. In order to retain the current models should the data support them, I propose a more general framework. In particular, I show that any link function, corresponding to a particular GLM, can be thought of as coming from a family of link functions. I parameterize this family through two parameters α and δ and show that all existing GLMs correspond to particular values of them. To ensure identifiability and equivalency to existing models, without loss of generality I focus the methodology to depend on only one parameter α^*, a function of $\{\alpha, \delta\}$. Accordingly, below I first present the generalized link function, followed by an application of it to the Logit generalized link. To motivate identifiability of this new link function specification, I first prove under what circumstances uniqueness and existence is guaranteed for the Logit. From it we may deduce and expand the proof of existence and uniqueness to any GLM in the discrete outcome case for the model and proceed accordingly[4].

[4]Please note that all non-obvious vectors and matrices are represented using bold notations.

2.3.1 Generalized Link Function

Consider any link function, g(), that satisfies the regularity conditions (continuous, real valued and analytic) for any specification of the error distribution, F(), for LV or BO models. Let $\lambda(\mathbf{X}, \beta) = \mathbf{c}(\mathbf{X})\beta$, where $\mathbf{c}$ is a continuous, bounded, real-valued function of $\mathbf{X}$, the $(n \times (k+1))$ matrix of covariates or explanatory variables (thus, $\mathbf{c}(\mathbf{X})$ is also $(n \times (k+1))$). Then by construction,

$$g(\mathbf{P}(\mathbf{X})) = \lambda(\mathbf{X}, \beta) = \mathbf{c}(\mathbf{X})\beta \iff \mathbf{P}(\mathbf{X}) - g^{-1}(\mathbf{c}(\mathbf{X})\beta) = \mathbf{0}_{n\times 1}. \tag{2.7}$$

Since we know that (2.7) is not always satisfied for the i^{th} observation, we would like to ensure that the constraint holds so that we can conditionally estimate β with more accuracy. Therefore, when it is not satisfied consider,

$$\mathbf{P}(\mathbf{X}) = (g^{-1}(\mathbf{c}(\mathbf{X}))\beta)^{\alpha^*} \iff \alpha^* = log(\mathbf{P}(\mathbf{X}))(log((g^{-1}(\mathbf{c}(\mathbf{X})\beta))))^{-1}, \alpha^* \in \mathbf{R^n}. \tag{2.8}$$

Since for certain link functions (2.8) cannot be uniquely identified, much of the contribution of this paper relates to how this non-trivial problem can be overcome while maintaining equivalency to the current framework if the data support them. As an example, the Logit has the log-odds ratio as the link function, meaning that for uniqueness, we must incorporate some restrictions on the numerator or the denominator of the odds function. However, such restrictions can easily diverge from current GLM specifications, and we need a more general definition of the link function.

Accordingly, let us hypothesize that the actual fitted link for a particular GLM belongs instead to a family of link functions with parameters $\{\alpha, \delta\}$ (Pregibon 1980) for each obser-

vation i. The principle assumption is that any fitted link, such as the log-odds for the Logit, belongs to a family of link functions for different values of $\{\alpha, \delta\}$. Let us then consider any GLM in which given an assumption imposed on the probability of success (Logistic, Standard Normal, Extreme Value Type I etc.), we wish to hypothesize a link function $g()$ such that for the i^{th} observation the following condition holds,

$$y_i = g(\mathbf{x}_i, \beta, \alpha_i, \delta_i) = \lambda(\mathbf{c}(x_i), \beta). \tag{2.9}$$

Critically, through this formulation, for differing parameter values we can induce a symmetric or asymmetric behavior for a particular link function specification. In particular through an assumption either on the probability of success or the error term we may hypothesize,

$$Hypothesized\ Link : g_0(\mathbf{x}_i, \beta; \alpha_i, \delta_i) = g(\mathbf{x}_i, \beta; \alpha_i = \alpha_0, \delta_i = \delta_0), \tag{2.10}$$

for specific values of $\{\alpha_0, \delta_0\}$. In reality however, our data may suggest a functional specification in the same link family but with different parameters, say $\{\alpha_i^*, \delta_i^*\}$,

$$Correct\ Link : g^*(\mathbf{x}_i, \beta; \alpha_i, \delta_i) = g(\mathbf{x}_i, \beta; \alpha_i = \alpha_i^*, \delta_i = \delta_i^*). \tag{2.11}$$

Crucially, (2.11) ensures that for some values of this family, the link condition will always hold with equality for any GLM for every observation of the regression model. To show existence of this specification, a necessary and sufficient condition is that the family of link functions is analytic under i.i.d. assumptions. Therefore, the estimation process becomes,

$$argmin_\beta (Y - \mathbf{c}(\mathbf{X})\beta)^d \ s.t.\ g^*(\mathbf{x}_i, \beta; \alpha_i, \delta_i) = \lambda(\mathbf{c}(\mathbf{x_i}), \beta),\ \forall i \in \{1, ..., n\}; 1 \leq d < \infty, \tag{2.12}$$

where p represents the appropriate p-norm in $L^p(E)$[5]. The proof of this statement can be

[5]Where for a measureable set E I define $L^p(E)$ to be the collection of measureable functions f for which $|f|^p$ has a finite integral over E.

found in Theorem 2.5. Since by construction, observations are independent, one can further impose

$$\mathbf{E}(\alpha^*) = \mathbf{E}(\alpha_i); \mathbf{E}(\delta^*) = \mathbf{E}(\delta_i), \tag{2.13}$$

which follows easily from our identically distributed assumption. Thus,

$$\mathbf{E}[g^*(\mathbf{x}_i, \beta; \alpha_i, \delta_i)] = \mathbf{E}[\lambda(\mathbf{c}(\mathbf{x_i}), \beta)] = \mathbf{E}[\lambda(\mathbf{c}(\mathbf{X}), \beta)], \forall i \in \{1, ..., n\}. \tag{2.14}$$

Therefore, we need only ensure that this assumption holds for each of the i^{th} observations. To show the importance of this formulation, below I first prove the existence and uniqueness of (2.11) for the Logistic model, through first a Generalized Odds function and then by the Generalized Log-Odds function. I then show that in this formulation the Generalized Logistic Link function is analytic, and therefore, we can approximate the link condition holding for each observation. From this specific application to the Logistic I then deduce and prove the existence and uniqueness results for all GLMs. This ensures that the link constraint can be approximated to hold across all observations, such that the parameters β can be conditionally estimated in the Bayesian framework or solved through constrained optimization. The results follow below where a bold notation indicates a vector or matrix unless already defined accordingly.

2.3.2 Generalized Odds Function

Consider the following specifications for the Odds function, where $\mathbf{P}_{n\times 1}$ is the probability of success and element-wise does not equal either 0 or 1 identically.

$$g_0(\mu, \alpha, \delta) = \frac{\mathbf{P}^{\alpha}}{(\mathbf{1}-\mathbf{P})^{\delta}}. \tag{2.15}$$

Theorem 2.1. *The Generalized Odds function is uniquely identified for some* $\alpha^* \in \mathbf{R^n} \setminus \{-\infty, \infty\}, \mathbf{P}_{n\times 1} \notin \{0,1\}_{n\times 1}$ *s.t.*

$$g_0(\mu, \alpha^*, \delta^* = \mathbf{1}) = \mathbf{P}^{\alpha^*}(\mathbf{1}-\mathbf{P})^{-1}. \tag{2.16}$$

Proof. To prove that the proposed family of functions can only be identified up to a monotonic transformation for either α or δ, but not both for the i^{th} observation consider,

$$g_0(\mu, \alpha_i, \delta_i)^{(1/\delta_i)} = \frac{P_i^{\alpha_i/\delta_i}}{(1-P_i)}, \tag{2.17}$$

where $\mathbf{P(x_i)} = P_i$. WLOG hold $\delta_i \in \mathbf{R} \setminus \{-\infty, \infty, 0\}$ fixed (since element-wise $\alpha^* = \frac{\alpha}{\delta} \neq \pm\infty, \delta \neq 0$ by construction). Since by construction $P_i \in (0,1)$,

$$\lim_{\alpha_i \to \infty} \frac{P_i^{\alpha_i/\delta_i}}{(1-P_i)} = 0 \; and \; \lim_{\alpha_i \to -\infty} \frac{P_i^{\alpha_i/\delta_i}}{(1-P_i)} = \infty. \tag{2.18}$$

Thus, α_i or δ_i cannot both be $-\infty$ or ∞ at the same time. Let us fix δ_i such that it is not $\infty,\ -\infty$ *or* 0. Then by the arguments preceding α_i can be ∞. However, since such a set

has lebesgue measure 0, we can safely restrict our attention to

$$\mathbf{x_i} \textit{ such that } \{P_i : \{\alpha_i, \delta_i\} \notin \{-\infty\ , \infty\} \textit{ and } \delta_i \neq 0\}. \tag{2.19}$$

Having restricted our attention to the constrained values of $\{\alpha_i, \delta_i\}$, let us fix δ_i. Then by the density of the rationals in the reals, for any

$$\alpha_i \in R \setminus \{-\infty, \infty\}, \tag{2.20}$$

if $\delta_i^* = 1$ there exists an $\alpha_i^* \in R$ such that $\alpha_i^* = \frac{\alpha_i}{\delta_i}$. Therefore, the generalized odds function can be given by,

$$g_0(\mu, \alpha_i, \delta_i)^{(1/\delta_i)} = g_0(\mu, \alpha_i^*, \delta_i^* = 1), \tag{2.21}$$

The n-dimensional result then easily follows under the independence assumption of each observation as needed. □

It is then straight forward to show that a generalized Logistic link family may be defined through a monotonic transformation of the Generalized Odds function.

Proposition 2.1. *There exists a family of link functions given by a monotonic transformation of the Generalized Odds function,* $\mathbf{P}^{\alpha^*}(\mathbf{1} - \mathbf{P})^{-1}$ *such that for* $\{\alpha^* = \mathbf{1}, \delta^* = \mathbf{1}\}$ *it represents the Generalized Logistic Link function for each observation.*

To ensure the Generalized Logistic Link function can be approximated through Taylor approximations, I also show that it is analytic. The rigorous proof of the statement is given in the Appendix (.1.1), and as a result the model can interpolate values for link conditions even

when the current GLM framework cannot ($\{-\infty, \infty\}$ can result in the current estimation processes).

Theorem 2.2. *The Generalized Logistic Link function, $log(\mathbf{P}^{\alpha^*}(\mathbf{1}-\mathbf{P})^{-1})$, is Analytic.*

This is a sufficient condition for the existence of Taylor approximations and convergence of regression parameters of interest, conditionally on α_i^*, for all observations. Thus, it remains to show that the results above hold for any specification of the Generalized Logistic Link function with extensions to all GLMs, under the assumptions of the current GLM framework. The theorems below establishe these results.

Theorem 2.3. *There is an unique solution to the link modification problem for the Generalized Logistic GLM formulation where the link constraint is binding for some $\alpha^* \in R^n \setminus \{-\infty, \infty\}$, given $\mathbf{P_i} \notin \{0,1\}$, $\mathbf{x_i} \notin \{0, \infty, -\infty\}$ for each $i \in \{1, ..., n\}$ and $\beta_j \notin \{\infty, -\infty\}$ with $j \in \{1, ..., (k+1)\}$.*

A somewhat technical proof of this result is given in the Appendix (.1.1). Using this result, I provide the foundations for the extension of the specific result to all GLMs below.

Theorem 2.4. *There is an unique solution to any link modification problem, where the link constraint holds with equality in the Generalized Linear Model Framework for some $\alpha^* \in R^n \setminus \{-\infty, \infty\}$, given $\mathbf{P_i} \notin \{0,1\}$, $\mathbf{x_i} \notin \{0, \infty, -\infty\}$ for each $i \in \{1, ..., n\}$ and $\beta_j \notin \{\infty, -\infty\}$ with $j \in \{1, ..., (k+1)\}$.*

Proof. Consider as before that $\mathbf{P_i} \neq \{0,1\}$ and $|\beta| < \infty$. Then if α_i^* is the unique value attained by fixing δ_i for each i,

$$\alpha^* = log(P(\mathbf{X}))(log((g^{-1}(\mathbf{c}(\mathbf{X})\beta))))^{-1}, \alpha^* \in \mathbf{R^n} \setminus \{-\infty, \infty\}. \tag{2.22}$$

Note that α^* is $n \times 1$. As long as element-wise,

$$\mathbf{P(X)} \neq \mathbf{0}_{n\times 1} \; and \; g^{-1}(\mathbf{c}(\mathbf{X})\beta)) \neq \mathbf{0}_{n\times 1}, \tag{2.23}$$

(2.22) has a specification and a solution, by the same argument as I had proceeded with the Generalized Logistic Link. Thus, it remains to show the immediately preceding two equations do not hold for any i. Note that by construction of a GLM through independence,

$$\mathbf{E}[g(\mathbf{P}(\mathbf{X}))] = \mathbf{E}[(\beta'\mathbf{c}(\mathbf{x_i}))]. \tag{2.24}$$

Therefore, a sufficient condition for (2.22) to hold means that $(\beta'\mathbf{c}(\mathbf{x_i})) \neq 0$. Let us conjecture otherwise and say that this does not hold. Then either $\mathbf{c}(\mathbf{x_i}) = \mathbf{0}_{(k+1)\times 1}$ or $\beta = \mathbf{0}_{(k+1)\times 1}$. One can safely discard the possibility of $\mathbf{c}(\mathbf{x_i}) = \mathbf{0}_{(k+1)\times 1}$, since that implies there is no explanatory variables to understand probability of success, i.e. there does not exist a GLM. If on the other hand, $\beta = \{\beta_j\}_{j=1}^{(k+1)} = \mathbf{0}_{(k+1)\times 1}$, for any $j \in ((k+1)\times 1)$, then that implies the explanatory variables used in the GLM are not adequate to describe a relationship to the dependent variable, and as such, should not be considered in the regression specification. Therefore, $\mathbf{E}[g(P(\mathbf{X}))] = (\beta'\mathbf{c}(\mathbf{x_i})) \neq 0$ under the model preliminaries and assumptions of GLMs. As such, the existence of the Taylor Approximation to the functional forms under consideration is implied by the existence of a GLM and its assumptions. Thus, we can proceed by the intermediate value theorem to show that there exists an unique solution to the link modification problem for any GLM, as needed. □

THEOREM 2.5. *For any continuous and bounded specification of a Generalized Linear Model there exists a solution to* $argmin_{\beta}(\mathbf{Y} - \mathbf{c}(\mathbf{X})\beta)^d$ *s.t.* $g^*(\mathbf{x}_i, \beta; \alpha_i, \delta_i) = \lambda(\mathbf{c}(\mathbf{x_i}), \beta),\ \forall i \in \{1, ..., n\}; 1 \leq d < \infty; \{|\mathbf{c}(\mathbf{X})|, |\beta|\} < \infty$.

Proof. To see this[6], note that $\lambda(\mathbf{c}(\mathbf{x_i}, \beta))$ is continuous by assumption of GLM. Thus, $\lambda(\mathbf{c}(\mathbf{x_i}, \beta))$ is lesbesgue measureable on our domain of choice $\mathbf{R} \setminus \{\infty, -\infty\}$. Let $E \subseteq \mathbf{R} \setminus \{\infty, -\infty\}$. Then for each continuous and bounded function f in $L^p(E)$ there exists closed and bounded intervals $[a, b]_j$ such that

$$\cup_{j=1}^{n}[a, b]_j = E,\ n \in \{1, 2, ...\} \tag{2.25}$$

where f vanishes outside of E when restricted to $\cup_{j=1}^{n}[a, b]_j$. Thus, f vanishes outside of E. Therefore, each f is the limit of a sequence of piecewise linear, continuous functions which can be represented by $\lambda(\mathbf{c}(\mathbf{x_i}, \beta))$. If F is taken to be the union of all of these approximating sequences, then F is dense in $\mathbf{R} \setminus \{-\infty, \infty\}$ and the statement follows. □

The above results show the theoretical foundations of the methodology are consistent with the existing GLM framework. However, in many cases no analytical solutions to β as a function of α^* may exist (for example in the Logistic formulation). Consequently, the convergence to true population parameters is also a non-trivial problem. Therefore, I now detail an estimation procedure for the Logistic regression application, in both the frequentist (Appendix [.1.2]) and in a full-probability Bayesian formulation (for BO or LV models), that guarantees the convergence to true population parameters with very few MCMC iterations[7].

[6]The result follows readily from the continuous, real valued assumptions on the GLM functional specification and I follow the standard arguments given in most graduate level Real Analysis books.

[7]The estimation procedures are further shown to be applicable in any GLM because of the uniqueness of α^* given Theorem 2.5.

2.4 Estimation

To illustrate the viability of the proposed model, I apply it to a specific Generalized Linear Model, the well known Logistic regression. Thus, in this section I present a Bayesian Hierarchical method to estimate the Logistic regression model under this new proposed methodology (the frequentist application can be found in the appendix [.1.2]). The extension of these algorithms to any GLM, follows similarly from the existence and uniqueness results discussed previously.

2.4.1 Estimation of the Generalized Logistic Link

Given the linear Logistic regression under either the BO or LV models, note that

$$log\left\{\frac{\mathbf{P}^{\alpha^*}}{(1-\mathbf{P})}\right\} = \lambda(\mathbf{c}(\mathbf{X}),\beta) \iff log\left\{\frac{F(\lambda(\mathbf{c}(\mathbf{X}),\beta))^{\alpha^*}}{(1-F(\lambda(\mathbf{c}(\mathbf{X}),\beta)))}\right\} - \lambda(\mathbf{c}(\mathbf{X}),\beta) = 0,$$

$$\implies \alpha^* = \frac{\lambda(\mathbf{c}(\mathbf{X}),\beta) + log(1-F(\lambda(\mathbf{c}(\mathbf{X}),\beta)))}{log(F(\lambda(\mathbf{c}(\mathbf{X}),\beta)))}. \tag{2.26}$$

Clearly, there is no analytical solution here for $\beta|\alpha^*$. However, for any particular value of β, we can solve for $\{\alpha^*|\beta, \delta^* = 1\}$ on a grid, through sequential iteration of a hill climbing algorithm or through Taylor Series approximation. Further, since the solution exists for all $\mathbf{P} \in (0,1)_{n\times 1}$, we can also proceed through MCMC in a Bayesian framework. Of particular interest is the conditional estimation of β, given the explanatory variables $\mathbf{x}_i$ such that the nonlinear link constraint (2.26) holds for every observation.

As such, the frequentist estimation may be done using parametric assumptions on the conditional distribution of $f(\alpha^*|\beta)$ for each observation in a joint MLE estimation procedure iteratively or for model checking [2.5] for a particular estimated value of β. Since, solving n such nonlinear constraints can be computationally expensive, I focus on and detail a new latent variable, Bayesian Hierarchical estimation procedure that can overcome this constraint. A frequentist estimation procedure is detailed in the appendix [.1.2] while the Bayesian formulation is given in Section [2.4.2].

2.4.2 Hierarchical Bayesian Estimation Algorithm for Proposed Logistic Regression

Because of the constrained optimization nature of the problem, the Bayesian Hierarchical estimation procedure allows substantial improvements in the correlations between $\mathbf{P}$, α^* and β. The central issue revolves around the fact that to have a full probability model we must specify a distributional assumption for $f(\alpha^*|\beta)$ or $f(\beta|\alpha^*)$. Since the only information at hand is the expected value of $\alpha^*|\beta$, to keep this assumption from being too restrictive, it is reasonable to specify a distribution for which both the mean and variance may be expressed as a function of $\alpha^*|\beta$. As such, consider a latent variable model similar to that given in Albert and Chib 1993, where[8]

$$y_i^* = \lambda(\mathbf{c}(x_i), \beta) + \epsilon_i^*, \tag{2.27}$$

$$y_i^* \overset{i.i.d.}{\sim} Logistic(\lambda(\mathbf{c}(x_i), \beta), \pi^2/3), \tag{2.28}$$

$$y_i = \begin{cases} 1 \ if \ y_i^* > 0 \\ 0 \ if \ y_i^* \leq 0. \end{cases} \tag{2.29}$$

[8]Please note that fixing the variance parameter is according to the formulation given in Albert and Chib in 1993. However, the current formulation may provide further avenues of research to overcome this constraint and is left open to be pursued in future research efforts.

Using an augmented joint posterior distribution, the full probability model can be written as,

$$p(\beta, \alpha^*|y) = \int_{y^*} p(\alpha^*, \beta, y^*|y) \implies p(\beta|\alpha^*, y) \propto L(X, \beta)p(\alpha^*|\beta, y)p(\beta). \tag{2.30}$$

Therefore, a sequential MCMC algorithm can be set up where by integrating out the sampled y^* values we can draw from the conditional distribution of $f(\alpha^*|\beta)$. Then we can draw from a suitable proposal density and get estimates of β, by iterating to completion. In particular, consider

$$F(\epsilon_i^*) \overset{i.i.d.}{\sim} Logistic(0, \pi^2/3), \tag{2.31}$$

$$f(\alpha_i^*|\mathbf{x_i}, \beta) \overset{i.i.d.}{\sim} \theta exp(-\theta g(\mathbf{x_i}, \beta)), \tag{2.32}$$

$$g(\mathbf{x_i}, \beta) = \frac{\lambda(c(\mathbf{x_i}), \beta) + log(1 - F(\lambda(c(\mathbf{x_i}), \beta)))}{log(F(\lambda(c(\mathbf{x_i}), \beta)))}, \tag{2.33}$$

$$f(\beta) \sim N(\mu_0, \sigma_0^2). \tag{2.34}$$

Thus, for suitable values of the hyper-parameters (section 2.6 and section 2.7) and given the existence and uniqueness of the functional specification, we can set up an appropriate MCMC algorithm with a Metropolis Hastings (MH) within Gibbs procedure, as follows.

---

1. Draw from the truncated Logistic distribution for each observation.
2. Given the realized values of $y_i^{*\prime}s$, draw from $f(\alpha_i^*|X, \beta)$, making any transformations as necessary.
3. Perform a MH step to accept the current draws of β's from a suitable proposal distribution, ensuring that the posterior is traversed accordingly to the mode.

4. Iterate to completion.

It is easy to see that though I have applied the methodology to the Logistic latent variable formulation, it can be applied to any GLM with a modification to $g(c(\mathbf{X}), \beta)$. This is guaranteed by the existence and uniqueness results above.

2.5 Asymptotics

One of the more useful outcomes of the proposed model is that it simply adds one extra parameter to be estimated. Furthermore, since we know $\mathrm{E}(\alpha^*|\beta)$ for existing models such as Logit ($\boldsymbol{\alpha}^* = \mathbf{1}$), we can use large-sample results under independence through Assumptions 2.1 and 2.4 to test the hypothesis that our model results vary from traditional GLM fits. In particular, we know for GLM,

$$E[\alpha^*|\beta] = log(P(\mathbf{X}))^{-1}(log((g^{-1}(\mathbf{c}(\mathbf{X})\beta)))). \tag{2.35}$$

While the X's are held fixed, $\bar{\alpha^*}$ is both asymptotically unbiased, consistent and asymptotically normal by the central limit theorem and i.i.d. assumptions. This is an assertion which holds as long as β is consistent and asymptotically unbiased. Given $\bar{\alpha^*}$, we can thus estimate the asymptotically consistent estimates of the variance of α^* as well using these facts and the central limit theorem then we have

$$\alpha^* \sim N(E(\alpha^*|\beta^*), E(\alpha^* - E(\alpha^*|\beta^*)|\beta^*)^2), \tag{2.36}$$

$$\hat{\bar{\alpha}}^* \overset{asymp.}{\sim} N\left(\frac{\sum_{i=1}^{n}\alpha_i}{n}, \frac{\sum_{i=1}^{n}(\alpha_i-\bar{\alpha})^2}{(n-1)}\right). \tag{2.37}$$

β^* above represents the optimized estimated value. Thus, we can check our hypothesis that $\bar{\alpha^*} = k$, for some $k \in \mathbf{R} \setminus \{-\infty, \infty, 0\}$ for any particular GLM as follows.

---

1. Perform a t-test on $\bar{\alpha^*}$, with the appropriate null hypothesis values, and accept/reject model fit assumptions (for example $H_0 : \bar{\alpha^*} = 1$ for the Logit).

2. Thus,

   (a) Under rejection, the existing GLM is not adequate given assumptions on the model specification and the proposed model should be used.

   (b) Otherwise, the existing GLM is adequate and it can be used for inference and prediction (classification) accordingly[9] (taking into account comparative MIP performances of the models considered as needed).

This framework can similarly be extended to the likelihood ratio test, under the appropriate null values. For example, for the Logistic specification the null values are $\mathbf{E}[\alpha^*] = \mathbf{E}[\delta^*] = 1$.

[9]Note however, that model fit, prediction and inference criteria should be evaluated on a wholistic basis to arrive at a choosen model even if the null hypothesis is not rejected.

2.6 Monte Carlo Simulation

In order to validate the robustness of the proposed methodology the Generalized Logistic Link Function is used in the Bayesian framework for extensive simulation studies on various DGP's, both symmetric (Logit and Probit) and asymmetric (Complementary Log-Log). For this purpose, datasets were generated from the standard normal distribution for different sample sizes ($n = \{100, 500, 1000\}$) for three different models,

$$\mathbf{Y} = Intercept + \mathbf{X_1} + (\mathbf{X_2})^2, \tag{2.38}$$

$$\mathbf{Y} = Intercept + \mathbf{X_1} + exp(\mathbf{X_2}), \tag{2.39}$$

$$\mathbf{Y} = Intercept + exp(\mathbf{X_1}) + sin(\mathbf{X_2}). \tag{2.40}$$

The different model specifications are needed to understand the performance of the proposed model when the data are linear, non-linear or a mixed specification in the X's. All datasets had 3 parameters to estimate, for the intercept (β_1) and for two explanatory or independent variables drawn from the standard normal ($\{\beta_2, \beta_3\}$) with the appropriate transformations indicated above. Then for fixed and known β values, either a Probit, Logit or a Complementary Log-Log DGP was used to generate outcomes (dependent variable $\mathbf{Y}$), that varied in the number of 1's that were present[10]. That is, the known β values were used with the known $\mathbf{X}$'s in a regression model to create the dependent variable $\mathbf{Y}$. Furthermore, some additional changes were done to make sure that in-sample and out-of-sample simulated data were comparable in regards to their means.

In particular, the known $\{x, \beta\}$ values along with each functional form above (the Probit, Logit or Complementary Log-Log) can be used to calculate the probability of each obser-

[10]If the probability calculated under a DGP for a particular observation was greater than the median, it is considered to be 1.

vation for each specific model (for example, $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 (x_{i2})^2$), where x_{ij} indicates value of the j^{th} independent variable ($j \in \{1, 2\}$) for the i^{th} row [11,12]. Thus, we can consider the calculated $\mathbf{Y}$ values along with the generated $\mathbf{X}$'s as the data on which we can fit our chosen statistical models for each DGP. We can then evaluate the performance of the proposed model against other popular existing baseline models[13].

Finally, another step was done to create datasets which had different numbers of successes as opposed to failures[14]. Thus, the unbalancedness of the data were varied between $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Here, 0.5 indicates equal number of successes and failures (balanced), 0.4 indicates 10% fewer successes than failures and so forth. This alteration was done for each of the different sample sizes, for each of the three DGPs of Probit, Logit and Complementary Log-Log, as well as for each of the three models specified (linear, non-linear or mixed). Thus, for each sample size there are five different unbalanced datasets, each of which has three parameters or β's to estimate for each of the three DGPs for each of the models specified (linear, non-linear or mixed). As such, for each model, there are 45 different datasets, each with 3 parameters to estimate, for a total of $135 \times 3 = 405$ parameters to estimate, compare and contrast[15].

On these synthetic datasets a simple MLE based Logistic, a Bayesian Latent Probit, the proposed Generalized Logistic model in the Bayesian latent framework, and an MLE based Penalized Logistic model were run. The final comparisons were based on both in-sample and out-of-sample (last 20% of each synthetic dataset) data, confidence intervals of estimated

[11] Naturally, the values achieved from each functional specification of the DGP are necessarily different for each function.

[12] Then we may create a success as those observations for which a particular functional form of the DGP predicted a probability greater than the median.

[13] This construction means that if the data were generated using a Logistic DGP, then when we fit the Logistic model to this synthetic data, its model fit and inference results should be better than the other models fitted to the data.

[14] As iterated above, if the number of successes and failures in the dataset differ, then the data are considered to be unbalanced.

[15] Note also that by construction, we know what the true β's are, and therefore, can use these true values to understand the performance of each of the models fitted to each dataset.

β's compared to actual β's, number of MCMC iterations required and Akaike Infomration Criteria (AIC). Where the AIC is defined as,

$$2 \times (\text{-}log - likelihood + Number\ of\ Parameters)/Number\ of\ Observations \quad (2.41)$$

Evidently, the AIC statistic penalizes those models which are more complex or have more parameters to estimate[16]. Consequently, lower AICs are considered better than higher ones and they can be computed for all models for which a likelihood can be computed. In the simulation study results below, this is an important criterion for determining the model which fits the simulated data the best. Note, however, that for inference, standard errors of each model and the confidence intervals which they give are more important for choosing the best model. Indeed, these are distinctly different tasks and as such requires the consideration of the appropriate statistics to measure their effectiveness separately.

A summary of the results below shows the efficiency, robustness and superior model fits of the proposed methodology, both in-sample and out-of-sample. In almost all circumstances for the Logit DGP, the Probit DGP or even the asymmetric Complementary Log-Log DGP, the proposed model out-performs the existing methodologies with respect to at least one of the comparison criteria AIC, confidence interval or most importantly, the number of times the confidence intervals contained the true β's. There are 405 specific β's to estimate and compare and the summary based on averages are given below in Table 2.1 and Figure 2.1 for all linear, non-linear and mixed models specified.

The MLE Logistic model fits are extremely poor with multiple confidence interval ranges being very large. On the other hand, the proposed model has the lowest average AIC both in-sample and out-of-sample for all DGPs (Figure 2.1). However, most importantly, the proposed model contained the true parameters 84.20% (341 out of 405 total) of the time, as

[16]Thus, it naturally considers the Occam's razor bias in its estimation.

Table 2.1: Simulation Summary of Model Fits for All DGPs

	Bayesian Latent Probit	MLE Logistic	Penalized Logistic	Proposed Logistic
In-Samp. AIC	1.39	1.37	3.27	1.17
Out-of-Samp AIC	1.57	1.60	3.62	1.27
$\#\beta_1$ in C.I. (max. 135)	99	33	12	97
$\#\beta_2$ in C.I. (max. 135)	79	35	75	116
$\#\beta_3$ in C.I. (max. 135)	109	35	71	128
β_1 C.I. Rng.	3.07	1,849.95	7.65	4.76
β_2 C.I. Rng.	1.96	279.44	4.54	3.95
β_3 C.I. Rng.	2.33	9,174.67	12.14	4.54

Note: This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 15 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 45 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 135 parameters per DGP for a total of 405 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets. The AIC of the proposed method were on average 21.31% better (1.22 vs. 1.48) for in-sample and out-of-sample datasets combined, in comparison to the next best model in terms of AIC, the LV existing Bayesian Probit model. The confidence intervals (C.I.'s) of the proposed model were far more reasonable with a range of about 4.42, as opposed to a range of only 2.45 for the Bayesian Probit model. As such, the proposed model had 18.82% more of the true parameters than the Bayesian Probit and almost 331.07% more of the true parameters in its C.I.'s than the MLE Logistic (which is widely recognized to be the baseline model for binary outcomes). That this was attained in only 8,000 MCMC iterations with a 4,000 burn-in period is even more poignant in regards to the efficiency and robustness of the proposed methodology (as opposed to 12,000 iteration and 6,000 burn-in period for the existing Bayesian Probit).

opposed to 70.86% (287 out of 405) and 39.01% (158 out of 405) of the time for the Bayesian Probit and Penalized Logistic models, respectively. In fact, this level of coverage is attained using a smaller confidence interval than the Penalized Logistic, which contained the third highest number of true β's in its confidence intervals. The proposed model, in comparison to the Bayesian Probit, had on average 18.82% (341 vs. 287) more of the true parameters. In comparison to the Penalized Logistic, the proposed model had on average 54.50% (4.42 vs. 8.11) smaller confidence intervals, while containing 215.82% (341 vs. 158) more of the true parameters. In comparing to the MLE Logistic, the proposed model had on average 331.07% (341 vs. 103) more of the true parameters, even if we ignore the unsupportable confidence intervals for the MLE Logistic due to several extremely poor fits.

Clearly the proposed model has a significant advantage over the existing models compared. However, the analysis also highlights that it is possible to have a low AIC, as the MLE

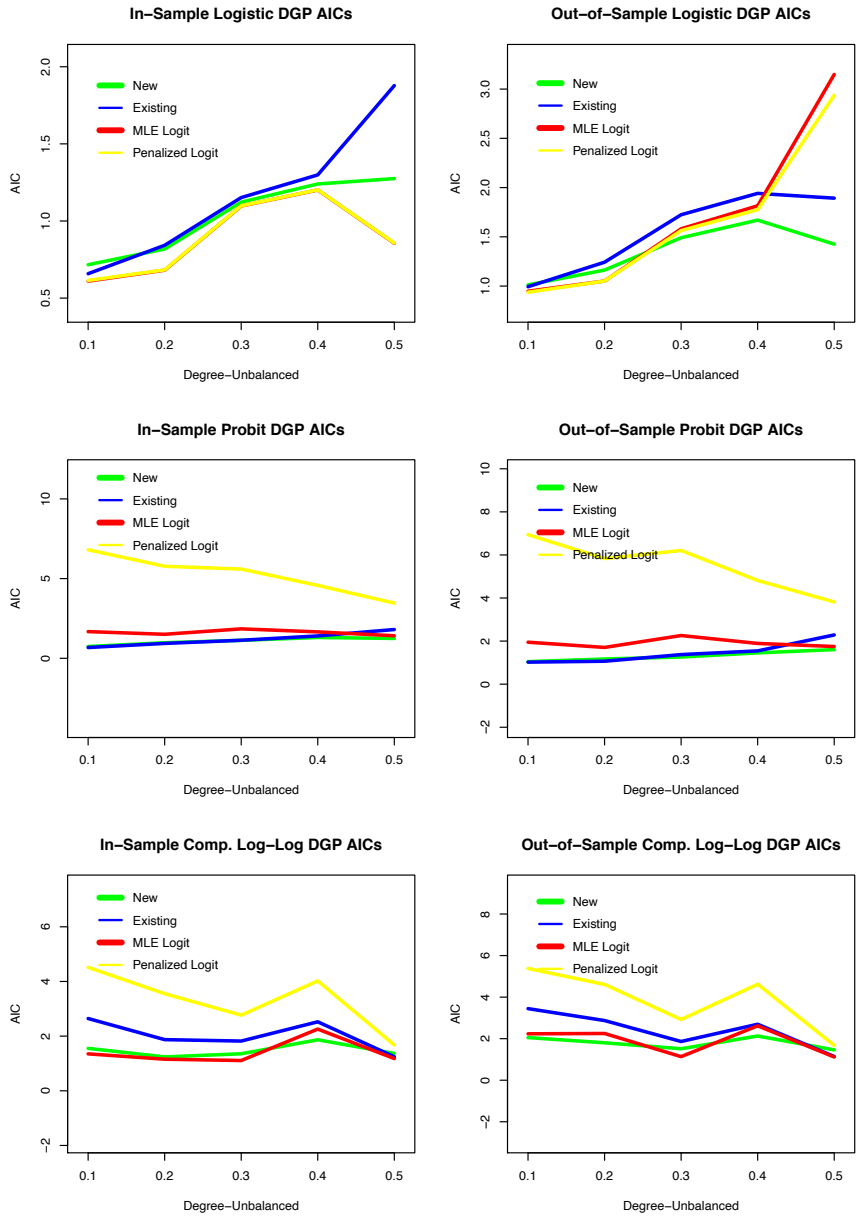
Logistic does for both in-sample and out-of-sample data over the Penalized Logistic, yet give a confidence interval which does not contain the true parameter. One reason for this discrepancy which would impact inference most-of-all, could be that many existing models overfit the data. The proposed model suffers less from this issue. Consequently, not only does it contain the true parameters more often, it also has uniformly better average AICs in both in-sample and out-of-sample data. Additionally, this performance level is attained with far fewer iterations needed than the existing latent Bayesian Probit model for parameter convergence. A more detailed breakdown along DGP, observation and unbalancedness is available upon request.

2.7 Empirical Application

In order to apply the theoretical constructs above, I apply the Logistic formulation to the data from Hu et al. (2020) to understand the importance of author-defined keywords for articles to be highly-cited in the Management Information Systems (MIS) field. In particular, I apply it to those articles which they identified as being in the top 25^{th} percentile of citation counts for all articles considered in the MIS field. This is done for 6 separate years for two different training dataset sizes. The first of which used 80% of the observations available for each year while the latter used only 25% of the total data available for training purposes. Thus, there are 12 specific datasets to compare and contrast. For the MIS field "three top influential" journals were considered for identifying highly cited papers, Information Systems Research (ISR), MIS Quarterly (MISQ) and Journal of Management Information Systems (JMIS) for all papers published between 2009 to 2012. Below I give a summary of how the data were created.

The preprocessing of the texts occurred based on the title, abstract and keywords from Web of Science (WOS), creating an "article-term matrix" through Latent Dirichlet Allocation

Figure 2.1: Simulation Results Summary For All Three DGPs In-Sample and Out-of-Sample For All Three Linear and Non-Linear Models Considered.



Note: The AICs for MLE-Logit and Penalized Logit were ∞ for multiple datasets and thus only finite AIC values are graphed here. Comp. Log-Log refers to Complementary Log-Log, Logistic to Logistic and Probit to the Probit Data Generating Processes (DGPs). New: Proposed Methodology, Existing: Bayesian Probit, MLE Logit: MLE Logistic Regression, Penalized Logit: Penalized Logistic Regression. The results are summarized over all observations and unbalanced datasets created, graphed in order of decreasing unbalancedness (unless ∞) for in-sample and out-of-sample datasets. The results are presented as a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} summarized over the linear, non-linear and mixed models specified. Thus, there are 135 different datasets to consider with a total of 405 parameters estimated over the entire simulation study. The results are summarized by average over all simulated datasets according to the amount of unbalancedness in the datasets. Where 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth. While in the in-sample datasets the proposed model in the Hierarchical Bayesian Logistic application had AICs very close to the Penalized Logistic regression, the out-of-sample AICs for the proposed methodology were almost uniformly better than the existing methods compared.

(LDA) on the tokenized texts, to obtain the best keyword candidates[17]. A standard Parts-of-Speech tagger was applied to identify nouns (NN), proper nouns (NP) and adjectives (JJ) to convert each article into a vector listing of these parts-of-speeches[18]. This was followed by further dimensionality reduction procedures according to Phan and Nguyen 2008. Thus, the CDFs of ϕ the keywords within a topic and θ the topic within each paper were generated to yield the article-keyword matrix for each paper, for each of the six years, from one year after publication to six years after publication for every article in the data. Finally, with the use of web crawlers and Application Programming Interface (APIs) for ReseachGate, Google Scholar and Google Trends, each keyword was searched and the popularity measures were calculated.

The binary dependent variables for each article, for each year considered were classified to either fall within the top 25^{th} percentile of total citation counts (a success or 1) or not (a failure or 0) for the year under consideration[19]. While the original study considered journal, author and several keyword features, the efficacy of the proposed model meant that in the current application only journal impact factor (JIF) and one keyword feature (PP) needed to be considered (according to the best model fit outcomes), while still being consistent with the original results of the Hu et al. 2020 paper. Thus, for journal features, journal impact factor (JIF) was the main attribute considered (based on existing well established results in the field; see for example Bai et al. 2019 and Wang et al. 2019). For the keyword parameters, five specific measures or variables were considered namely, topic popularity (TP), published popularity (PP), news popularity (NP), web page popularity (WPP) and video popularity (VP). Below I elaborate on their computations in greater detail.

Let $j \in \{2009, 2010, 2011, 2012, 2013\}$ be the year of publication of an article, let M be the

[17]The keyword candidates themselves were retrieved from various search engines and I elaborate more on this shortly.

[18]These parts-of-speech are considered more indicative of the academic publishing content in the field.

[19]Thus, the analyses done here is a cross-sectional analyses for the years considered for the MIS field for an article since its initial publication year.

number of topics and let N be the number of keywords in each topic obtained from the LDA analysis mentioned above. Define $k_{m,n}$ $\{m \in M,\ n \in N\}$ as the n^{th} keyword for the m^{th} topic. Thus, from ResearchGate for each $k_{m,n}$ we can obtain the number of questions ($q_{m,n}$) related to it. From Google Scholar we can obtain the number of search results (sorted by year) related to each $k_{m,n}$ (which I further denote as $p^j_{m,n}$ here). Similarly, from Google Trends for each $k_{m,n}$ (specifically using Google News, Google Web Pages and YouTube), we can also obtain the counts for news popularity ($e^j_{m,n}$), web page popularity ($w^j_{m,n}$) and video popularity ($v^j_{m,n}$) respectively for each article. Finally, we can define each article in a year j, by the index $i \in \{1, ..., I\}$, with i defined as the total number of articles in the j^{th} year. Accordingly, for the j^{th} year the measures can be defined as follows:

$$TP^i_j = \sum_{m=1}^{M}\sum_{n=1}^{N} q_{m,n}\theta^i_m\phi^m_n \tag{2.42}$$

$$PP^i_j = \sum_{m=1}^{M}\sum_{n=1}^{N} p^j_{m,n}\theta^i_m\phi^m_n \tag{2.43}$$

$$NP^i_j = \sum_{m=1}^{M}\sum_{n=1}^{N} e^j_{m,n}\theta^i_m\phi^m_n \tag{2.44}$$

$$WPP^i_j = \sum_{m=1}^{M}\sum_{n=1}^{N} w^j_{m,n}\theta^i_m\phi^m_n \tag{2.45}$$

$$VP^i_j = \sum_{m=1}^{M}\sum_{n=1}^{N} v^j_{m,n}\theta^i_m\phi^m_n. \tag{2.46}$$

Therefore, TP is a weighted mean of the numbers of questions from ResearchGate matching article keywords, PP is a weighted mean of the number of search results of the article keywords from Google Scholar, NP is a weighted mean of the degree of news popularity of article keywords from Google Trends, WP is a weighted mean of the degree of web page popularity of article keywords from Google Trends and VP is a weighted mean of the degree of video popularity of article keywords from Google Trends all with corresponding probabilities.

Therefore, the goal is to understand how well JIF and the keyword parameters predict which article for any given year will be highly-cited[20]. Accordingly, I use the same AIC statistic as in the simulation (2.41) to compare model fits across the various models considered. Separately, to better evaluate prediction or classification performance, there are many accepted statistics based on the confusion matrix (which is recreated below for convenience). However, for the current application I use the ROC-Statistic given in Chowdhury 2019.

Table 2.2: Confusion Matrix.

		Fitted Model Prediction	
		Highly-Cited	Not Highly-Cited
True Classification in Data	Highly-Cited	True Positive (TP)	False Negative (FN)
	Not Highly-Cited	False Positive (FP)	True Negative (TN)

The statistic is defined as follows,

$$ROC - Statistic = \frac{FP}{TP}, \tag{2.47}$$

which spans between $[0, \infty)$, with a lower number indicating better prediction results. As such, please note that in any model a lower number for AIC and ROC-Statistic indicates a better model fit or prediction results respectively. They can further be computed for training and test datasets separately to see how well the in-sample results compare to out-of-sample results. This is done, because we would like to recreate the performance of in-sample MIP's performances to that from other samples from the true populaiton DGP (represented by the out-of-sample hold-out data). Therefore, the underlying assumptions is that MIP's performances based on true population parameters should be more robust and give better models fits out-of-sample, in addition to having reasonable MIP's performances in-sample.

Thus, in what follows I apply the proposed methodology to this dataset along with the Bayesian Probit, MLE Logistic, Penalized Logistic and ANN to understand their model fit,

[20]Where again an article is considered highly cited if it is in the top 25^{th} percentile of citation counts for all publications in a particular year and not highly cited otherwise.

inference and prediction performances based on these criteria. In doing so, I show that the proposed methodology can be used for both prediction and inference without overfitting the data, a known weakness of many AI and ML methods such as ANN or SVM. I further show that unlike in Hu et al. 2020, who find that the best model fits are attained by combining several keyword parameter variables with other Journal or Author variables a similar prediction results can be attained by not including correlated variables in the model specification. This implies that for the MIS dataset we need not sacrifice between the prediction, inference or model fit criteria because the proposed model finds the requisite balance between them to give generalized results that can outperform widely used AI models such as the ANN.

2.7.1 Classification of Highly Cited Papers

Given the highly correlated nature of the various explanatory variables considered, the final analysis on the datasets consisted of the following model,

$$Highly\ Cited = Intercept + Journal\ Impact\ Factor + Popularity\ Parameter. \tag{2.48}$$

A brief version of the algorithm is given below.

---

1. Draw from the truncated Logistic distribution for each observation.
2. Given the realized values of $y_i's$ draw from $f(\alpha_i^*|X, \beta)$, performing any transformation as necessary.
3. Perform an MH step with the t-distribution as the proposal, with 10 degrees of freedom (WLOG).
4. Iterate to completion, for total draws of 8,000 with 4,000 burn-in samples.

This was done for both a diffuse prior (normal prior with 0 mean and a variance of 10) and a more informative prior where the β's were considered to have positive normal prior mean of 0.5 and a variance of 10. In addition, the MLE Logistic, Bayesian Latent Probit, Penalized Logistic and ANN models were also run to compare the robustness of the proposed procedure. In total the Bayesian Probit was run for 12,000 maximum iterations with 6,000 burn-in period, in contrast, the proposed model was run for only two-thirds the number of iterations with 4,000 burn-in period.

To showcase the flexibility of the model when the dataset is small and unbalanced, the above mentioned models were fitted to two separate datasets. The first dataset contained 80% of the total available observations per year for training the models, and the latter was a smaller training dataset containing only 25% of the total observations available. A summary of the ROC-Statistics can be found in Table 2.3 and the AIC based summary is given in Table 2.4. The estimates with the relevant standard errors can be found in Table 2.5 and Table 2.6.

In the 80% in-sample dataset, the proposed model despite having a higher AIC than the other models, beat their classification performances uniformly out-of-sample. However, the out-of-sample classification performance of all the models were very close in this application. Therefore, given how close the proposed model's ROC-Statistic is to that of ANN and the Penalized Logistic, the performances for it can be considered to be essentially equal to them (or at best slightly better) for this dataset. The results also showcase the need to treat model fit (AIC) and classification (ROC-Statistic) separately in applications. For example, in the 80% in-sample dataset the average AIC of the proposed model under both prior specifications were higher than the other models considered, yet it gave classification results superior to the aforementioned models.

In comparison, in the 25% in-sample data, though the proposed model had the highest average AIC for the training dataset, it beat the other models in the test dataset on average by nearly 44.90%. In other words, in regards to model fit, when the dataset is small

Table 2.3: ROC-Statistic for Management Information System.

Management Information Systems (25%)							Management Information Systems (80%)					
Dataset	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
IS-1	0.51	0.51	0.46	0.20	0.46	0.46	0.25	0.23	0.34	0.01	0.13	0.23
IS-2	0.44	0.41	0.44	0.07	0.44	0.41	0.15	0.17	0.48	0.08	0.06	0.17
IS-3	0.49	0.49	0.49	0.33	0.49	0.49	0.26	0.15	0.30	0.00	0.08	0.16
IS-4	0.48	0.48	0.48	0.04	0.48	0.43	0.24	0.20	0.29	0.04	0.08	0.19
IS-5	0.58	0.52	0.47	0.17	0.52	0.52	0.20	0.20	0.30	0.01	0.04	0.18
IS-6	0.49	0.49	0.44	0.09	0.44	0.49	0.14	0.28	0.33	0.08	0.06	0.18
Mean (IS)	0.50	0.48	0.46	0.15	0.47	0.47	0.21	0.20	0.34	0.04	0.08	0.18
OS-1	0.23	0.19	0.53	0.01	0.62	0.47	0.09	0.09	0.09	0.05	0.17	0.09
OS-2	0.18	0.15	0.35	0.11	0.33	0.24	0.11	0.35	0.35	0.18	0.35	0.35
OS-3	0.11	0.12	0.17	0.14	0.17	0.15	0.00	0.00	0.09	0.16	0.09	0.00
OS-4	0.08	0.09	0.11	0.14	0.11	0.10	0.19	0.19	0.10	0.05	0.19	0.19
OS-5	0.10	0.11	0.21	0.83	0.2	0.17	0.19	0.00	0.33	0.46	0.19	0.19
OS-6	0.05	0.05	0.09	0.01	0.09	0.08	0.10	0.05	0.24	0.05	0.20	0.00
Mean (OS)	0.12	0.12	0.24	0.21	0.25	0.20	0.11	0.11	0.20	0.16	0.20	0.14

Note: (1): Informative Prior N(0.5, 10); (2) Diffuse Prior N(0, 10); (3): Bayesian Probit (Inform. Prior); (4): Artificial Neural Network (ANN); (5): MLE Logistic; (6): Penalized Logistic; (IS): In-Sample; (OS): Out-of-Sample. IS-1 (OS-1) to IS-6 (OS-6): 1 year after publication to 6 years after publication, with IS indicating in-sample and OS indicating out-of-sample. 25%(80%) implies 25%(80%) of each dataset was kept as in-sample or training data.

Table 2.4: AIC for Management Information Systems for Varying Training Data Size.

Management Information Systems (25%)							Management Information Systems (80%)					
Dataset	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
IS-1	2.20	2.20	1.15	1.09	1.15	1.17	1.64	1.64	1.17	0.91	1.14	1.16
IS-2	2.19	2.19	1.12	0.89	1.12	1.14	1.42	1.56	1.14	0.90	1.13	1.15
IS-3	1.58	1.58	1.1	1.01	1.08	1.10	1.31	1.40	1.10	0.86	1.05	1.07
IS-4	1.61	1.61	1.09	0.8	1.15	1.16	1.28	1.37	1.64	0.88	1.04	1.07
IS-5	2.09	2.09	1.11	0.84	1.19	1.20	1.47	1.47	1.41	0.93	1.08	1.11
IS-6	2.09	2.09	1.12	0.71	1.20	1.21	1.39	1.39	1.50	0.93	1.06	1.09
Mean (IS)	1.96	1.96	1.11	0.89	1.15	1.16	1.42	1.47	1.33	0.90	1.09	1.11
OS-1	1.74	1.74	3.86	1.14	2.97	2.02	0.96	0.96	0.97	1.41	0.71	0.81
OS-2	1.59	1.59	3.16	1.03	3.16	2.09	1.06	1.02	0.86	1.12	0.80	0.90
OS-3	1.37	1.37	1.90	1.55	2.07	1.78	1.26	1.20	0.86	1.08	0.77	0.86
OS-4	1.25	1.25	1.86	1.29	1.75	1.61	1.00	0.92	1.15	1.10	0.70	0.80
OS-5	1.62	1.62	1.68	7.04	1.88	1.66	0.93	0.93	1.13	0.99	0.70	0.81
OS-6	1.53	1.53	2.34	1.39	1.97	1.72	0.95	0.95	1.06	0.86	0.72	0.85
Mean (OS)	1.52	1.52	2.46	2.24	2.30	1.81	1.03	1.00	1.01	1.09	0.73	0.84

Note: (1): Informative Prior N(0.5, 10); (2) Diffuse Prior N(0, 10); (3): Bayesian Probit (Inform. Prior); (4): Artificial Neural Network (ANN); (5): MLE Logistic; (6): Penalized Logistic; (IS): In-Sample; (OS): Out-of-Sample. IS-1 (OS-1) to IS-6 (OS-6): 1 year after publication to 6 years after publication, with IS indicating in-sample and OS indicating out-of-sample. 25%(80%) implies 25%(80%) of each dataset was kept as in-sample or training data.

or more unbalanced, in real-world applications the proposed model out-of-sample handily outperformed all the other models considered including Neural Networks. On average in out-of-sample data, the proposed model for both specifications had the lowest AIC (1.52 vs. 2.24 for ANN and 2.46 for the Bayesian Probit), and, therefore, it was the best model in regards to model fit in the more truncated dataset (as hypothesized). In addition, for classification in this dataset, the proposed model outperformed all other models out-of sample by 91.67% (0.12 vs. 0.23).

Furthermore, the inference results and the significance outcomes tell a related yet separate story. While for the 80% training dataset all explanatory variables (other than the Intercept) are significant at the 0.05 $\alpha-$level, the proposed model uniformly found both the JIF and PP metrics to be relevant to citation outcomes for all years under consideration. This shows the versatility of the proposed model since not only does it have uniformly better out-of-sample prediction results, but it also finds that for the MIS field, both JIF and PP parameters may be more important than in other fields of the social sciences. However, in the 25% in-sample data application, the JIF parameter is never found to be significant, yet the PP metric is always significant and positive. This finding is consistent with Choi et al. 2011, who find that as the MIS field is more interdisciplinary, there can be rapid changes in its domain over other fields. As such, it stands to reason that keyword popularity measures such as PP would be at least as important for such fields as JIF. In addition, the number of years after publication is also a critical factor in predicting highly-cited papers in the MIS field. However, unlike in Hu et al. 2020 who find that the fifth and sixth year dataset prediction performances were better using Journal, Author and Keyword parameters, I find that the importance of PP and JIF can extend from some where between two to three years after publication onwards. The conclusion follows from the proposed model's prediction results being perfect for identifying highly-cited papers out-of-sample three years after publication, compared to the prediction performance two years after publication in the 80% in-sample dataset. In the truncated dataset I also find that prediction performance in the second year

was almost 78.26% better than the first year. Thus, it seems reasonable to conclude that the PP parameter can be important for the MIS field somewhere between two to three years after publication. This seems a reasonable finding since papers based on new ideas can take two or three years from idea inception, working draft creation, to finally publication after going through a thorough review process in high JIF journals.

Table 2.5: Summary of Model Fits - Management Information Systems (MIS) for All Years for 80% Training Dataset.

Variable	Proposed Model (Diffuse Prior) Estimates	Proposed Model (Subjective Prior) Estimates	Penalized Logistic Estimates	MLE Estimates	Bayesian Latent Probit Estimates
			MIS-Year 1		
Intercept	-0.08 (0.10)	0.04 (0.11)	-0.13 (0.16)	-0.46** (0.18)	0.17*** (0.07)
JIF	1.36*** (0.10)	1.76*** (0.12)	1.03*** (0.17)	1.17*** (0.19)	0.55*** (0.06)
PP	0.41*** (0.11)	1.04*** (0.13)	0.36** (0.17)	0.39** (0.18)	0.16*** (0.05)
Alpha	1.17 (1.31)	1.59 (1.8)	1.00	1.00	NA
			MIS-Year 2		
Intercept	-0.14 (0.11)	-0.06 (0.13)	-0.06 (0.17)	-0.38** (0.18)	0.34 (0.06)
JIF	1.63 *** (0.15)	0.92*** (0.14)	1.00 *** (0.18)	1.14 *** (0.20)	0.59*** (0.05)
PP	0.46*** (0.11)	0.74*** (0.11)	0.53*** (0.19)	0.56*** (0.19)	0.19*** (0.05)
Alpha	1.12 (1.61)	1.24 (1.47)	1.00	1.00	NA
			MIS-Year 3		
Intercept	-0.11 (0.13)	-0.42 (0.11)	-0.10 (0.18)	-0.49** (0.19)	0.24 (0.07)
JIF	1.43*** (0.15)	1.61*** (0.11)	1.25*** (0.19)	1.40*** (0.21)	0.60*** (0.06)
PP	0.56*** (0.10)	0.57*** (0.11)	0.46** (0.19)	0.50** (0.2)	0.09*** (0.05)
Alpha	1.04 (1.9)	1.91 (2.36)	1.00	1.00	NA
			MIS-Year 4		
Intercept	0.01 (0.1)	-0.01 (0.11)	0 (0.18)	-0.41** (0.2)	0.14*** (0.07)
JIF	1.66*** (0.13)	1.59*** (0.12)	1.24*** (0.2)	1.42*** (0.22)	0.53*** (0.05)
PP	0.76*** (0.12)	1.06*** (0.10)	0.52** (0.2)	0.55*** (0.21)	0.16*** (0.04)
Alpha	1.05 (1.71)	1.63 (1.81)	1.00	1.00	NA
			MIS-Year 5		
Intercept	0.06 (0.11)	0.03** (0.08)	-0.07 (0.18)	-0.44** (0.20)	0.02*** (0.07)
JIF	1.72*** (0.12)	1.77*** (0.10)	1.12*** (0.19)	1.29*** (0.22)	0.56*** (0.06)
PP	0.66*** (0.11)	0.66*** (0.13)	0.50** (0.20)	0.53*** (0.20)	0.17*** (0.05)
Alpha	1.41 (1.48)	1.2 (1.4)	1.00	1.00	NA
Intercept	0.33** (0.11)	0.03 (0.10)	0.04 (0.18)	-0.34* (0.20)	0.31*** (0.06)
JIF	1.78*** (0.12)	1.60*** (0.12)	1.14*** (0.19)	1.33*** (0.22)	0.53*** (0.05)
PP	0.72*** (0.10)	0.53*** (0.08)	0.57*** (0.20)	0.61*** (0.21)	0.13*** (0.05)
Alpha	1.41 (2.31)	0.97 (1.18)	1.00	1.00	NA

Note: *** indicates significance at $\alpha = 0.01$, ** indicates significance at $\alpha = 0.05$ and * indicates significance at $\alpha = 0.10$. Please note $\alpha \neq \alpha^*_{n\times 1}$. α is the significance criteria.

Table 2.6: Summary of Model Fits - Management Information Systems (MIS) for All Years for 25% Training Dataset.

Variable	Proposed Model (Diffuse Prior) Estimates	Proposed Model (Subjective Prior) Estimates	Penalized Logistic Estimates	MLE Estimates	Bayesian Latent Probit Estimates
			MIS-Year 1		
Intercept	-2.43 (0.16)	-2.86 (0.15)	-1.12*** (0.31)	-1.50*** (0.43)	-0.47 (0.06)
JIF	-0.34 (0.14)	-0.41 (0.11)	-0.62* (0.33)	-1.39** (0.7)	-0.44 (0.07)
PP	0.67*** (0.12)	0.52*** (0.13)	0.38 (0.31)	0.54 (0.42)	0.08*** (0.06)
Alpha	4.56** (2.62)	4.47* (3)	1.00	1.00	NA
			MIS-Year 2		
Intercept	-1.67 (0.09)	-2.17 (0.14)	-1.16*** (0.34)	-1.53*** (0.45)	-0.46 (0.07)
JIF	-0.76 (0.14)	-0.27 (0.17)	-0.70** (0.35)	-1.59** (0.76)	-0.44 (0.07)
PP	1.12*** (0.16)	0.94*** (0.14)	0.59* (0.33)	0.83* (0.45)	0.20*** (0.05)
Alpha	3.17*** (1.19)	3.51*** (1.66)	1.00	1.00	NA
			MIS-Year 3		
Intercept	-2.51 (0.10)	-1.92 (0.11)	-1.26*** (0.33)	-1.27*** (0.41)	-0.56 (0.06)
JIF	0.50 (0.20)	0.29 (0.17)	-0.23 (0.32)	-0.55 (0.69)	-0.17 (0.07)
PP	0.96*** (0.17)	0.89*** (0.16)	0.71** (0.33)	1.04** (0.46)	0.20*** (0.05)
Alpha	4.19* (2.72)	3.64** (2.12)	1.00	1.00	NA
			MIS-Year 4		
Intercept	-1.78 (0.16)	-1.63 (0.14)	-1.07*** (0.33)	-1.05*** (0.40)	-0.40 (0.08)
JIF	0.03 (0.21)	0.60 (0.17)	-0.13 (0.32)	-0.30 (0.68)	-0.08 (0.07)
PP	0.96*** (0.15)	1.22*** (0.12)	0.75** (0.33)	1.01** (0.42)	0.21*** (0.05)
Alpha	3.94*** (1.65)	3.34** (1.82)	1.00	1.00	NA
			MIS-Year 5		
Intercept	-1.78 (0.15)	-1.73 (0.14)	-1.08*** (0.32)	-1.12*** (0.38)	-0.49 (0.05)
JIF	-0.08 (0.2)	-0.06 (0.18)	-0.21 (0.32)	-0.47 (0.67)	-0.29 (0.08)
PP	0.81*** (0.13)	0.77*** (0.14)	0.40 (0.32)	0.60 (0.46)	0.12*** (0.06)
Alpha	4.03* (2.52)	4.07*** (1.87)	1.00	1.00	NA
			MIS- Year 6		
Intercept	-1.18 (0.12)	-1.78 (0.19)	-1.06*** (0.32)	-1.11*** (0.38)	-0.56 (0.06)
JIF	-0.33 (0.17)	-0.01 (0.17)	-0.23 (0.32)	-0.53 (0.68)	-0.23 (0.07)
PP	0.54** (0.17)	0.63*** (0.14)	0.42 (0.32)	0.63 (0.45)	0.18*** (0.05)
Alpha	3.24*** (1.59)	3.89* (2.45)	1.00	1.00	NA

Note: *** indicates significance at $\alpha = 0.01$, ** indicates significance at $\alpha = 0.05$ and * indicates significance at $\alpha = 0.10$. Please note $\alpha \neq \alpha^*_{n\times 1}$. α is the significance criteria.

2.8 Discussion

The simulation results are indicative of the efficacy of the methodology even when the assumptions of the GLM model specifications are not violated, in the presence of unbalanced data. However, the most noteworthy result was that the model contained nearly 84.20% of the true parameters while having the lowest AIC's among all the models. Consequently, the proposed model does not overfit the data, while maintaining accuracy and numerical consistency, even when the sample size is small, and does so with far shorter confidence intervals than widely used existing methodologies compared. That this level of performance was attained using only two-thirds the number of iterations of the Bayesian Probit is further testament to its applicability to a multitude of scientific contexts.

The results for different training data size applications to identify highly-cited papers is also informative. In regards to the MIS dataset which kept 80%[21] of all data as training, the results among the methodologies are largely consistent, with both JIF and PP parameters being significant. The proposed methodology also finds both the JIF and PP parameters to be significant for every year. This indicates that in general for the MIS field both PP and JIF are very good predictors of whether a paper will be highly-cited. In addition, the use of the methodology yields interesting results in terms of when Keyword popularity measures, such as PP, are more predictive of highly-cited papers. In the 80% in-sample training dataset for the diffuse prior application, the prediction results are worse for articles considered after only one or two years of publication. However, from the third to sixth years, the ROC-Statistics are on average 366.67% better than the first two years. This clearly indicates that both JIF and PP are more significant in predicting which articles are more likely to be highly-cited somewhere between two to three years after publication and beyond, a result which is novel to the field.

[21] Year-1: 251 observations, Year 2: 233 observations, Year 3: 240 observations, Year 4: 224 observations, Year 5: 223 observations, Year 6; 219 observations.

The efficacy of the proposed model is evident here, since it outperforms all models, even more widely used AI and ML applications such as ANN, in multiple in-sample and out-of-sample datasets in regards to model fit (AIC), as well as prediction (ROC-Statistic). Since it is generally recognized that such AI and ML methods give better model fit and prediction results, and only on rare occasions would they be outperformed by more traditional methodologies, it is one of the more interesting findings to consider here. As such, the results show that this assumption does not necessarily have to hold given a particular model specification considered and may indeed be the opposite! Therefore, the proposed methodology gives the empirical researcher a better baseline against which the performance of AI and ML methods can be compared and contrasted and improved upon as needed.

One of the more important applications of the proposed methodology is when the data size is small or the data is unbalanced. For the truncated data, the proposed model outperforms all models on average, out-of-sample in regards to prediction/classification (Table 2.3) and model fit (Table 2.4) including ANN. The out-of-sample AICs are 61.84% better than the Bayesian Probit (1.52 vs. 2.46) and 47.37% better than ANN (1.52 vs. 2.24), with roughly two-thirds the number of iterations needed for convergence as the Bayesian Probit. This illustrates the versatility and robustness of the methodology as it outperforms existing widely used methods including the ANN, on both model fit (AIC) and prediction (ROC-Statistic) for the test dataset. Further, this performance is attained while maintaining interpretability of its parameters while needing fewer iterations than the existing Bayesian Probit. In addition, the methodology is demonstrated to have robust confidence intervals which contain the true parameters more often in the simulation study. These results were attained with very general assumptions on the hyperparameters and can likely be further improved as needed by considering different specifications in the estimation process.

In fact, the proposed methodology has definite advantages over all the methods compared and contrasted in regards to inference. This conclusion, in the case of the Bayesian Probit,

MLE Logistic and Penalized Logistic is apparent from the simulation results. However, they are reinforced by the empirical application as the standard errors of the parameters of the proposed method are not as large as the MLE methods or indeed as small as the Bayesian Probit. As such, it produces more realistic confidence intervals than these models in the empirical application, which is more likely to contain the unknown true parameter[22]. Thus, overall the proposed model outperforms all methods in regards to both model fit and classification in the smaller training dataset example out-of-sample. It further matches (or slightly improves) the best performing models for classification in the larger training dataset application and has more robust confidence intervals as demonstrated in the simulation studies. As such, its usefulness becomes even more apparent when the data are unbalanced or smaller as theorized.

However, it should be noted that the goal of the methodology proposed is not to replace existing AI and ML methods, but rather to guide their application in a more focused way for better model fit and prediction. As an example, consider (as is the norm in empirical applications) the MLE Logistic as the baseline model used to compare against other AI methods such as ANN. If we relied on this and not consider the proposed methodology as a baseline, we may stop our analysis as being adequate in a cursory application of an ANN model. Yet the ANN application can be improved by considering other specifications of hidden layers (here a maximum of two hidden layers with two neurons per layer were considered), for better model fits and prediction. This is a task which in general has infinitely many specifications of the ANN model to consider. Yet, by using the proposed methodology, we can improve the baseline against which these AI and ML models can be compared to further improve statistical and scientific conclusions beyond that possible by only considering the MLE Logistic regression as the baseline. Consequently, the proposed model is a better baseline model for comparison and should therefore lead to better scientific conclusions regarding questions of

[22]In comparison to ANN, all the other methods, including the proposed methodology, have better interpretability of model parameters, and ANN, therefore, is naturally less useful for inference.

importance to Informetrics, Informatics or the sciences in general. Furthermore, even if after many specifications of existing AI and ML methods, they outperform the proposed model in regards to model fit and/or classification, it does not suffer from a lack of interpretability of the parameter estimates. This is especially important since its confidence intervals are more robust than existing non-AI or ML methods. As such the researcher may decide to use it even if it is outperformed in regards to AIC (model fit) or ROC-Statistic (classification or prediction) for inference purposes. Accordingly, it finds a balance between AI, ML and non-AI or non-ML methods to provide a valuable tool for the analyst in addition to being a better baseline model for comparison.

One of the more useful results of the model is the ability to compare the α^* values to benchmark DGPs such as the Logistic. In the truncated MIS dataset, the large sample test on α^* rejected the dataset coming from a Logistic DGP. Therefore, it is one of the reasons why both the AIC and ROC-Statistic were lower (and therefore better) for the proposed method in that application in comparision to the MLE Logistic methods. Yet in the 80% in-sample MIS data, α^* failed to reject the DGP being Logistic. Therefore, the excellent classification results out-of-sample for this dataset for both the MLE and Penalized Logistic are entirely complementary and consistent with the proposed methodology. Please note that in the Hu et. al. 2020 paper, the Logistic regression gave the best classification performance for this dataset. Accordingly, the findings of the proposed method and the application results here are entirely consistent with existing findings in the literature. This then provides further validation of the proposed methodology as being complementary to existing non-AI and non-ML methods widely used in the sciences.

In addition, given the advantage of the model in regards to both inference and classification in out-of-sample data and model fits for certain datasets, it leaves little doubt that the methodology outperforms the other methods in the empirical applications overall here (while giving similar results to existing methods if the data support them). In particular, the results

highlight the distinction between model fit (AIC), inference (p-value/significance/confidence intervals/scientific significance) and classification/prediction that should be carefully considered by every scientist using statistical methods. It is particularly important to recognize that it is possible for a model to outperform another on any one of these criteria while underperforming in the other(s). This can be through (among others) overfitting or having outliers in the data. For example, a particular model can have a lower AIC, in-sample or out-of-sample, yet have poor classification and/or inferential results, when the estimates and standard errors of the model are used for scientific interpretation or inference.

Therefore, in application the proposed model finds the appropriate balance between these model evaluation criteria without overfitting the data. Consequently, in regards to scientific applicability it seems to have demonstrable advantages over existing widely used methods compared here that can further guide AI and ML applications. Evidently, the proposed model's ability to identify such differences, under varying realizations of the underlying stochastic processes, under minimal assumptions, makes it ideal to inform decision making and answer scientific questions. As such, if we were to consider the multinomial or non-parametric extensions of this baseline construct, then it also, through better all around model fit, inference and classification, should be able to add to the current scientific framework. However, these results are left open to be pursued in future efforts.

In fact the α^* values and the accompanying test can be even more informative to the researcher for AI and ML applications such as ANN. This is because, as in most AI and ML applications, there is a need for functional specification in the estimation process. If α^* rejects the null that the data came from a Logistic DGP, one can then specify a different functional form for estimating ANN. In doing so, the researcher can then focus on optimizing model fit and prediction with respect to the number of hidden layers and/or neurons specified in each layer. As such, the methodology can be used in a number of complementary ways as a baseline to improve answers to scientific questions of interest to the researcher.

Evidently, while it is possible that in a particular dataset existing neural network and machine learning techniques (SVM) can outperform the proposed methodology, it is also possible that the proposed methodology can outperform existing AI and ML methods (Hu et al. 2020). As such, few models can be claimed to be superior without further contextual comparison of the dataset and application. However, it is crucial to point out that by changing the input criteria, any particular model may outperform another (the classic issue of data mining, p-hacking etc.). The true robustness of a particular method, therefore, should depend on the results attained in simulation and real-world applications without any such changes to the inputs. Furthermore, this should be done under completely diffuse assumptions for general robustness checks. The results attained above were consistent with this philosophy of scientific inquiry. As such the results point to the viability and robustness of the methodology under a wide array of applications[23].

In addition, the proposed model nests the Box-Cox transformation (Guerrero and Johnson 1982), as the Bayesian implementation ensures congruency to the MLE result of the paper under particular prior specifications. More specifically, if we specify a non-informative prior then the proposed model is similar to the MLE method on which the convergence and uniqueness of the Box-Cox transformation results mentioned above rely. However, the proposed method is more general in the sense that if we were to specify an informative prior, as we have done above with the subjective and diffuse priors of $N(0.5, 10)$ and $N(0, 10)$ respectively, then existence and uniqueness results follow both from Theorem 4 as well as Irreducibility[24] and Ergodocity[25] of the MCMC implementation above for any link modification problem. These conditions are easily satisfied under i.i.d. assumptions. It is important to also highlight that while the application of the methodology is through the Logistic link

[23]Furthermore, such applications can be done either on a standalone basis or for comparison purposes to better train AI and ML methods as well.

[24]A stochastic process that is Irreducible can visit all neighborhoods in the appropriate $\sigma - algebra$ with positive probability (Fouque et al. 2007).

[25]A stochastic process is Ergodic if the average of a function of the process goes to the ensemble average as time grows large ([Ibid]).

modification problem, it can equally be applied to any generalized linear model in the presence of the appropriate link modification. Consequently, the Box-Cox transformation [Ibid] is a specific version of the generalized framework presented here and is, therefore, nested within it!

While the convergence of the functional forms and their proofs provide solid foundations, it also must be acknowledged that no simulation result can be thought of as a proof in general. Though this is never the purpose of a simulation study alone, there is room here for further empirical verification of the results portrayed. However, as the simulation is done over both linear, non-linear and mixed model specifications, it provides a more complete picture than perhaps performing it on only one type of data. Nevertheless, the results are still dependent on the created covariates. However, since in GLM it is customary to consider the $\mathbf{X}$'s as fixed, it seems reasonable that this will be less of an issue here than in other model specifications. As an example, the empirical test was done on a relatively small dataset of the MIS field. Thus, it would be interesting to compare the results to other such datasets both in Informatics, Informetrics and more broadly in the sciences. Naturally, further applications of the methodology to varied datasets of different sizes and complexities in diverse fields are necessary to better understand its efficacy[26]. Furthermore, as with any numerical procedure, the convergence of the proposed methodology can also vary based on a multitude of criteria depending on the data. Though, this is the case for any numerical estimation process, and, therefore, other than for relatively rare boundary conditions it seems unlikely to change the conclusions above.

There are many extensions of this model which are left open for future researchers. For example, in terms of time series analysis, the efficacy of the model must be ascertained. Furthermore, there are numerous ordered and unordered multinomial models in the presence of heterogeneity and varying scientific phenomena, which need to be extended in this

[26]The author hopes to make general statitical packages available to the greater scientific community to apply the methodology.

framework and are left as open questions to be answered. There are also multiple AI applications as well for the methodology, from image recognition to understanding behavior of algorithms under varying input criteria. Given the excellent inference and classification results, and the fact that this was attained even in the presence of very few iterations, there is also much potential for the method to be used in large data contexts. Yet, another open area of research is to consider the groupings of α_i^*'s as representative of the various types of groups in the data, and thus can be thought of as a means of understanding the behavior within each sub-group as well.

One of the more useful results is that the findings are robust to when there are fewer choices or the frequency of success is low in a dataset. This is especially relevant for Informetrics and the physical, biomedical and social sciences. For example, one rarely sees the same number of articles being highly-cited as those that are not, or a dataset which has exactly the same number of agents choosing an alternative as those not choosing an alternative. Thus, the proposed method provides a reasonable step forward in modeling efforts under these scenarios, with natural extensions to today's large datasets.

2.9 Conclusion

In summary, the proposed methodology for Generalized Linear Models, and for the Generalized Logistic Link estimation procedure in particular, is seen to give equal or better model fit, inference and classification/prediction results, to existing methodologies when the assumptions of the model are relevant and the link condition is satisfied. Yet the methodology can give much better model fits, inference and prediction results especially for out-of-sample data when the traditional assumptions for GLMs are violated, even in comparison to AI and ML methods such as the ANN. Therefore, it is shown to be more flexible to violations of the assumptions on the error distribution, in both simulation and real-world applications.

Consequently, it is more robust, with better classification and inference outcomes compared to existing methodologies and can be used to understand relationships between scientific variables of interest far more scientifically. As such, it provides an expanded tool-set with which scientists, statisticians, mathematicians, analysts, researchers and managers can hone their correlational or causal understandings between variables. Furthermore, the results hold even in large data settings, where estimation can proceed with small sample sizes over each MCMC iteration, even in the presence of low frequency of successes observed in the data. However, as with any new methodology, the efficacy of the model still has much room for verification empirically in other contexts and is left up to future applications to the greater scientific community.

Chapter 3

Nonparametric Application

3.1 Introduction

As mentioned earlier, binary outcome models continue to be relevant for Artificial Intelligence (AI) and Machine Learning (ML) applications (see for example, Li et al. (2018); Hu et al. (2020); Chowdhury (2021a)), as they serve as the building blocks for various Multinomial extensions (e.g,. Allenby and Rossi (1998); Murad et al. (2003)). The type of usage whether latent variable (LV) or binary outcome (BO) is field specific. For example, in Econometrics LV models have been used to understand behavior of the average individual within a population (see Greene (2003) for a summary), for calculating propensity scores for causal interpretation and program evaluation (see Imbens and Rubin (2015) for an excellent summary), and also to understand heterogeneity through finite and infinite mixture distributions (Andrews et al. (2002)). In Psychology (Hofmans (2017)), Experimental Economics (Edelman et al. (2017)), Biomedical Sciences (Zhang et al. (2017)), and in the Physical Sciences (Hattab et al. (2018)) there is a extensive history of each formulation.

From Chapter 2, since the underlying assumptions of BO vs. LV models are distinctly different, it is difficult to reconcile divergence in model fit, inference or prediction (MIP) performance between them. Further complexities arise if the data is unbalanced as then the assumptions of Logistic or Probit models need not hold. Thus unsurprisingly, it is well established that the parameter estimates in these models, in either BO or LV models are susceptible to bias and inconsistency, (e.g., Simonoff (1998), Abramson et al. (2000), Maity et al. (2018)). To overcome some of these issues, (Chowdhury (2021a)) presented a parametric extension of the current Genralized Linear Model (GLM) framework. The work had multiple contributions. Applied in the Logistic formulation Chapter 2 showed that it can give equivalent performance to existing GLMs such as the Logit if the data supported their assumptions, but could give better MIP results if they were violated. The methodology also showed that it can give results better or equivalent to popular Artificial Intelligence (AI) methods such as Artificial Neural Network (ANN) under a wide range of circumstances with-

out loss of interpretability of parameter estimates. In addition, the methodology introduced a large-sample diagnostic test which could be used to improve existing AI methods. As such, the work presented a better baseline against which popular AI and machine learning (ML) methods could be compared with better coverage probabilities than existing widely used methodologies such as the maximum likelihood (mle) Logistic regression.

Further, the functional specification was shown to be highly flexible, since the link condition automatically adjusts to violations of the link constraint. This is because the link constraint is imposed to hold conditionally for all observations. However, the underlying probability of success in its formulation was assumed to be parametric in design. Thus, despite a flexible link function, the estimation of the model when the distribution on the underlying latent error differs from the parametric specification can potentially lead to minor techinicalities. As such, this paper adds to this parametric version presented in Chowdhury (2021a) using a nonparametric application which is shown to have certain advantages under focused applications. Though the parametric version remains relevant for inference, especially if the underlying DGP is symmetric, the nonparametric application is shown to improve upon it for classification purposes in training datasets and can outperform it in test datasets if the underlying DGP is asymmetric. The simulation studies also paint a more nuanced picture of when the parametric or nonparametric versions are more (or less) useful in comparison to existing AI, ML models or GLMs.

Thus, this chapter presents six meaningful extensions of the extant literature that spans all three aspects of MIP. In particular, for convergence results, in simulation studies I show that the convergence of the nonparametric application takes longer if the underlying DGP is asymmetric but has very similar performance to the parametric setting if the DGP is symmetric. For prediction, if the DGP is symmetric it can outperform the parametric version for training datasets but has very similar prediction performances in test datasets to the existing parametric version. Furthermore, for symmetric DGPs it has largely equivalent

or at best nominally better inference performance to the parametric methodology. However, if the data are asymmetric it has better overall prediction and inference performance to the existing parametric version. To better comprare classification performance among the various methodologies considered, I also introduce a new ROC-Statistic based statistical test (Chowdhury (2019)). This large-sample test allows for comparison of model performance to understand if the model results are statistically different. In addition, in one of the more useful applications of the methodology, through a separate large-sample test we can test whether any particular parametric assumption on the DGP actually fit the observed data, without any a priori assumption on the DGP itself. As such I show it to be an extension of Liu and Zhang (2018) in understanding model diagnostics for categorical data analysis. In what follows we first present the mathematical foundations of the methodology in Section 3.2, I then perform extensive simulation studies in Section 3.3 followed by multiple applications of the methodology to real-world datasets in Section 3.4. I then discuss the findings in Section 3.5 and finally end with some concluding thoughts in Section 3.6.

3.2 Methodology

Following the notation of Agresti (2003) we have that a GLM expands ordinary regressions to nonnormal response distributions and modeling functions. It is identified by three components, the random component for the response $\boldsymbol{y}$, a systematic component that outlines how the explanatory variables are related to the random components, and a link function. The link function specifies how a function of $E(\boldsymbol{y})$ relates to the systematic component. The random components of $\boldsymbol{y}$ are considered independent and identically distributed (i.i.d.). Henceforth, uppercase letters indicate matrices and bolded letters indicate vectors. Thus, consider a $n \times 1$ outcome variable $\boldsymbol{y}$ which is related to a $n \times (k+1)$ set of explanatory variables, $X = (\mathbf{1}, \boldsymbol{x_1}, ..., \boldsymbol{x_k})$, with $\boldsymbol{x_k}$ and $\mathbf{1}$ each $n \times 1$, through a continuous, bounded,

real valued function $c(X)$ of the same dimensions, namely $n \times (k+1)$. The usual $(k+1) \times 1$ parameters of interest are $\boldsymbol{\beta} = \{\beta_0, ..., \beta_k\}$, where each $\{\beta_k\}$ can be a vector.

In particular, in the one dimensional classical formulation if we have a sample of size n, $\{(y_1, \boldsymbol{x_1}), ..., (y_n, \boldsymbol{x_n})\} = \{(y_i, \boldsymbol{x_i})\}_{i=1}^n$, from the exponential family its pdf can be written as,

$$f(y_i; \beta_i) = a(\beta_i)b(y_i)exp[y_i q(\beta_i)]. \tag{3.1}$$

The values of $\boldsymbol{\beta}_i$ varies as a function of the explanatory variables and $q(\beta)$ is called the natural parameter. While a GLM is usually considered to be linear in both the explanatory variables and β, for example $\eta_i = \sum_{ij} \beta_j x_{ij}$[1] we may expand the systematic components to encompass a much broader array of functions and linear spaces in the random parameters model where,

$$\boldsymbol{\beta_i} = \{\boldsymbol{\beta}_{i0}, ..., \boldsymbol{\beta}_{ik}\}. \tag{3.2}$$

In particular, we may define,

$$\eta_i = \sum_j \boldsymbol{\beta}_{ij} c(x_{ij}). \tag{3.3}$$

Consequently, this model formulation may be thought of as a basis expansion on the usual regular topology and naturally subsumes the traditional formulation over the random parameter $\boldsymbol{\beta_i}$. Thus, a link function may be defined as

$$g(\mu_i) = \boldsymbol{\beta}_i' c(\boldsymbol{x}_i). \tag{3.4}$$

In order to retain the current models should the data support them, I follow the more general framework as in Chowdhury (2021a). For completeness, I restate the latent variable

[1]Further, if $g(\mu_i) = q(\beta_i)$, we call it the canonical link function.

formulation as considered in (Albert and Chib (1993)) here. Let $\boldsymbol{y}^*$ be a latent or unobserved continuous random variable. Then an index function model for binary outcome gives the GLM,

$$\boldsymbol{y}^* = c(X)\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.5}$$

where

$$y_i = \begin{cases} 1 \; if \; y_i^* \geq 0, \\ 0 \; if \; y_i^* < 0, \end{cases} \tag{3.6}$$

with the threshold 0 being a normalization. The two approaches have their strengths and weaknesses, and the purpose of this contribution is to better align the advantages of both models in a rigorous fashion using some of the findings of Chowdhury (2021a). For example, Albert and Chib [Ibid] clearly point out that in the binary regression case in the Frequentist interpretation any observed error can only take two values, either 1 or 0. On the other hand, in the latent variable formulation the existence of such an unobserved variable $\boldsymbol{y}^*$ is not guaranteed in the current formulation. The reasoning why the application still holds especially in the symmetric DGP case is due to Tanner and Wong (1987) as in Albert and Chib [Ibid], $\boldsymbol{y}^*$ is integrated out. In making such an assumption it is also necessary to fix the variance of the latent distribution for identification purposes. An approach which under most circumstances would be considered restrictive.

The proposed methodology has two main goals. It first seeks of a way to incoporate the strengths of both the latent variable and the binary regression in a mathematically rigorous form. Secondly, it seeks to overcome restrictive assumptions on the latent distribution such as having a constant variance or assuming a particular distribution on the probability of success. Below I first outline how the two methodologies may be combined. Then I outline

a different methodology for the latent variable formulation using signed measures which has distinct advantages over the current latent variable formulation. I then show how both of these methodologies can be combined in a unified framework.

3.2.1 Equivalency of Binomial Regression and Latent Variable Formulations

To see the equivalency of the models, first note that by construction of the binary regression,

$$E(\boldsymbol{y}|X) = F(\boldsymbol{y} = \mathbf{1}|X) = \boldsymbol{p}(X), \tag{3.7}$$

where F is a prespecified cdf under the binary regression case. Going forward, for notational simplicity I will suppress the express dependence of $\boldsymbol{p}$ on X unless otherwise stated. Accordingly, this induces a pmf of,

$$f(y_i|\boldsymbol{x}_i) = p_i^{y_i}(1-p_i)^{(1-y_i)}. \tag{3.8}$$

Clearly, symmetry and the cutoff of 0 are important for the equivalency of the binomial regression and its latent variable application. Let F^* denote the symmetric cdf of the latent variable, and F the cdf of the binomial regression model, then if we consider the random parameter formulation we have,

$$\boldsymbol{p}_i = F(\boldsymbol{x'}_i\boldsymbol{\beta}_i) = F^*[y_i^* > 0] = F^*[-\epsilon_i < \boldsymbol{x'}_i\boldsymbol{\beta}_i] = F^*[\boldsymbol{\epsilon}_i < \boldsymbol{x'}_i\boldsymbol{\beta}_i] = F^*[\boldsymbol{x'}_i\boldsymbol{\beta}_i]. \tag{3.9}$$

Of course, if we set the latent variable and the binomial regression probability of successes to be the same we get pointwise equivalency between the two models under the assumptions above. Further, under i.i.d. assumptions, for the binomial regression WLOG for $k \leq n$

successes,

$$L(X|\boldsymbol{y},\boldsymbol{\beta}) = \prod_{i=1}^{n} F(y_i,\boldsymbol{x}_i,\boldsymbol{\beta})^{y_i}(1-F(y_i,\boldsymbol{x}_i,\boldsymbol{\beta}))^{1-y_i}, \tag{3.10}$$

where we can interchange the LV and BO model distributions if the assumptions on the latent variable and its existence actually hold.

Indeed, if the assumptions above are not true regarding the latent variable formulation the congruency between the two models become less clear since the errors for the binomial regression case again can only take the values of 0 or 1 in the Frequentist formulation. As an example, if the latent variable distribution is not symmetric it is entirely possible that,

$$p_i = F(\boldsymbol{x'}_i\boldsymbol{\beta}_i) \neq F^*[-\epsilon_i < \boldsymbol{x'}_i\boldsymbol{\beta}_i], \tag{3.11}$$

given the observed data, and therefore equivalency is necessarily lost.

As such, a more flexible latent variable formulation is needed to ensure that equivalency holds. Accordingly, consider a more flexible methodology using a general functional analysis rooted perspective. For its application note that by construction,

$$E(y_i) = F(\boldsymbol{x'}_i\boldsymbol{\beta}_i), \tag{3.12}$$

which is consistent only if the specification of F is the true distribution function for the probability of success. This is under most circumstances, unknown. On the other hand, regardless of the assumed distribution for the latent probability of success, even if it is assumed to be the same as that assumed for F, asymmetry or unbalancedness in the data may imply (3.11). Therefore, it is reasonable to expect different model fit, prediction and inference result from each. Unsurprisingly, therefore this is exactly what is seen in application throughout the sciences. Yet there are multiple virtues of the Bayesian approach to binary

and polychotomous data as illustrated in Albert and Chib (1993). Chief among these are the continuous nature of the latent error and the ability to use a data augmentation approach as in Tanner and Wong (1987).

One of the principle contribution of the current work is in detailing how to utilize these virtues while still overcoming the identifiability concerns addressed. In particular, not only do we wish to fit a more flexible nonparametric distribution on the latent probability of success, but we also want to allow its distributional parameters to be free from artificial constraints, such as the need to fix its variance for identification of these same parameters. In addition, we would like to implement such a methodology so that it is *equivalent* in some manner to the true known data likelihood, namely binomial under i.i.d. assumptions.

Accordingly, note that if due to reasons above (3.11) occurs we may consider a pointwise transformation such as,

$$(F^*[-\epsilon_i < \boldsymbol{x'}_i\boldsymbol{\beta}_i])^{\boldsymbol{\alpha}^*} = E(y_i) = F(\boldsymbol{x'}_i\boldsymbol{\beta}_i). \tag{3.13}$$

With such a transformation, which holds pointwise for some $\boldsymbol{\alpha}^* \in \mathbf{R} \setminus \{-\infty, \infty\}$, we may specify a fully nonparametric distribution on the latent probability of success while maintaining pointwise equivalency to the underlying binary model. This relationship should hold whether the true latent distribution is symmetric or asymmetric. In the asymmetric case we need not then ensure that the cutoff for a probability of success is some particular prefixed support point such as 0, but rather may let this cutoff be based on a probability as a function of the observed data. As such, if we are able to specify such a distribution nonparametrically we can ensure even without any artificial preconceived restrictions on the distributional parameters that pointwise, the probabilities of successes match.

Further note that beyond the pointwise convergence of the binary regression and latent variable probability of success in an observed sample we would like to say more about the

equivalency of these two models overall in terms of the likelihood. That is to say that since pointwise convergence by iteself certainly does not guarantee strong or almost sure convergence, one would like to relate these two models in a more concrete way. In fact, it is clear that under the two specifications the likelihoods for the latent variable and that for the binary outcome models are not necessarily the same. Yet note that the link condition as expressed above in (3.13) ensures that the probability of successes are pointwise convergent for both models. Accordingly, if we define $L(\boldsymbol{\beta}|\boldsymbol{y})$ as the likelihood w.r.t. the observed data and $L(\boldsymbol{\beta}|\boldsymbol{y^*})$ as the likelihood w.r.t. the latent outcomes then by Birnbaum's Theorem (Casella and Berger (2002), Theorem 6.3.6) we may write,

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \boldsymbol{c}(\boldsymbol{y}, \boldsymbol{y^*})L(\boldsymbol{\beta}|\boldsymbol{y^*}), \tag{3.14}$$

for some constant $\boldsymbol{c}(\boldsymbol{y}, \boldsymbol{y}^*)$. While the mathematical results follow below, intuitively this seems a very reasonable condition since $\boldsymbol{y}^*$ is a function of $\boldsymbol{y}$ through both $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$. In fact, that such a relationship should hold for any two experiments or sample observations that claim to explain the same underlying stochastic process is also entirely logical. Thus, in the Bayesian formulation it straightforwardly follows that the posterior distribution of $\boldsymbol{\beta}$, should satisfy $p(\boldsymbol{\beta}|\boldsymbol{y}^*) \propto p(\boldsymbol{\beta}|\boldsymbol{y})$. As such, we must have that the inferences drawn from such a combination of the binary and latent formulations should be identical.

Note however, that while Birnbaum's Theorem ensures the inferences drawn are identical under the two likelihoods, model fit and prediction results need not be identical at all. To ensure the existence and uniqueness of the latent variable specification, it is necessary to go further to ensure the equivalency of the two models. For this purpose, we need to consider a signed measure.

Thus, in what follows I lay the groundwork for the viability of the model specifications. To ensure identifiability and equivalency to existing models, without loss of generality I focus the

nonparametric methodology to depend on only one parameter α^*, such that the generalized link condition holds for each observation. Below I first discuss the methodology and then introduce a new (to the best of the author's knowledge) statistic called Adjusted ROC-Statistic (ARS) for classification in categorical models. I then discuss how the methodology can be used for model diagnostics for any parametric assumption on the underlying DGP such as Logistic or normal specifications. Then I expand on the nonparametric methodology and give the necessary proofs.

3.2.2 Discussion on Existence and Uniqueness of Signed Measure

To begin our discussion consider the usual Logistic regression for the binary specification when we specify

$$F(\mathbf{x}_i, \boldsymbol{\beta}) = (1 + exp(\mathbf{x}_i\boldsymbol{\beta})^{-1})^{-1}. \tag{3.15}$$

In this formulation we get the Logit link function as the familiar log-odds ratio,

$$log\left(\frac{\boldsymbol{F}(\boldsymbol{y}, X, \boldsymbol{\beta})}{\mathbf{1} - \boldsymbol{F}(\boldsymbol{y}, X, \boldsymbol{\beta})}\right) = \boldsymbol{\lambda}(X, \boldsymbol{\beta}) = c(X)\boldsymbol{\beta}, \tag{3.16}$$

where following Chowdhury (2021a) λ is a continuous, bounded function of the explanatory variables in the GLM framework. In the nonparametric setting this condition need not hold if F, the probability of success is not the Logistic distribution. If the underlying probability of success does not follow a Logistic distribution this formulation should not hold. Furthermore, the underlying probability of success and its distribution is something that is inherently unknown and as before is a function of the observed X's. In fact, even assuming a nonparametric distribution on p_i is insufficient to ensure almost sure convergence. Furthermore, simply assuming a nonparametric distribution is also not enough to ensure equivalency between a binary regression framework and its latent variable specifica-

tion (Proposition 3.5). These insights follow straightforwardly from the convergence in Law or distribution of the nonparametric distribution to the true distribution. This is because convergence in distribution of a nonparametric estimator is of course not enough to ensure convergence in probability or almost surely. Thus, ensuring almost sure convergence utilizing the current latent variable framework requires a more robust formulation. To facilitate this we must impose a measure theoretic construct on the latent variable formulation. In particular, we must consider signed measures (which can take both positive and negative values) over the link constraint support on $\lambda(X, \boldsymbol{\beta})$.

Thus, consider any measure space $(X_0, \mathcal{M}_0, \nu_0)$ and a restriction of the $\sigma-$algebra $\mathcal{M}_0$ to $\Sigma \subseteq \mathcal{M}_0$ such that the link condition holds for the linear space $X \subseteq X_0$. Then the results below show that (X, Σ, ν) is also a measure space where ν is the restriction of ν_0 to Σ. In particular, note that by Skohorod we may define a random variable $\lambda(X, \boldsymbol{\beta})$, such that $(\lambda(X, \boldsymbol{\beta}), \Sigma)$ is also a measureable space. As such consider a signed measure $\boldsymbol{\nu}$ on $\{\zeta \in \boldsymbol{\lambda}(X, \boldsymbol{\beta})\}$ for any given $\boldsymbol{\beta}$ and define $\mathbf{A} = \{\boldsymbol{y}^* \in \zeta : \boldsymbol{\lambda}(X, \boldsymbol{\beta}) = \boldsymbol{y}^* \geq \boldsymbol{\kappa}\}$ and $\boldsymbol{B} = \{\boldsymbol{y}^* \in \zeta : \boldsymbol{\lambda}(X, \boldsymbol{\beta}) = \boldsymbol{y}^* < \boldsymbol{\kappa}\}$, where $\boldsymbol{\kappa} \in (0, 1)$ thus $A \cap B = \emptyset$. Further let $\mathbf{A}$ and $\mathbf{B}$, be a Hahn Decoposition of ζ, for which $\mathbf{A}$ is positive w.r.t. $\boldsymbol{\nu}$ and $\mathbf{B}$ is negative w.r.t. $\boldsymbol{\nu}$. Define, $\boldsymbol{\nu}^+(E) = \boldsymbol{\nu}(E \cap A)$ and $\boldsymbol{\nu}^-(E) = -\boldsymbol{\nu}(E \cap B)$ for some arbitrary $E \in \zeta$. Then by using the Hahn Decomposition Theorem as well as the Jordan Decomposition Theorems we have,

$$\boldsymbol{\nu} = \boldsymbol{\nu}^+ - \boldsymbol{\nu}^-, \tag{3.17}$$

with the mutually singular measures $\{\boldsymbol{\nu}^+, \boldsymbol{\nu}^-\}$ unique. Thus, there exists a τ^* such that element wise,

$$\nu(\boldsymbol{x}_i, \boldsymbol{\beta})^{\tau *} = \lambda(\boldsymbol{x}_i, \boldsymbol{\beta}) \; and \; \tau \in \mathbf{R} \backslash \; \{\text{-}\infty, \infty\}, \tag{3.18}$$

holds for each observation. The existence of such a measure and two positive measure

$\{\nu^+, \nu^-\}$, are guaranteed by the construction of the link function condition in (3.13) and will be discussed in greater detail in the mathematical results below. For the current purpose, note that the uniqueness of the positive measures allows us an extremely useful way to ensure the link condition holds pointwise for each observation. Specifically, we know that if the probability of success is given by a nonparametric distribution $\hat{\boldsymbol{F}}(\boldsymbol{x}_i, \boldsymbol{\beta})$ then the probability of failure can be given by $\mathbf{1} - \hat{\boldsymbol{F}}(\boldsymbol{x}_i, \boldsymbol{\beta})$. Thus, if $\boldsymbol{\nu}^+ = \hat{\boldsymbol{F}}(\boldsymbol{x}_i, \boldsymbol{\beta})$, then $\boldsymbol{\nu}^- = (\mathbf{1} - \hat{\boldsymbol{F}}(\boldsymbol{x}_i, \boldsymbol{\beta}))$. Since the labels of success or failure are arbitrary, the above formulation can easily be reversed if the probability of failure is considered as a success and as such the link function formulation with $\boldsymbol{\nu}^+ = (\mathbf{1} - \hat{\boldsymbol{F}}(\boldsymbol{x}_i, \boldsymbol{\beta}))$ would also be valid if a probability measure exists that can represent $\boldsymbol{\nu}$ in such a way.

For the present discussion let us proceed under the assertion that such a probability measure exists (please see the mathematical results section for rigorous proofs of this assertion and other relevant results). Then it must be that such a measure exists and that it is unique. Then using Skohorod one may easily define a nonparametric distribution $\hat{\boldsymbol{F}}(X, \boldsymbol{\beta})$ such that $\hat{\boldsymbol{F}}(A) = \int_A \boldsymbol{\nu}^+$ and $\hat{\boldsymbol{F}}(B) = 1 - \int_A \boldsymbol{\nu}^+ = \int_B \boldsymbol{\nu}^-$, which may or may not be symmetric by construction and therefore the probabilities of successes and failures need not necessarily approach 1 or 0 at the same rate. Therefore, the link function is also not necessarily symmetric by construction, and is dependent on the observed data, as it should be. As such, the methodology encompasses the existing latent variable formulations if the data support them.

To illustrate the methodology, let us continue with the Logistic regression example and note that if $\boldsymbol{\lambda}(\boldsymbol{x}_i, \boldsymbol{\beta}) \geq \mathbf{0}$,

$$log\left(\frac{\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x}_i, \boldsymbol{\beta})}{\mathbf{1} - \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x}_i, \boldsymbol{\beta})}\right) = \boldsymbol{\lambda}(\boldsymbol{x}_i, \boldsymbol{\beta}) = \hat{\boldsymbol{F}}(\boldsymbol{x}_i, \boldsymbol{\beta})^{\boldsymbol{\alpha}^*}, \tag{3.19}$$

holds for some α^* pointwise. Therefore, the link modification is valid for any binary latent

variable formulation. If $\boldsymbol{\lambda}(\boldsymbol{x}_i, \boldsymbol{\beta}) < \mathbf{0}$ the indicies of success or failure can be reversed such that 3.19 again can hold for every observation. In fact, this allows for a more flexible latent formulation since we do not have to assume a particular probability of success for F such as the Logistic. Nor do we have to fix the variance parameter of a parametric distribution such as the Logistic for identification purposes.

Thus, the preceding discussion shows how Birnbaum's Theorem and the Jordan Decomposition Theorems can be used to relate the binary regression and any latent variable formulation and extend the current latent variable formulations accordingly. Below I now expand on formal mathematical proofs that verify the assertions above. First I give the results that lay the mathematical foundations of this new methodology which ensures equivalency between the two formulations, and also ensures the unique identification of the true but unknown distribution of the probability of success.

3.2.3 Mathematical Results

Below, I first present some relevant definitions and then present the mathematical foundations for the discussions above. For a discussion on when the latent variable and binomial regression formulations are equivalent, I refer the reader to Chapter 4. For the current formulation I assume that the formulations are equivalent with any differences elaborated accordingly in the results. While the circumstances, under which the equivalency is lost is more general, the mathematical foundations below give a more general functional analysis perspective on the Bayesian Hierarchical formulation which ensures almost sure convergence if the formulations are equivalent. Thus, I take the results of the previous chapter and the first two sections of the current chapter to present a more coherent framework that unifies the methodologies in a mathematically rigorous way.

3.2.3.1 Definitions

The following definitions can be found in any graduate level measure theory book and are restated for completeness.

Definition 3.1. *A signed measure ν_0 on a measureable space (X_0, M_0) is defined as a real-valued function $\nu_0 : M_0 \to [-\infty, \infty]$ such that*

1. *ν_0 attains at most one of the values ∞ or $-\infty$.*
2. *$\nu_0(\emptyset) = 0$.*
3. *Finite and countable additivity holds for disjoint measureable sets and all measureable sets respectively.*

It is well established from analytic theory that the restriction of a measure space to a subset of the measureable space is also measureable. However, for the current GLM construction we want to focus on a particular subset, that of the link function. Note that from previous discussions we know that by construction a link function relates the systematic components to the mean of an observation in a specified manner,

$$\eta_i = \lambda(\boldsymbol{x}_i, \boldsymbol{\beta}). \tag{3.20}$$

The novelty of this methodology is in considering a signed measure over the $\sigma - algebra$ defined on the support of the link function.

3.2.3.2 Mathematical Foundations of the Proposed Methodology

Accordingly, first note that the restriction of a measure space to a subspace of the $\sigma - algebra$ is itself a measure space. For the purpose at hand we seek to restrict our attention to the subspace of the sample space that ensures that the link condition holds pointwise. Thus, consider the measure space (X, Σ_0, ν_0) restricted to a subspace of X for which the link condition holds for any particular $\boldsymbol{\beta}$. Then from elementary analysis we know that $(X, \Sigma_{0|\boldsymbol{\lambda}(X,\boldsymbol{\beta})} = \Sigma, \nu_{0|\boldsymbol{\lambda}(X,\boldsymbol{\beta})} = \nu)$ is also a measure space. The results below outline this in a more rigorous way.

Proposition 3.1. *For the measureable space* $(\lambda(X, \boldsymbol{\beta}), \Sigma)$ *There exists a signed measure* ν *such that,*

$$y^* = \begin{cases} 1 \ if \ \lambda \geq 0 \\ 0 \ if \ \lambda < 0, \end{cases} \tag{3.21}$$

where WLOG $\lambda \in [-\infty, \infty)$.

Proof. Let us first consider the finite case. Note that by construction of the latent variable formulation, the measure of success is given by

$$E[\boldsymbol{y}^*|X] = \boldsymbol{\nu}(X, \boldsymbol{\beta}) = c(X)\boldsymbol{\beta} = \boldsymbol{\lambda}(X, \boldsymbol{\beta}). \tag{3.22}$$

But $\boldsymbol{\lambda}(X, \boldsymbol{\beta}) \in R \setminus \{-\infty, \infty\}$, and thus $\boldsymbol{\nu}(X, \boldsymbol{\beta}) \notin \{-\infty, \infty\}$. Therefore, ν may be finite and the measureable space $(\lambda(X, \boldsymbol{\beta}), \Sigma)$ may be represented as a Bounded Finitely Additive measure space.

It remains to show then that $\boldsymbol{\nu}$ may also be $\sigma - finite$. Note that $\boldsymbol{\lambda}(X, \boldsymbol{\beta})$ is a bounded

continuous functional specification. Note, WLOG

$$\{X, \boldsymbol{\beta} : \boldsymbol{\lambda}(X, \boldsymbol{\beta}) < c\} \; for \; any \; c \in \mathbf{R} \setminus \{\infty\} \tag{3.23}$$

is a measureable set. But $\mathbf{R}$ can be expressed as countable unions of measureable sets for each E_k. Thus, $\boldsymbol{\lambda}(\boldsymbol{x}, \boldsymbol{\beta}) \subseteq \cup_{k=1}^{\infty} E_k \; s.t. \; \boldsymbol{x} \in X \in E_k$ given $\boldsymbol{\beta}$. But from the preceding discussion $\nu(E_k) < \infty$, through the link constraint holding for each observation. Therefore,

$$\lim_{n \to \infty} \nu(E_k) \to -\infty, \; where \; \boldsymbol{\lambda}(X, \boldsymbol{\beta}) < 0. \tag{3.24}$$

Therefore the measureable space $(\lambda(X, \boldsymbol{\beta}), \Sigma)$ may also be $\sigma-$finite, as needed.

Now WLOG let S^+ be a collection of subsets of $\lambda \in [0, \infty)$ and S^- be a collection of subsets of $\lambda \in [-\infty, 0)$ and consider a $\sigma-$finite measure space. Since $\nu(S^+) \geq 0$ and $\nu(S^-) < 0$, we have $S^+ \cup S^- = \mathbf{R} \setminus \{-\infty, \infty\}$ and $S^+ \cap S^- = \emptyset$. Thus, (S^+, S^-) is a Hahn Decomposition of $\mathbf{R} \setminus \{-\infty, \infty\}$.

We have thus established existence. $\square$

Above and for the remainder of this manuscript for notational simplicity I employ λ to represent $\lambda(X, \boldsymbol{\beta})$. The preceding proposition established the existence result. The forthcoming proposition establishes the uniqueness of this construction for the finite case.

Proposition 3.2. *For the measureable space (λ, Σ) there exists an unique decompostion of the signed, finite measure ν as a function of two positive measures ν^+ and ν^- such that,*

$$\nu = \nu^+ - \nu^- \; where, \tag{3.25}$$

$$y_i^* = \begin{cases} 1 \; if \; \lambda \geq 0 \\ 0 \; if \; \lambda < 0, \end{cases} \tag{3.26}$$

and $\lambda \in (-\infty, \infty)$.

Proof. As before, let S^+ be a countable collection of subsets of $\lambda \in [0, \infty)$ with $S^+ \subseteq \lambda_{|\geq 0}$ and S^- be a countable collection of subsets of $\lambda \in (-\infty, 0)$ with $S^- \subseteq \lambda_{|<0}$. Since $\nu(S^+) \geq 0$ and $\nu(S^-) < 0$, we have $S^+ \cup S^- = \mathbf{R} \setminus \{-\infty, \infty\}$ and $S^+ \cap S^- = \emptyset$. Thus, (S^+, S^-) is a Hahn Decomposition of $\mathbf{R} \setminus \{-\infty, \infty\}$ and further define,

$$\nu^+ = \nu(E \cap S^+), \text{ and } \nu^- = -\nu(E \cap S^-), \text{ for } E \in \Sigma. \tag{3.27}$$

Thus, $\nu^+(S^-) = \nu^-(S^+) = 0$ and the positive measures are mutually singular. Therefore, by the Jordan Decomposition Theorem the pairs $\{\nu^+, \nu^-\}$ are unique. □

Proposition 3.3. *Let* (λ, Σ, ν) *be a measure space as above and* ν *a finite signed measure on it. Then,*

$$|\bar{\nu}|(\Sigma) = \bar{\nu}^+(S^+) + \bar{\nu}^-(S^-) \tag{3.28}$$

is a probability measure, where $\{\bar{\nu}^+, \bar{\nu}^-\}$ *are positive measures and* S^+ *is positive, but* S^- *is negative w.r.t. the signed measure* ν.

Proof. First note that since by construction ν is finite, we must have that both ν^+ and ν^- must also be finite. Further Σ is a semiring, and thus using the Caratheodory-Hahn Theorem, we know there exists a $\mu : \Sigma \to [0, \infty)$. Then the Caratheodory measure defined as $|\bar{\nu}| = \bar{\nu}^+(S^+) + \bar{\nu}^-(S^-)$ induced by μ is an extension of μ and $|\bar{\nu}|$ is an unique extension and it is a finite measure. The statement of the proposition then is a straightforward consequence. □

Before going to the proof of the next result we need another consequence of an infinite measure space with some collection of measureable sets which are finite.

Theorem 3.1. *Let (X, Σ, μ) be any measure space with $\{E_k\}_{k=1}^{\infty} \subseteq A \subset \Sigma$ a collection of measureable sets for which $\mu(A) = \infty$. Then there exists some E_j where j is a countable collection of some k, such that*

$$\mu(\cup_j E_j) < \infty. \tag{3.29}$$

Proof. Case I: Let A consist of singleton sets of infinite measure.

Then there is nothing to prove as

$$\mu(\cup_{i=1}^{n} A_i \sim A_{j \neq i}) < \infty, \tag{3.30}$$

where A_j indicate the collection of singleton sets of infinite measure.

Case II: Let A be dense in Σ. We seek to find a countable collection of measureable sets that have finite measure while satisfying the condition of the theorem.

By construction,

$$\sum_{k=1}^{\infty} \mu(E_k) \leq \infty. \tag{3.31}$$

Choose a number $m_{ik} \in \boldsymbol{R} \setminus \{\infty, \infty\}$, such that

$$E_{ik} = x \in A : \mu(E_{ik} < 1/m_{ik}). \tag{3.32}$$

Then by the finite additivity and countable monotonicity of a measure, there exists a disjoint collection of sets, $E_{ik} \subset E_k$, such that there exists an enumeration of E_{ik} to the natural

numbers where the following condition holds,

$$\mu\left(\cup_n \cap_{ik\geq n} E_{ik}\right) = \sum_{i=1}^{n} \mu(E_{ik}) \leq 1. \tag{3.33}$$

Further, we know by Borel-Cantelli there exists measureable sets $E_{jk} \subset E_k$ such that

$$\mu(\cap_n \cup_{jk\geq n} E_{jk}) = 1. \tag{3.34}$$

Thus, there exists a disjoint countable collection of measureable sets in A such that $\mu(\cup_{jk=1}^{n} E_{jk}) = 1$. Therefore, there exists a set E_k with measure 1. Now by assumption $\mu(A) = \infty$, therefore there exists some n_2 and n_3 belonging to the naturals such that,

$$\mu\left((\cup_{k=1}^{n_2} E_k) \cup_{n_3\geq n_2}^{\infty} (A \sim (\cup_{k=1}^{n_2} E_k))\right) \leq \mu(A). \tag{3.35}$$

Take $n_2 \to \infty$ to get by finite additivity of measure that,

$$\lim_{n_2\to\infty} \sum_{n_2} \mu(E_k) + \sum_{n_3\geq n_2} (A \sim (\cup_{k=1}^{n_2} E_k)) \leq \infty, \tag{3.36}$$

thus,

$$\lim_{n_2\to\infty} \sum_{n_2} \mu(E_k) \to \infty. \tag{3.37}$$

But k was arbitrary, and therefore denote by $C = \cup_{k=1}^{j} E_k \subset A$ to get

$$\lim_{(k\to j)} \mu(C) \to \infty \implies \mu(C \sim A) < \infty, \tag{3.38}$$

as needed. □

The existence of this result will be important in dealing with signed measures, which can take one of the nonfinite values which is not $\sigma-$finite. Therefore, utilizing this result we can now prove one of the more important results and corollaries of Theorem 3.1, which can be important for the construction of finite or $\sigma-$finite measure spaces for all GLMs. It is detailed below.

Proposition 3.4. *Let (λ, Σ) be a measureble space with μ a measure which is neither finite or $\sigma - finite$ such that $\nu(\Sigma) = \infty$ and let $(\lambda, \Sigma, |\bar{\nu}|)$ be a $\sigma - finite$ measure space. Then if the signed measure ν takes one of the values of $\{-\infty, \infty\}$ then either $y = 1$ or $y = 0$ for every observation w.r.t. the $\sigma - finite$ measure.*

Proof. By definition the signed measure does not take both values of $-\infty$ or ∞, and the labels of $y = 1$ and $y = 0$ are arbitrary. Thus, WLOG let $\mu = -\infty$ over S^- where S^- is defined as before. Further, note that ν^+ must be a finite measure and ν^- a $\sigma-$finite measure. Choose ϵ and define,

$$E_n^+ = \{y \in \Sigma | \nu_k^+(y) > (1 - \epsilon)\phi^+(y) \ \forall \ k \geq n\}, \tag{3.39}$$

and its complement

$$E_n^c = \{y \in \Sigma | \nu_z^-(y) > (1 - \epsilon)\phi^-(y) \ \forall \ z \geq t\}, \tag{3.40}$$

where ϕ^+ is a simple approximation of ν^+ and ϕ^- is a simple approximation to ν^- w.r.t. the signed measure ν. Let us define $M^+ > 0$ then to be the maximum over all values of $\{\phi^+, \phi^-\}$. Choose an index $N = max\{N^+, N^-\}$ where N^+ is s.t. $\nu^+(S^+ \sim E_n^+) < \epsilon$ for all $n_1 \geq N^+$ and N^- is chosen s.t. $\nu^-(S^- \sim E_n^c) < \epsilon$ for all $n_2 \geq N^-$. Then by additivity over

domains of integrals, linearity and monotonicity and using Fatou's Lemma, we may define,

$$|\bar{\nu}| \leq lim\ inf \frac{\sum_{i=1}^{n} \nu^{-}(E_i^c)}{\int_{S^-} \phi_n^-\ d\nu} + lim\ inf \frac{\sum_{i=1}^{n} \nu^{+}(E_i^+)}{\int_{S^+} \phi_n^+\ d\nu}. \tag{3.41}$$

Thus by Theorem 3.1 we have,

$$|\bar{\nu}| \leq 0 + lim\ inf \frac{\sum_{i=1}^{n} \nu^{+}(E_i^+)}{\int_{S^+} \phi_n^+\ d\nu}. \tag{3.42}$$

But $|\nu| \in [0, \infty]$ by definition and $\nu^+ \in [0, \infty)$ by construction. Thus,

$$|\bar{\nu}| \geq 0 + lim\ sup \frac{\sum_{i=1}^{n} \nu^{+}(E_i^+)}{\int_{S^+} \phi_n^+\ d\nu}. \tag{3.43}$$

Therefore,

$$|\bar{\nu}| = \frac{\sum_{i=1}^{n} \nu^{+}(E_i^+)}{\int_{S^+} \phi_n^+\ d\nu} \leq 1. \tag{3.44}$$

Since $\int_{S^+} \phi_n^+\ d\nu < \infty$ (3.44) is well defined and the assertion follows. $\square$

This result has some very important consequences on the usual regression analysis widely used in the sciences. For example, we may now consider a finite measure such that for a measureable set $E \in \Sigma$,

$$|\bar{\nu}| = |\nu|\delta_{E \cap S^+},\ where\ \delta_{E \cap S^+} = \begin{cases} 1\ if\ E \in S^+ \\ 0\ o.w. \end{cases} \tag{3.45}$$

The usefulness of the result follows from the unique Jordan Decomposition of a signed measure. If f is a Lebesgue integrable function then existing results from analysis can be used through the translation invariance property of the measure, to find the unique functional specification as demonstrated in a forthcoming corollary.

The astute reader no doubt realizes that in the case that the signed measure takes one of the $\{-\infty, \infty\}$ values, over a measure space which is neither finite nor $\sigma - finite$, then we may have information loss. The resulting reformulation to the restricted measure space given in Theorem 3.1 and Proposition 3.4, can be overcome to address this information loss concern, and the following proposition addresses this. To that end, let us define the following.

Definition 3.2. *For p any real-valued function defined on a linear space X, we say it is positive homogeneous if*

$$p(ax) = ap(x), \tag{3.46}$$

for all $a > 0$ for every x in X.

Definition 3.3. *For p any real-valued function defined on a linear space X, we say it is subadditive if*

$$p(x + y) \leq p(x) + p(y), \tag{3.47}$$

for every x and y in X.

These definitions can be used with the Hahn-Banach Theorem to define a linear functional over all integrable functions in $L^p(\lambda, \Sigma, |\bar{\nu}|)$ with $1 \leq p < \infty$.

Proposition 3.5. *Consider a Hahn-Decomposition of the measure space (λ, Σ, ν) into $\{S^+, S^-\}$ as defined before, where the signed-measure ν WLOG takes the value of $-\infty$ but is not $\sigma-$finite. Then there exists a linear functional $\mathcal{L}$ which extends any measure ν^+ over S^+ to*

all of $L^p(Q, |\bar{\nu}|)$, with $|\bar{\nu}|$ as in Proposition 3.4, and Q is the measureable space (λ, Σ) with $1 \leq p < \infty$.

Proof. This is a consequence of the Hahn-Banach Theorem. First note let f be a non-negative function on L^p restricted to S^+. Let ψ be a Simple Function on λ the subspace of $L^p(X_{|S^+}, \Sigma, |\bar{\nu}|)$. Let f be any bounded, continuous function on the subspace λ of X. Thus, from elementary measure theory we know by the simple approximation theorem that there exists a sequence ψ_n such that,

$$|\psi_n - f|^p \leq 2^p.|f|^p \text{ on } \lambda \text{ for all } n. \tag{3.48}$$

Further by construction $|f|^p$ is integrable so by Lebesgue Dominated Convergence we know that $\{\psi_n\}$ converges. Therefore the simple functions are dense and subadditive for the metric induced by the norm on L^p. That $\{\psi_n\}$ is positively homogeneous is straightforward and thus, the result is asserted without proof here.

Therefore, by the Hahn-Banach Theorem we have that there exists a linear functional $\mathcal{L}$ such that,

$$\mathcal{L}(\lambda) \leq \{\psi_n\}(\lambda), \tag{3.49}$$

and further that it can be extended to all of X with the same norm. □

Using these results then we have some useful existing results from Real Analysis which can be restated for the specific purpose at hand. They are detailed below.

Corollary 3.1. *Let $(\lambda, \Sigma, |\bar{\nu}|)$ be a $\sigma-$finite measure space, where $|\bar{\nu}|$ is defined as in proposition 3.4. Let $\{f_n\}$ be a sequence of bounded Lebesgue measureable functions finite a.e. that converges p.w. a.e. on the set $E \in \Sigma \setminus S^-$ to f which is also finite a.e. on E. Then,*

$$\{f_n\} \to f, \tag{3.50}$$

in measure.

Proof. Note that by construction $|\bar{\nu}| < \infty$. Therefore, the result follows from elementary analysis since using Egoroff's Theorem the sequence of functions $\{f_n\}$ is uniformly integrable. Then an use of the Vitalli Convergence theorem for Lebesgue Integrable functions proves the theorem. The uniqueness of this measure follows from the Hahn Decomposition Theorem.□

Corollary 3.2. *Let $(\lambda, \Sigma, |\nu|)$ be a $\sigma-$finite measure space, where $|\nu|$ is defined as in proposition 3.3. Let $\{f_n\}$ be a sequence of bounded Lebesgue measureable functions finite a.e. that converges p.w. a.e. on the set $E \in \Sigma$ to f which is also finite a.e. on E. Then,*

$$\{f_n\} \to f, \tag{3.51}$$

in measure.

Proof. Note that by construction $|\nu| < \infty$. Therefore, the result follows from Corollary 3.1 straightforwardly.□

While convergence in probability is useful, below I show a stronger result in Theorem 3.3. Furthermore, these results have several nonintuitive applications to non-binary analysis and the remarks below highlight some of them.

Remark 3.1. *First, note that the decomposition above is unique, and as such the existence of a latent variable implies the existence an unique pair of positive measures that can represent it, and vice versa.*

Remark 3.2. *If the signed measure takes one of the values of* $\{-\infty, \infty\}$ *but is not* $\sigma-$*finite, we may represent any continuous generalized linear model in one of the outcomes with possibly a linear transformation that ensures all observed* $(\boldsymbol{y}, \boldsymbol{x})$ *as a function of* $\boldsymbol{\beta}$ *are positive or negative (WLOG). As such the traditional regression formulation can similarly be improved using the link-constraint condition holding for each observation!*

Remark 3.3. *That a* $\sigma-$*finite signed measure can be extended to a complete measure space* (λ, Σ, ν) *with* ν *a restriction of the outermeasure on* Σ *follows from elementary analysis results. In addition, while the traditional formulation assumes a symmetric distribution around the mean (for example* $N(\lambda, 1)$*), the current formulation allows far more flexibility. This is because, instead of fixing the variance of an unimodal symmetric distriution we may instead fix the value based on a probability as a function of* λ*. As such, the latent variable formulation can take a symmetric or asymmetric distributional form around 0, and thus the probabilities of success do not necessarily have to approach either 0 or 1 at the same rate.*

Remark 3.4. *That a finite signed measure can be extended to a complete measure space* $(\lambda, \Sigma, |\nu|)$ *with* $|\nu|$ *a restriction of the outermeasure on* Σ *follows from elementary analysis results. In addition, while the traditional formulation assumes a symmetric distribution around the mean (for example* $N(\lambda, 1)$*), the current formulation allows far more flexibility. This is because, instead of fixing the variance of an unimodal symmetric distriution we may instead fix the value based on a probability as a function of* λ*. As such, the latent variable formulation can take a symmetric or asymmetric distributional form around 0, and thus the probabilities of success do not necessarily have to approach either 0 or 1 at the same rate.*

In the forthcoming, I elaborate on the convergence properties of an estimation methodology for the binary outcome case. The extension to the continuous GLM formulation are also

briefly discussed. However, first I give some foundational results for the uniqueness of the link constraint.

Theorem 3.2. *Let (λ, Σ) be a measureable space. Then there is an unique solution to any link modification problem, where the link constraint holds with equality in the Generalized Linear Model Framework for some $\alpha^* \in R \setminus \{-\infty, \infty\}$, given $\hat{F}_i \notin \{0, 1\}$, $X \notin \{0, \infty, -\infty\}$ element wise for each $i \in \{1, ..., n\}$ and $\boldsymbol{\beta}_j \notin \{\infty, -\infty\}$ with $j \in \{1, ..., (k+1)\}$.*

Proof. *Case I: $(\lambda, \Sigma, |\nu|)$ is a finite measure space with $|\nu|$ finitely additive and countably monotone.*

First note that by construction,

$$\{|\nu|, \lambda(\mathbf{x}, \beta)\} < \infty \implies |\nu|^{\alpha^*} = \lambda(\mathbf{x}, \beta), \tag{3.52}$$

holds for some $\alpha^* \in R \setminus \{-\infty, \infty\}$ by the density of the reals since,

$$||\nu|| = sup \sum_{i=1}^{n} |\nu|(E_i) < \infty. \tag{3.53}$$

Thus, $||\nu||$ is of bounded variation and $|\nu|$ may be represented as the difference of two monotonic functions. As such, there exists a function $g \in R \setminus \{-\infty, \infty\}$ such that

$$g_{|\nu|} = |\nu|[I] = \hat{\mathbf{F}} \tag{3.54}$$

where I is any countable collection of measureable sets covering $R \setminus \{-\infty, \infty\}$. Such a covering exists from the compactness of the support on $|\nu|$ and the assertion follows.

Case II: (λ, Σ, ν) is a signed measure space.

Subcase A: ν is finite a.e. on Σ.

In this case we are back at *Case I* above and the results hold.

Subcase B: ν is not finite a.e. on Σ.

Consider the Caratheodory-Hahn extension to the measuareable space $(\lambda, \Sigma, |\bar{\nu}|)$. From Proposition 3.3 and Proposition 3.4, we know such an extension exists for λ lebesgue measureable. Consequently, using the results of *Case I* again we arrive at the desired conclusion. □

I now discuss the almost sure convergence property of this methodology.

Theorem 3.3. *Given $\alpha^* \in R \setminus \{-\infty, \infty\}$, and $\hat{F}_i \notin \{0, 1\}$, $X \notin \{0, \infty, -\infty\}$ elementwise for each $i \in \{1, ..., n\}$ and $\boldsymbol{\beta}_j \notin \{\infty, -\infty\}$ with $j \in \{1, ..., (J+1)\}$ subject to the link constraint holding for each observation,*

$$\hat{\beta} \xrightarrow{a.s.} \beta. \tag{3.55}$$

Proof. Consider an MCMC framework and the folllowing cases below.

Case I: Let (λ, Σ, ν) be a finite measure space with ν a signed measure.
First note that the link condition holding for each observation implies following Kass and Steffey (1989) we may write for the appropriate measureable space $(\lambda, \Sigma, |\nu|)$,

$$p(y^*, \alpha^*, \beta|y) \propto p(y|\beta) f(y^*|\alpha^*, \beta, y) f(\alpha^*|\beta) f(\beta) \tag{3.56}$$

as the posterior distribution. Therefore,

$$p(\beta|y) \propto \int_{y^*} \int_{\alpha^*} p(y|\beta) f(y^*|\alpha^*, \beta, y) f(\alpha^*|\beta) f(\beta). \tag{3.57}$$

Then[2] considering all observations we may write,

$$p(\hat{\beta}^{(j)}|y) \propto f_n\left(\lambda(\hat{\beta}^{(j)})|y^{(j)*}, \alpha^{(j)*}\right). \tag{3.58}$$

For brevity I denote $f_n\left(\lambda(\hat{\beta}^{(j)})|y^{(j)*}, \alpha^{(j)*}\right)$ as f_n^j going forward. Let T be an integral transform such that,

$$Tp(y|\beta)f(y^*|\alpha^*, \beta)f(\alpha^*|\beta)f(\beta) = p(\beta|y) \propto \int_{y^*}\int_{\alpha^*} p(y|\beta)f(y^*|\alpha^*, \beta)f(\alpha^*|\beta)f(\beta). \tag{3.59}$$

Clearly, T is a bounded linear operator in L_1 by construction (the L_p case will be considered shortly) and define,

$$f_n^{(j+1)} = (Tf_n^{(j)})(\beta^{(j)}). \tag{3.60}$$

Then, define

$$X_k = \{\lambda : g_n^{(j)} = |f_n^{(j+1)} - (Tf_n^{(j)})(\beta^{(j)})| > k\}. \tag{3.61}$$

I claim $g_n^{(j)}$ is an uniformly integrable sequence of functions w.r.t. $|\nu|$ over Σ. It is clear that by construction $|\nu|$ is finite over Σ. Thus, we may choose a natural number N such that if $k \geq N$ we have,

$$|\nu|(\cup_{k=N}^{\infty} E_k) < \frac{1}{N}, \tag{3.62}$$

for each k. Let $\tilde{N}$ be the maximum of these indicies such that for all k (3.62) holds. Further by the continuity of the measure we may choose a disjoint collection of such sets. Let $g_{nk}^{(j)}$ be the restriction of $g_n^{(j)}$ to E_k. We know it is finite over E_k, so by the simple approximation

[2]Following Tanner and Wong (1987) I assume that $\{y^*, \alpha^*\}$ both have compact support as the case for discrete support can be proved similarly.

lemma there is a simple function $g_{nk}^{(j)}$ such that,

$$\int_{E_k} g_n^{(j)} - \int_{E_k} g_{nk}^{(j)} < \epsilon/2. \tag{3.63}$$

But $g_n^{(j)} - g_{nk}^{(j)} \geq 0$ thus,

$$\int_{\Sigma} g_n^{(j)} - \int_{\Sigma} g_{nk}^{(j)} \leq \sum_{k=1}^{\infty} \int_{E_k} |g_n^{(j)} - \int_{E_k} g_{nk}^{(j)}| \leq \tilde{N}|\nu|(\Sigma) + \epsilon/2. \tag{3.64}$$

By letting $|\nu|(\Sigma) = \epsilon/(2\tilde{N})$ we get our desired result. Thus, $g_n^{(j)}$ is uniformly integrable. Further, by contruction the link condition holding for each observation implies $g_n^{(j)} \xrightarrow{p.w.} |\nu|$ for each j. Therefore, by the Vitali Convergence Theorem we have,

$$\lim_{j \to \infty} g_n^{(j)} = p(\beta|y). \tag{3.65}$$

Thus, we are done. □

This result has some ready extensions to the L^p spaces and the result below highlights one of those results.

Corollary 3.3. *Under the conditions of Theorem 3.3 we have that for* $1 \leq p < \infty$,

$$\{g_n\} \xrightarrow{a.s.} p(\beta|y). \tag{3.66}$$

Proof. This result follows from the Vitali L^p Convergence Criterion, from uniform integrability and finiteness. □

One of the more useful results of the methodology is that the latent variable nonparametric

distributional assumptions do not need any restrictions on the variance parameter. The next corollary puts this in more concrete terms. Thus, we can assert the following regarding the variance of the nonparametric latent variable formulation.

Corollary 3.4. *Given* $\alpha^* \in R \setminus \{-\infty, \infty\}$*, and* $X \notin \{0, \infty, -\infty\}$ *for each* $i \in \{1, ..., n\}$ *and* $\boldsymbol{\beta}_j \notin \{\infty, -\infty\}$ *with* $j \in \{1, ..., (k+1)\}$*, the variance of the latent variable distribution* y^* *need not be fixed for identification subject to the link constraint holding for each observation.*

Proof. The result follows naturally from the a.s. convergence property of β. Since the Hahn Decomposition is unique, a.s. convergence implies that the variance cannot be fixed for all p.w. convergent sequences of $g_n^{(j)}$ while guaranteeing uniqueness of the mutually singular measures. □

3.2.4 Nonparametric Latent Adaptive Hierarchical EM Like (LAHEML) Algorithm

The methodology described above is extremely versatile as pointwise convergence is ensured through the link constraint holding for each observation, which in concert with a data augmentation framework in the latent Bayesian formulation can be used with MCMC to give almost sure convergence, under very general conditions. Accordingly, I refer to the methodology as a Latent Hierarchical Adaptive EM Like Algorithm (LAHEML), which is nevertheless more general than the EM algorithm. It further has the advantage of being able to be used for both model selection and MIP comparisons concurrently. However, the estimation process can be rather involved. Therefore, this section outlines detailed general algorithms, one in the unpenalized case and the other for the penalized model selection application for a particular model formulation in the regression framework. Thus, below I first outline the unpenalized algorithm in 3.2.4.1, and the penalized version is outlined in 3.2.4.2.

3.2.4.1 Unpenalized Application of LAHEML

A particular model formulation of the nonparametric methodology in the usual linear regression framework is given below.

1. Please note, the posterior is given by

$$p(\boldsymbol{\beta}|\boldsymbol{x}) \propto \int_{\boldsymbol{y^*}} \int_{\boldsymbol{\alpha^*}} L(\boldsymbol{y^*}, \boldsymbol{x}|\boldsymbol{\beta}) g(\alpha^*|\boldsymbol{\beta}) f(\boldsymbol{\beta}), \tag{3.67}$$

   thus, perform an MH step to optimize the posterior for the current value (jth value) of $\boldsymbol{\beta}^{(j)}$.

2. To draw from the latent variable y^* instead of running a parametric normal or Logistic regression on the latent variable, I propose a nonparametric regression, such that,

$$E(y_i|\boldsymbol{x}_i = \boldsymbol{x}) = m(\boldsymbol{x}). \tag{3.68}$$

   From which we get the distribution of $\boldsymbol{y}^{(j)*}$ by

$$\hat{\boldsymbol{F}}(\boldsymbol{y}^{(j)*}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}^{(j-1)}) = \sum_{x=-k}^{k} \frac{\sum_{i=1}^{n} 1(|\boldsymbol{x_i} - x|\boldsymbol{\beta}^{(j-1)} \leq h) y_i}{\sum_{i=1}^{n} 1(|\boldsymbol{x_i} - x|\boldsymbol{\beta}^{(j-1)} \leq h)}, \tag{3.69}$$

   and drawing from $\{X, \boldsymbol{\beta} : X\boldsymbol{\beta}^{(j)} \in (-\infty, \int \hat{F}(\boldsymbol{y^*}|\boldsymbol{y}, X) = \kappa] \}$ if $y_i = 0$ and from $\{X, \boldsymbol{\beta} : X\beta \in (\int \hat{F}(y^*|\boldsymbol{y}, X) = \kappa, \infty) \}$ if $y_i = 1$ to get $y_i^{(j)*}$ for each observation $(\boldsymbol{y}^{(j)*})$. Of course, the binary values for y_i create no issues as the distribution of failure probabilities is simply the success probabilities subtracted from 1. Further, we can also use continuous kernel estimates here.

3. Thus, we can compute a similar nonparametric continuous distribution on $\boldsymbol{\epsilon}^{(j)*} = \boldsymbol{y}^{(j)*} - X\boldsymbol{\beta}^{(j)}$ to get the current draws of the probability of success.

4. We can then compute the numerical estimates of $\boldsymbol{\alpha}^{(j)*}$, making any transformations as necessary for the link condition as discussed above in

$$\left(\hat{\boldsymbol{F}}(\boldsymbol{\epsilon}^{(j)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)})\right)^{\boldsymbol{\alpha}^{(j)*}} = X\boldsymbol{\beta}^{(j)}, \tag{3.70}$$

by solving

$$\left(\left(\hat{\boldsymbol{F}}(\boldsymbol{\epsilon}^{(j)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)})\right)^{\boldsymbol{\alpha}^{(j)*}} - X\boldsymbol{\beta}^{(j)}\right)^{d} = 0, \tag{3.71}$$

for some $d \in \{1, 2, ...\}$.

5. Compute the nonparametric density estimate of

$$\hat{\boldsymbol{g}}(\boldsymbol{\alpha}^{(j)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)}) = \sum_{\tau=-k}^{k} \frac{\sum_{i=1}^{n} 1(|\boldsymbol{\alpha}_i^{(j)} - \tau| \leq h)}{\sum_{i=1}^{n} 1(\boldsymbol{\alpha}_i^{(j)} - \tau \leq h)} \tag{3.72}$$

where $\{\tau\} \in R \setminus \{-\infty, \infty\}$. Or we can do a kernel density estimation here as well for $\hat{\boldsymbol{g}}(\boldsymbol{\alpha}^{(j)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)})$.

6. Now we can treat $\boldsymbol{\alpha}^{(j)*}$ as a latent variable itself and can randomly draw from it if $\boldsymbol{y} = \mathbf{1}$ such that $\hat{\boldsymbol{g}}(\boldsymbol{\alpha}^{(j+1)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)}) \geq \kappa$ and from $\hat{\boldsymbol{g}}(\boldsymbol{\alpha}^{(j+1)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)}) < \kappa$ if $\boldsymbol{y} = \mathbf{0}$ for every observation to get our estimates of $\boldsymbol{\alpha}^{(j+1)*}$ for the next iteration.

7. Go to step 1, repeat and iterate to convergence.

The construct above ensures that $\{\boldsymbol{\beta}^{(j)}\}$ converges to its true distributions given the data $\boldsymbol{y}$. Since $\boldsymbol{\alpha}^{*}$ is a function of $\boldsymbol{\beta}$, the convergence results hold for its distribution as well. The bias of the nonparametric density estimates are corrected by ensuring the link condition holds for each observation.

It is worthwhile to consider that the cutoff points of $\kappa \in (0, 1)$ a probability, can also be a parameter to be estimated here. Since the Jordan decomposition remains valid, such a

model specification of the methodology should be especially relevant for asymmetric DGPs. This is pursued in the nonparametric simulation datasets, where the cutoff was based on the observed distributions of successes and failures and not necessarily fixed at the median for $f(\alpha^*|\boldsymbol{\beta}, \boldsymbol{y}^*, \boldsymbol{y})$. The results of the simulation studies can be found in 3.3.

3.2.4.2 Penalized Application of LAHEML

For a penalized application I consider The Bayesian Adaptive Lasso as in Leng et al. (2014). This penalized version of the methodology is contingent on a different prior specification than described above. The prior is given by,

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{k=1}^{p} \frac{\lambda_k}{2\sqrt{\sigma^2}} e^{\lambda_k|\beta_k|/\sqrt{\sigma^2}}. \tag{3.73}$$

For the current application we can move forward with either a Hierarchical formulation or an empirical bayes application as given in Leng et al. [Ibid] and I follow a Hierarchical methodology accordingly. This is because it also requires the estimation of the prior hyper-parameters on the shrinkage parameters, which can no longer be considered uninformative. Leng et al. (2014) consider a gamma prior on the λ_k's and I also follow this same formulation. Thus, the prior on the shrinkage parameters can be given by

$$\pi(\lambda_k^2) = \frac{\delta^r}{\Gamma(r)} (\lambda_k^2)^{r-1} e^{-\delta\lambda_k^2}. \tag{3.74}$$

Further note that Lehmann and Casella (2006), point to the parameters deeper in the hierarchy having less of an impact on the estimation process. As such, for the present application I set the δ and r hyperparameters both equal to some small number such as 0.1. Accordingly the penalized estimation algorithm is given below.

1. Please note, the posterior is given by

$$p(\boldsymbol{\beta}|\boldsymbol{x}) \propto \int_{\boldsymbol{y^*}} \int_{\boldsymbol{\alpha^*}} L(\boldsymbol{y^*}, \boldsymbol{x}|\boldsymbol{\beta}) g(\alpha^*|\boldsymbol{\beta}) f(\boldsymbol{\beta}), \tag{3.75}$$

   thus, perform an MH step to optimize the posterior for the current value (jth value) of $\boldsymbol{\beta}^{(j)}$.

2. To draw from the latent variable y^* instead of running a parametric normal or Logistic regression on the latent variable, I propose a nonparametric regression, such that,

$$E(y_i|\boldsymbol{x}_i = \boldsymbol{x}) = m(\boldsymbol{x}). \tag{3.76}$$

   From which we get the distribution of $\boldsymbol{y}^{(j)*}$ by

$$\hat{\boldsymbol{F}}(\boldsymbol{y}^{(j)*}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}^{(j-1)}) = \sum_{x=-k}^{k} \frac{\sum_{i=1}^{n} 1(|\boldsymbol{x_i} - x|\boldsymbol{\beta}^{(j-1)} \leq h) y_i}{\sum_{i=1}^{n} 1(|\boldsymbol{x_i} - x|\boldsymbol{\beta}^{(j-1)} \leq h)}, \tag{3.77}$$

   and drawing from $\{X, \boldsymbol{\beta} : X\boldsymbol{\beta}^{(j)} \in (-\infty, \int \hat{F}(\boldsymbol{y^*}|\boldsymbol{y}, X) = \kappa] \}$ if $y_i = 0$ and from $\{X, \boldsymbol{\beta} : X\beta \in (\int \hat{F}(y^*|\boldsymbol{y}, X) = \kappa, \infty) \}$ if $y_i = 1$ to get $y_i^{(j)*}$ for each observation $(\boldsymbol{y}^{(j)*})$. Of course, the binary values for y_i create no issues as the distribution of failure probabilities is simply the success probabilities subtracted from 1. Further, we can also use continuous kernel estimates here.

3. Thus, we can compute a similar nonparametric continuous distribution on $\boldsymbol{\epsilon}^{(j)*} = \boldsymbol{y}^{(j)*} - X\boldsymbol{\beta}^{(j)}$ to get the current draws of the probability of success.

4. We can then compute the numerical estimates of $\boldsymbol{\alpha}^{(j)*}$, making any transformations as necessary for the link condition as discussed above in

$$\left(\hat{\boldsymbol{F}}(\boldsymbol{\epsilon}^{(j)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)})\right)^{\boldsymbol{\alpha}^{(j)*}} = X\boldsymbol{\beta}^{(j)}, \tag{3.78}$$

by solving

$$\left(\left(\hat{\boldsymbol{F}}(\boldsymbol{\epsilon}^{(j)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)})\right)^{\boldsymbol{\alpha}^{(j)*}} - X\boldsymbol{\beta}^{(j)}\right)^{d} = 0, \tag{3.79}$$

for some $d \in \{1, 2, ...\}$.

5. Compute the nonparametric density estimate of

$$\hat{\boldsymbol{g}}(\boldsymbol{\alpha}^{(j)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)}) = \sum_{\tau=-k}^{k} \frac{\sum_{i=1}^{n} 1(|\boldsymbol{\alpha}_i^{(j)} - \tau| \leq h)}{\sum_{i=1}^{n} 1(\boldsymbol{\alpha}_i^{(j)} - \tau \leq h)} \tag{3.80}$$

where $\{\tau\} \in R \setminus \{-\infty, \infty\}$. Or we can do a kernel density estimation here as well for $\hat{\boldsymbol{g}}(\boldsymbol{\alpha}^{(j)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)})$.

6. Now we can treat $\boldsymbol{\alpha}^{(j)*}$ as a latent variable itself and can randomly draw from it if $\boldsymbol{y} = \mathbf{1}$ such that $\hat{\boldsymbol{g}}(\boldsymbol{\alpha}^{(j+1)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)}) \geq \kappa$ and from $\hat{\boldsymbol{g}}(\boldsymbol{\alpha}^{(j+1)*}|\boldsymbol{y}^{(j)*}, X, \boldsymbol{\beta}^{(j)}) < \kappa$ if $\boldsymbol{y} = \mathbf{0}$ for every observation to get our estimates of $\boldsymbol{\alpha}^{(j+1)*}$ for the next iteration.

7. Perform an MH step to compute the value of

$$(\lambda_k^{(j+1)}|\boldsymbol{\alpha}^{(j+1)*}, \boldsymbol{\beta}^{(j+1)}, \boldsymbol{y}^{(j)*}), \tag{3.81}$$

from a suitable candidate density such as the t-distribution with 10 degrees of freedom.

8. Go to step 1, repeat and iterate to convergence.

3.2.5 Asymptotic Model Diagnostics

One of the more useful outcomes of the proposed model in both the nonparametric and parametric implementations is that it simply adds one extra parameter to be estimated. In the parametric case, since we know $\mathrm{E}(\alpha^*|\beta)$ for existing models such as the Logit ($\alpha^* = 1$ in the parametric case), we can use large-sample results under i.i.d. assumptions to test the hypothesis that our model results vary from traditional GLM model fits. Indeed the nonparametric methodology is even more useful in this regard. Consider that if $y = 1$ we have the specified link condition implies,

$$F(\boldsymbol{\beta}|y) = \lambda(X, \boldsymbol{\beta}). \tag{3.82}$$

This implies that if we parametrically assume a distribution for $\hat{F}$ (which we know converges to the true F) such as the Logistic or Probit distribution and input the estimated $\hat{\boldsymbol{\beta}}$'s into the functional specification and calculate the divergence of $\hat{\boldsymbol{F}}(\hat{\boldsymbol{\beta}}|\boldsymbol{y})$ from 3.82. For example, we may compute the value of α^*, say $\bar{\alpha^*}$ that minimizes,

$$\left(\hat{\boldsymbol{F}}(\boldsymbol{\beta}|\boldsymbol{y})^{\boldsymbol{\alpha}^*} - \boldsymbol{\lambda}(X, \boldsymbol{\beta})\right). \tag{3.83}$$

In particular, we know for GLM, if the convergence has occured to the true distribution then α^* should equal 1. While the X's are held fixed, $\bar{\alpha^*}$ is both unbiased, consistent and asymptotically normal by the central limit theorem and i.i.d. assumptions, as long as $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically unbiased. The accompanying proofs ensure that this is the case. Accordingly, given $\bar{\boldsymbol{\alpha}^*}$, we can thus estimate the aymptotically unbiased and consistent estimates of the variance of α^* as well to get,

$$\alpha^* \sim N(1, E(E(\alpha^*|\boldsymbol{\beta}^*) - 1|\boldsymbol{\beta}^*)^2), \tag{3.84}$$

$$\implies \hat{\bar{\alpha}}^* \overset{asymp.}{\sim} N\left(1, \frac{\sum_{i=1}^{n}(\alpha_i - 1)^2}{n-1}\right). \tag{3.85}$$

$\boldsymbol{\beta}^*$ above represents the optimized estimated value. Thus, we can check our hypothesis that $\bar{\alpha}^* = 1$ for any particular parametric specification on the probability of success.

---

1. Perform a t-test on $\hat{\bar{\alpha}}^*$, with the appropriate null hypothesis values, and accept/reject model fit assumptions.

2. Thus,

   (a) Under rejection, the existing GLM is not adequate given assumptions on the model specification and the proposed model should be used.

   (b) Otherwise, the existing GLM is adequate and it can be used for model fit, inference and prediction (classification) accordingly[3].

This framework can similarly be extended to the likelihood ratio test, under the appropriate null values.

[3]Note however, that model fit, prediction and inference criteria should be evaluated on a wholistic basis to arrive at a choosen model even if the null hypothesis is not rejected.

3.2.6 Asymptotic Distribution of Adjusted ROC-Statistic

In order to analyze adequacy of classification performance, there are many existing statistics such as the Receiver-Operating Curve. Here I consider Adjusted ROC-Statistic (ARS) instead, based on Chowdhury (2019), as not only does it allow for interpretable estimates, it also has well known closed form large sample distributions. This allows at least two advantages over existing statistics. First, ARS can be represented as a simple interpretable ratio of observed classification outcomes, without the need for a likelihood. Second, the classification performance of any two models can be tested to see if they differ statistically. Therefore, the mathematician can employ bootstrap or other methods to compare the performance difference between models. To aid in the discussion, the confusion matrix is reproduced below.

Table 3.1: Confusion Matrix.

		Fitted Model Prediction	
		Success	Failure
True Classification in Data	Success	True Positive (TP, n_{11})	False Negative (FN, n_{10})
	Failure	False Positive (FP, n_{01})	True Negative (TN, n_{00})

Further let, G = Ground True, $S(t)$ = Fitted Prediction Subject to Some Parameter t, D = Entire Dataset, then we may define the following quantities.

$$True\ Positive = \frac{|S(t) \bigcap G|}{|G|}, \tag{3.86}$$

$$True\ Negative = \frac{|\neg S(t) \bigcap \neg G|}{|\neg G|}, \tag{3.87}$$

$$False\ Positive = \frac{|S(t) - G|}{|D - G|}, \tag{3.88}$$

$$False\ Negative = \frac{|\neg S(t) - G|}{|G|}. \tag{3.89}$$

Then,

$$ARS = \frac{\frac{|S(t)-G|}{|D-G|} + \frac{|\neg S(t)-G|}{|G|}}{\frac{|S(t)\bigcap G|}{|G|} + \frac{|\neg S(t)\bigcap \neg G|}{|\neg G|}} \tag{3.90}$$

or simply the ratio of incorrectly identified vs. correctly identified elements according to the model fitted. Define,

$$A = \{x : Fitted\ Model\ Correctly\ Identifies\ Observed\ Sample\}. \tag{3.91}$$

$$B = \{x : Fitted\ Model\ Incorrectly\ Identifies\ Observed\ Sample\}. \tag{3.92}$$

Then from elementary probability theory, where, the entire sample space, S = {A, B} and

$$P(A) + P(B) = P(S) = 1. \tag{3.93}$$

Then remarkably,

$$P(ARS) = P\left(\frac{P(B)}{1 - P(B)}\right) \tag{3.94}$$

is nothing but the odds ratio of the probability of incorrectly identifying observed sample divided by probability of correctly identifying the observed sample! This then has an asymptotic distribution (Bland, Martin J. et al. 2020) given by

$$log(ARS) \sim N(log(Oddsratio), \sigma^2),\ where\ \sigma = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}} \tag{3.95}$$

where n_{ij} = counts within each cell in table 3.2.6 . Note here that it is entirely plausible that one or more of the cells will be 0. Thus, to avoid dividing by 0, I recommend including some small $\epsilon \neq 0$ for inference. Another approach could be to impose each cell being at least 1 for identifiability.

3.2.6.1 Inference

Clearly, $ARS \in [0, \infty)$. Then using a slight aberration of the usual hypothesis testing procedure, let

$$H_0 = \text{Incorrect and correct identification are equally likely.}$$
$$H_A = \text{Incorrect and correct identification are not equally likely.}$$

If incorrect and correct identification are equally likely, then the test statistic becomes,

$$\frac{\sqrt{n}\ log(ARS)}{\sigma}) \sim N(0,1). \tag{3.96}$$

This framework can be used in a two sample t-test as well to compare any two models fitted to the data. With multiple models, the test can be expanded accordingly. As an example, when the variances for the log-odds attained for ARS for two different samples are assumed to be the same we can do a pooled t-test,

$$\text{Test Statistic} = \frac{\bar{\kappa_1} - \bar{\kappa_2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}, \tag{3.97}$$

where the subscripts indicate each estimate under the relevant model specifications. If the

two models are deemed to be dependent in some manner a matched pair test can be similarly done above. Furthermore, multiple models can be compared using the Wald test.

3.2.6.2 A Likelihood Ratio Based Test Statistic

Since the Wald test has a number of weaknesses (see for example, King and Goh 2002), I also detail a likelihood ratio test based on ARS in this section. Suppose that an asymptotic test is determined to be unsuitable by the mathematician. Consider first that for any model fitted to binary data, we must specify a cutoff point, such that if the fitted probability is greater than this value, the fitted outcome is $\hat{y}_i = 1$ and 0 otherwise. Therefore, our model is in fact, a function of not only our chosen functional form F(.), but also this cutoff. Let this cutoff be κ (usually held at 0.5 as is customary according to the literature Greene 2003). Then for all distributions, symmetric or otherwise,

$$Pr(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}, \kappa) = 1 - F(-\boldsymbol{x}_i'\boldsymbol{\beta}) \; s.t. \; \{1 - F(-\boldsymbol{x}_i'\boldsymbol{\beta}) \geq \kappa\}. \tag{3.98}$$

Thus,

$$Pr(y_i = 1|\boldsymbol{x}_i, \kappa, 1 - F(-\boldsymbol{x}_i'\boldsymbol{\beta}) \geq \kappa) = \frac{1 - \int_{\{x:\lambda \geq x^*(\kappa)\}}^{\infty} F(-\boldsymbol{x}_i' \times \boldsymbol{\beta})dx}{(1 - F(-X^*(\kappa)\boldsymbol{\beta}))}. \tag{3.99}$$

Under this general formulation using LAHEML κ can be chosen by the scientist based on how they want to prioritize Type-I or Type-II error. Thus, if the null is: $H_0 : \hat{\boldsymbol{\beta}} = \boldsymbol{\beta_0}$ a

likelihood ratio statistic (LRT) can be given by,

$$\tilde{\lambda} = \frac{L(\hat{\boldsymbol{\beta}}_{\mathbf{0}}|X)}{L(\hat{\boldsymbol{\beta}}|X)}. \tag{3.100}$$

Here $\tilde{\lambda}$ represents the likelihood-ratio statistic and should not be confused with the link constraint condition.

3.2.7 Semiparametric Estimation of ARS

A fairly simple semi-parametric estimation procedure of the asymptotic distribution of ARS can also be given using bootstrap. A tentative algorithm for this is given below.

- Draw randomized $\hat{\boldsymbol{\beta}}$ from estimated $\boldsymbol{\beta}$s.
- Recalculate the bias corrected ARS.
- Find the bootstrapped distribution of ARS.
- Calculate the 95% credible interval of the bias corrected bootstrapped distribution of ARS.
- Calculate ARS on a holdout set using estimated $\hat{\boldsymbol{\beta}}$.
- If calculated ARS on the holdout set does lie in the 95% credible interval, reject the null that the observed values and the fitted probabilities are independent.

3.3 Monte Carlo Simulation

In order to validate the methodology and the mathematical results, extensive simulation studies were done for both the penalized and unpenalized versions. In particular, to validate the robustness of the Proposed Nonparametric methodology a Bayesian framework is used for the simulation studies on various DGP's, both symmetric (Logit and Probit) and asymmetric (Complementary Log-Log). For this purpose, datasets were generated from the standard normal distribution for different sample sizes
($n = \{100, 500, 1000, 2000\}$) and models,

$$\boldsymbol{y} = Intercept + X_1 + X_2, \tag{3.101}$$

$$\boldsymbol{y} = Intercept + X_1 + exp(X_2), \tag{3.102}$$

$$\boldsymbol{y} = Intercept + exp(X_1) + sin(X_2). \tag{3.103}$$

The different model specifications are needed to understand the performance of the proposed model when the data are linear, non-linear, or a mixed specification in the X's. All datasets had 3 parameters to estimate, for the intercept (β_1) and for two explanatory or independent variables drawn from the standard normal ($\{\beta_2, \beta_3\}$), with the appropriate transformations indicated above. Then for known β values, a Probit, Logit, or a Complementary Log-Log DGP was used to generate outcomes (dependent variable $\boldsymbol{y}$) which varied in the number of 1's that were present.

In particular, the known $\{X, \boldsymbol{\beta}\}$ values along with each functional form above can be used to calculate the probability of each observation for each specific model. Thus, we can consider the calculated $\boldsymbol{y}$ values along with the generated X's as the data on which we can fit our chosen statistical models for each DGP. We may then evaluate the performance of the

proposed model against other popular existing baseline models.

Finally, another step was done to create datasets which had different numbers of successes as opposed to failures. Thus, the unbalancedness of the data were varied between $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Here, 0.5 indicates equal number of successes and failures (balanced), 0.4 indicates 10% fewer successes than failures and so forth. Accordingly, for each sample size there are five different unbalanced datasets, each of which has three parameters or β's to estimate for each of the three DGPs for each of the models specified (linear, non-linear or mixed). As such, for each sample size, there are 60 different datasets, each with 3 parameters to estimate, for a total of $180 \times 3 = 540$ parameters to estimate, compare and contrast[4].

The results are extremely encouraging for both the penalized and unpenalized applications. The Proposed Nonparametric methodology not only outperforms existing models (including the Proposed Parametric methodology) in inference, but also in classification in regards to ARS. Indeed, the classification results even outperform ANN, on average over all of the many different datasets considered. The results below also show that the near perfect coverage results were attained with a smaller confidence interval than the Penalized Logistic. Two separate simulation runs were done, one in an unpenalized and the other in a penalized formulation. The summaries are given below.

3.3.1 Unpenalized Application Results

In this section I first consider the unpenalized results, and the summaries can be found in Table 3.2, Table 3.3, and Table 3.4 below. The unpenalized application almost uniformly contained the true parameters more often, and thus had better coverage. It was able to attain this by having confidence interval ranges which were smaller than the Penalized Logistic, which had the worst coverage performance among the methods compared here. The Proposed

[4]Note also that by construction, we know what the true β's are, and therefore, can use these true values to understand the performance of each of the models fitted to each dataset.

Nonparametric application in comparison to the Proposed (Parametric) Logistic, Bayesian Probit and Neural Net methods, uniformly outperformed them. It did so without considering functional specifications which lack scientific interpretability as in Neural Networks. This is because deeper networks with more complex basis expansions can lead to scientifically uninterpretable models, at the cost of better classification outcomes.

Accordingly, for Neural Net, since not all model specification and layers could always be fitted, a range of between two to five layers were considered with two neurons in each layer. While a more complicated model could have been used for comparison, the same could be said for the Proposed Nonparametric and Proposed Logistic methods as well. As such, following Chowdhury (2021a) to keep model performance comparable, more complicated model formulations were not deemed appropriate for the NN. In addition, Logistic and Penalized Logistic formulations were not considered for the classification performance comparison given that Chowdhury [Ibid] and Chowdhury (2021d) show that they are outperformed by the other methodologies compared.

Table 3.2: Simulation Coverage (in Percentage) Summary for Proposed Unpenalized DGPs (at 1% Significance Level)

	Proposed Nonpara.	Prop. Para.	Bayesian Probit	Penalized Logistic
LGR Covr. (NL)	98.33%	95.00%	83.33%	51.67%
PR Covr. (NL)	100.00%	95.00%	75.00%	46.67%
Comp. Lg. Covr. (NL)	100.00%	93.33%	80.00%	56.67%
LGR Covr. (Mx.)	98.33%	91.67%	71.67%	61.11%
PR Covr. (Mx.)	100.00%	95.00%	75.00%	25.00%
Comp. Lg. Covr. (Mx.)	100.00%	96.67%	81.67%	20.00%
LGR Covr. (L)	100.00%	96.67%	85.00%	63.33%
PR Covr. (L)	98.33%	96.67%	80.00%	66.67%
Comp. Lg. Covr. (L)	98.33%	96.67%	81.67%	66.67%

Note: This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000, 2000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 20 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 60 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 180 parameters per DGP for a total of 540 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets.

Table 3.3: Simulation Confidence Interval Range Summary for All DGPs (at 1% Significance Level)

	Penalized Logistic	Proposed Nonpara	Proposed Logistic	Bayesian Probit
LGR Covr. (NL)	6.47	5.66	5.37	2.00
PR Covr. (NL)	7.44	5.42	5.20	1.77
Comp. Lg. Covr. (NL)	7.77	5.65	4.89	1.88
LGR Covr. (Mx.)	7.66	5.75	5.40	2.07
PR Covr. (Mx.)	3.94	5.64	5.12	1.87
Comp. Lg. Covr. (Mx.)	2.27	6.12	4.93	1.84
LGR Covr. (L)	7.12	5.90	4.73	1.75
PR Covr. (L)	7.45	5.77	5.15	1.92
Comp. Lg. Covr. (L)	6.96	5.73	4.81	1.66

Note: This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000, 2000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 20 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 60 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 180 parameters per DGP for a total of 540 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets.

Table 3.4: Simulation Summary of ARS for All DGPs

	Proposed Nonpara.	Proposed Logistic	Bayesian Probit	Neural Net
Non-Linear	0.07	0.22	0.21	0.19
Mixed	0.11	0.22	0.22	0.22
Linear	0.08	0.19	0.23	0.20

Note: This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000, 2000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 20 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 60 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 180 parameters per DGP for a total of 540 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets.

3.3.2 Penalized Application Results

For the penalized application all aspects of the unpenalized version were retained with the addition of an extra explanatory variable drawn at random from the standard normal distribution. All other aspects of the simulation including the number of different models, observation numbers and unbalancedness criteria were all kept consistent for comparison across the penalized and unpenalized versions.

The results were again extremely encouraging, and consistent with the results from the previous Section. They can be found below in Table 3.5, Table 3.6, and Table 3.7. The Proposed Penalized Nonparametric methodology not only outperforms existing models (including the parametric methodology) in inference, but also in classification in regards to ARS. Indeed, the classification results even outperforms the unpenalized application and Neural Net, on average over all of the many different datasets considered. The inference results below again show that the near perfect coverage results were attained with smaller confidence intervals than the Penalized Logistic. The other aspects of the simulation from the unpenalized case in regard to Neural Net model specification were kept the same. However, in the current application, since an extra nuisance variable was considered, the Penalized Logistic model was also considered for comparison.

In summary, the results are indicative of the efficiency of the methodology and the mathematical results. The Proposed Nonparametric application contained the true parameters more often than the parametric application, which in turn was more efficient than the other existing methodologies. It did so while having smaller confidence intervals than the Penalized Logistic application. This same superior performance also extended to classification. While the Proposed Parametric application and the existing Bayesian Latent Probit gave classification accuracy similar to Neural Nets, the Proposed Nonparametric applications almost uniformly outperformed all other methodologies on average and were statistically significant

Table 3.5: Penalized Simulation Coverage Summary All DGPs (at 1% Significance Level)

	Unpenalized Nonpara.	Penalized Nonpara.	Prop. Para.	Bayesian Probit	Penalized Logistic
LGR Covr. (NL)	100.00%	100.00%	98.75%	92.50%	70.00%
PR Covr. (NL)	100.00%	98.68%	97.37%	84.21%	65.79%
Comp. Lg. Covr. (NL)	100.00%	98.75%	98.75%	81.25%	65.00%
LGR Covr. (Mx.)	100.00%	100.00%	100.00%	81.94%	70.83%
PR Covr. (Mx.)	97.22%	97.22%	94.44%	81.94%	70.83%
Comp. Lg. Covr. (Mx.)	100.00%	100.00%	96.67%	81.67%	72.06%
LGR Covr. (L)	98.75%	100.00%	100.00%	88.75%	75.00%
PR Covr. (L)	96.25%	95.00%	97.50%	80.00%	73.75%
Comp. Lg. Covr. (L)	100.00%	100.00%	100.00%	87.50%	75.00%

Note: This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000, 2000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 20 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 60 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 180 parameters per DGP for a total of 540 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets.

Table 3.6: Penalized Simulation Confidence Interval Summary for All DGPs (at 1% Significance Level)

	Unpenalized Nonpara.	Penalized Nonpara.	Proposed Logistic	Bayesian Probit	Penalized Logistic
LGR Covr. (NL)	5.96	5.77	5.52	1.87	5.64
PR Covr. (NL)	6.04	5.60	4.93	1.99	5.81
Comp. Lg. Covr. (NL)	5.88	5.43	5.63	2.03	6.18
LGR Covr. (Mx.)	5.64	5.87	5.67	1.80	6.22
PR Covr. (Mx.)	5.67	5.78	5.22	1.97	6.03
Comp. Lg. Covr. (Mx.)	5.97	5.44	5.44	1.97	6.07
LGR Covr. (L)	5.74	5.66	5.67	1.78	6.25
PR Covr. (L)	5.53	5.15	4.86	1.96	6.32
Comp. Lg. Covr. (L)	5.74	5.97	5.60	1.94	6.46

Note: This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000, 2000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 20 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 60 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 180 parameters per DGP for a total of 540 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets.

in the outperformance. Using these encouraging results I now apply the methodology to real-world dataset applications below and compare its performance to Random Forests, and deep neural networks.

Table 3.7: Summary of ARS for All DGPs Compared

	Unpenalized Nonpara.	Penalized Nonpara	Proposed Logistic	Bayesian Probit	Neural Net
Non-Linear	0.07	0.07	0.23	0.25	0.16
Mixed	0.17	0.07	0.23	0.27	0.21
Linear	0.08	0.05	0.19	0.23	0.21

Note: This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000, 2000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 20 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 60 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 180 parameters per DGP for a total of 540 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets.

3.4 Empirical Application

I make several empirical applications of the methodology discussed above. The first application is a biomedical one, where we seek to identify intoxicated individuals, based on phone accelerometer data, and the second is an application to identify exotic particles in high-energy Physics. They are detailed below.

3.4.1 Detecting Heavy Drinking Events Using Smartphone Data

To illustrate the efficacy of the model, we apply a simple model specification using its almost sure convergence property, to detect heavy drinking events using smartphone accelerometer data in Killian et al. (2019). Given the time series nature of the data the authors identified heavy drinking events within a four second window of their measured variable of Transdermal Alcohol Content (TAC) after various smoothing analyses on the accelerometer data. Their best classifier was a Random Forest with about 77.50% accuracy. A similar analysis was done on a far simpler model of TAC readings against the accelerometer reading predictors,

for all subject's phone placement in 3D space, for the x, y and z axes,

$$TAC = Intercept + x - axis\ reading + y - axis\ reading + z - axis\ reading. \quad (3.104)$$

TAC here was simply set to 1 if the measurement was over 0.08 and 0 otherwise. The same four second time window of accelerometer readings were used in the analysis with the assumption that the TAC readings were unlikely to change in such a small time interval. The results were extremely encouraging, with TeD (20% of the data) ARS classification accuracy of nearly 100.00%, with just $1,000$ iterations and 500 burn-in period, using some of the methodological contributions in Chowdhury (2021b) and Chowdhury (2021d) (the relevant plots can be found below in Figure 3.1 and Figure 3.2). In fact, the strength of

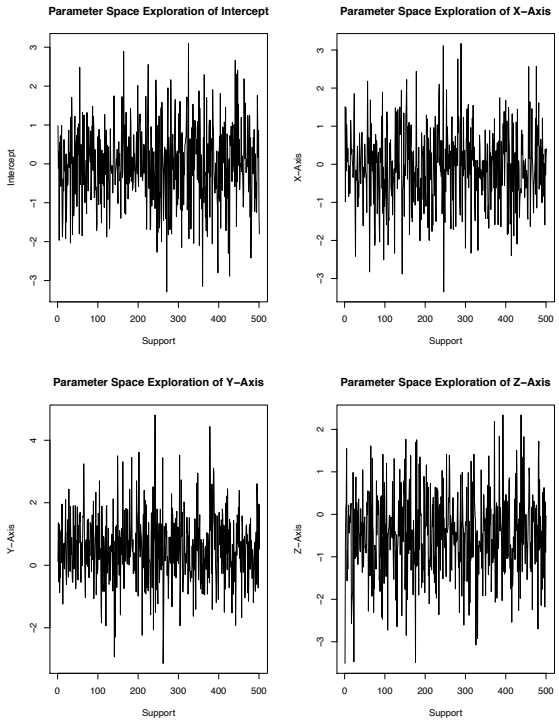


Figure 3.1: Sample Heavy Drinking Event Data Sample Space Exploration Plot for Nonparametric Methodology.

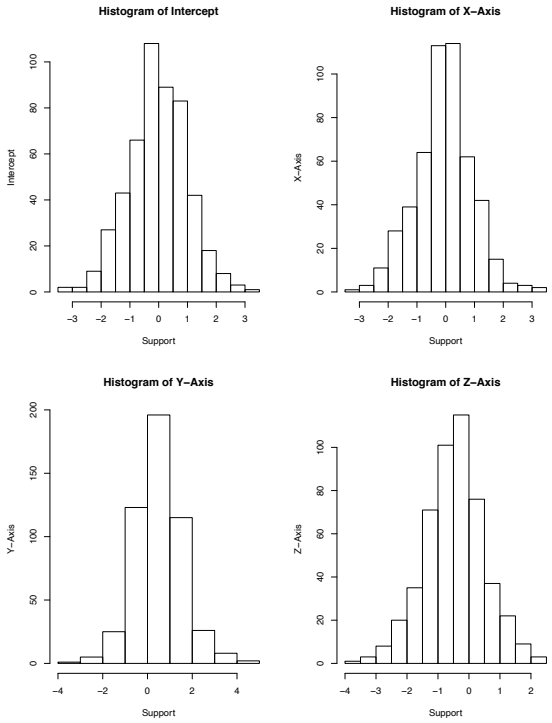


Figure 3.2: Sample Heavy Drinking Event Data Histogram of Parameters for Nonparametric Methodology.

the methodology may also allow us to perform model fit and model selection at the same time! To illustrate, a penalized methodology was applied using Adaptive Bayesian Lasso

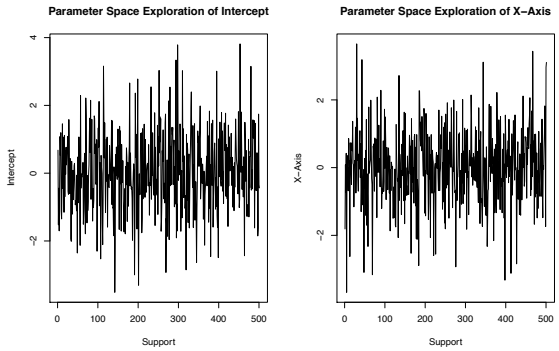


Figure 3.3: Sample Heavy Drinking Event Data Sample Space Exploration Plot for Nonparametric Penalized Methodology.

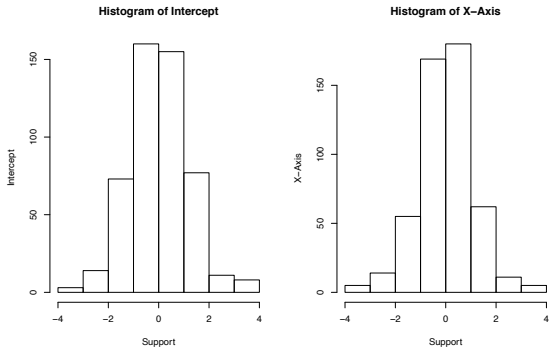


Figure 3.4: Sample Heavy Drinking Event Data Histogram of Parameters for Nonparametric Penalized Methodology.

(Leng et al. (2014)) in a Hierarchical framework on the same dataset. Contrary to the accepted norm that we cannot perform model fit and model selection at the same time, the TeD had perfect predictive accuracy (the relevant plots for the methodology are given in Figure 3.3 and Figure 3.4). However, it did have a slightly worse predictive performance in TrD (0.34 vs. 0.30). These new findings are "significant" in that they challenge and extend our discussion on scientific and statistical significance considerably. Accordingly, in the forthcoming discussion section I detail more of these advantages and disadvantages.

Table 3.8: Heavy Drinking Event Detection ARS Summary

	Unpenalized Nonpara. (TrD)	Unpenalized Nonpara. (TeD)	Penalized Nonpara. (TrD)	Penalized Nonpara. (TeD)
ARS	0.30	0.00	0.34	0.00

Note: Unpenalized Nonpara. (TrD) is the unpenalized application on the training data, and Unpenalized Nonpara. (TeD) is the unpenalized application on the test data (80% of the observations). Penalized Nonpara. (TrD) is the penalized application on the training data, and Penalized Nonpara. (TeD) is the penalized application on the test data (20% of the observations).

One further advantage of the methodology is that it allows us the ability to perform inference as the almost sure convergence of the parameter estimates retain their interpretability in the current model. Those results can be found in Table 3.9 below. The results bring to mind the image of a heavily intoxicated individuals trying to walk. The nature of the measured data for the Z-axis implies that the Proposed Nonparametric, the Penalized Nonparametric and the Proposed Parametric versions all find only the Z-axis as significant in explaining heavy drinking events. On the other hand the Bayesian Probit finds the Y-axis to be significant. The MLE Logistic and Penalized Logistic both indicated all variables to be significant. Thus, looking simply at the significance criteria, it is not clear which of the methodologies should be relied upon.

However, when we compare the model fits for the various methodologies, the nonparametric applications stand out as clear winners. The Proposed Nonparametric application had the lowest TeD AIC at 0.94. The next best model fit was for the Proposed Nonparametric Penalized (.95) application with Proposed Parametric Logistic (.97) coming in third in this regard. Accordingly, it is clear that in regards to MIPs the Proposed Nonparametric methodologies have a clear advantage in this application over the other existing methods compared.

Table 3.9: Intoxication Dataset Parameter Summary for All Relevant Methodologies

Methodology	Predictor	Estimates	CI-Low	CI-High
(1)	Intercept	0.24**	0.01	0.47
	X-axis	0.02	-0.22	0.25
	Y-axis	0.03	-0.19	0.25
	Z-axis	-0.54**	-0.83	-0.26
(2)	Intercept	0.22	-0.03	0.48
	X-axis	-0.07	-0.31	0.17
	Y-axis	0.21	-0.04	0.46
	Z-axis	-0.82**	-1.05	-0.59
(3)	Intercept	-0.13	-0.3	0.05
	X-axis	0.01	-0.19	0.2
	Y-axis	0.07	-0.13	0.27
	Z-axis	-0.21**	-0.37	-0.05
(4)	Intercept	-0.01	-0.14	0.11
	X-axis	-0.12	-0.32	0.08
	Y-axis	0.24**	0.06	0.43
	Z-axis	-0.02	-0.15	0.10
(5)	Intercept	-0.87***	-0.9	-0.85
	X-axis	-0.04*	-0.09	0
	Y-axis	0.17***	0.11	0.23
	Z-axis	0.00***	0.00	0.00
(6)	Intercept	-0.87***	-0.9	-0.84
	X-axis	-0.04*	-0.11	0.02
	Y-axis	0.17***	0.09	0.25
	Z-axis	0.00***	0.00	0.00

Note: (1) Nonprametric, (2) Penalized Nonparametric, (3) Parametric, (4) Existing Bayesian, (5) MLE Logistic, (6) Penalized Logistic.

3.4.2 Exotic Particle Detection Using Particle Accelerator Data

In order to see the applicability of the methodology across other scientific fields, I now apply the methodology to the identification of high-energy particles in Physics (Baldi et al. (2014)). There are 28 feature sets in the paper, of which the first 21 features are kinematic properties measured by detectors in the particle accelerator, and the last 7 are high-level features derived from the first 21 to discriminate between the two classes. The classes of 0 and 1 refer to noise and signal, respectively. In addition, the model also incorporates an intercept.

$$Signal/Noise = Intercept + \sum_{i=1}^{28} Feature_i. \tag{3.105}$$

For more information on the actual feature sets I refer the reader to the original paper, and here keep the discussion brief. Further note that, as the last seven features were nonlinear functions of the first 21, the specification remained valid, as inference is not the specific goal here. Given the far larger data size, over the Biostatistics application, I ran LAHEML for $5,000$ iterations with $2,500$ burn-in period. The convergence plots, along with the histograms of each parameter may be found below in Figure 3.5, Figure 3.6, Figure 3.7, and Figure 3.8. The penalized and unpenalized estimation formulations were identical to that for the Intoxication application for Biostatistics. Again, the classification outcomes were extremely encouraging, and can be found in Table 3.10 below.

Table 3.10: Signal/Noise Detection Summary of ARS for Nonparametric Application to Exotic Particle Detection Data.

	Unpenalized Nonpara. (TrD)	Unpenalized Nonpara. (TeD)	Penalized Nonpara. (TrD)	Penalized Nonpara. (TeD)
ARS	0.36	0.06	0.44	0.14

Note: Unpenalized Nonpara. (TrD) is the unpenalized application on the training data, and Unpenalized Nonpara. (TeD) is the unpenalized application on the test data (last 500,000 observations). Penalized Nonpara. (TrD) is the penalized application on the training data, and Penalized Nonpara. (TeD) is the penalized application on the test data (last 500,000 observations).

The unpenalized application was especially good for the Test Dataset (TeD), with the penalized version also giving excellent results in TeD, that appear to be an improvement on the initial publication. On average the unpenalized version idetified the correct Signal to Noise almost 79.23%, of the time, but in TeD it had an accuracy of almost 94.00%! Accordingly, the efficacy of the model is readily apparent in this application. The penalized application for this dataset did not have better results for the same number of iterations. However, since both formulations were only run for 5,000 iterations it seems plausible that the same pattern seen in the Biostatistics application may also be present here. This is because the penalized version is expected take longer to converge given the extra complexity of the estimation process.

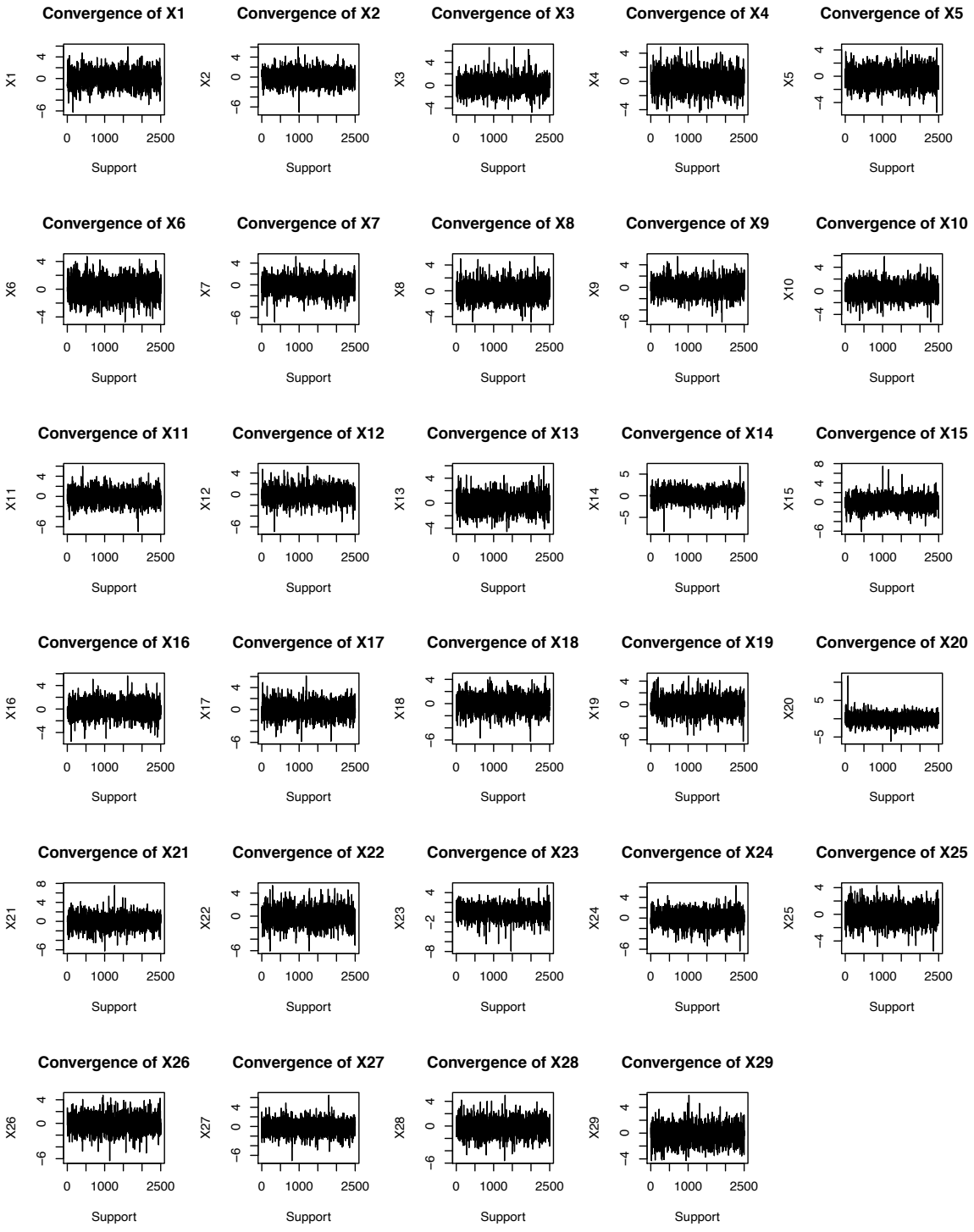


Figure 3.5: Unpenalized Convergence Plots of Nonparametric Application to Exotic Particle Detection Data. Note: The first plot in the upper left corner represents the intercept (X1). All other plots are sequential from left to right as presented in Baldi et al. (2014).

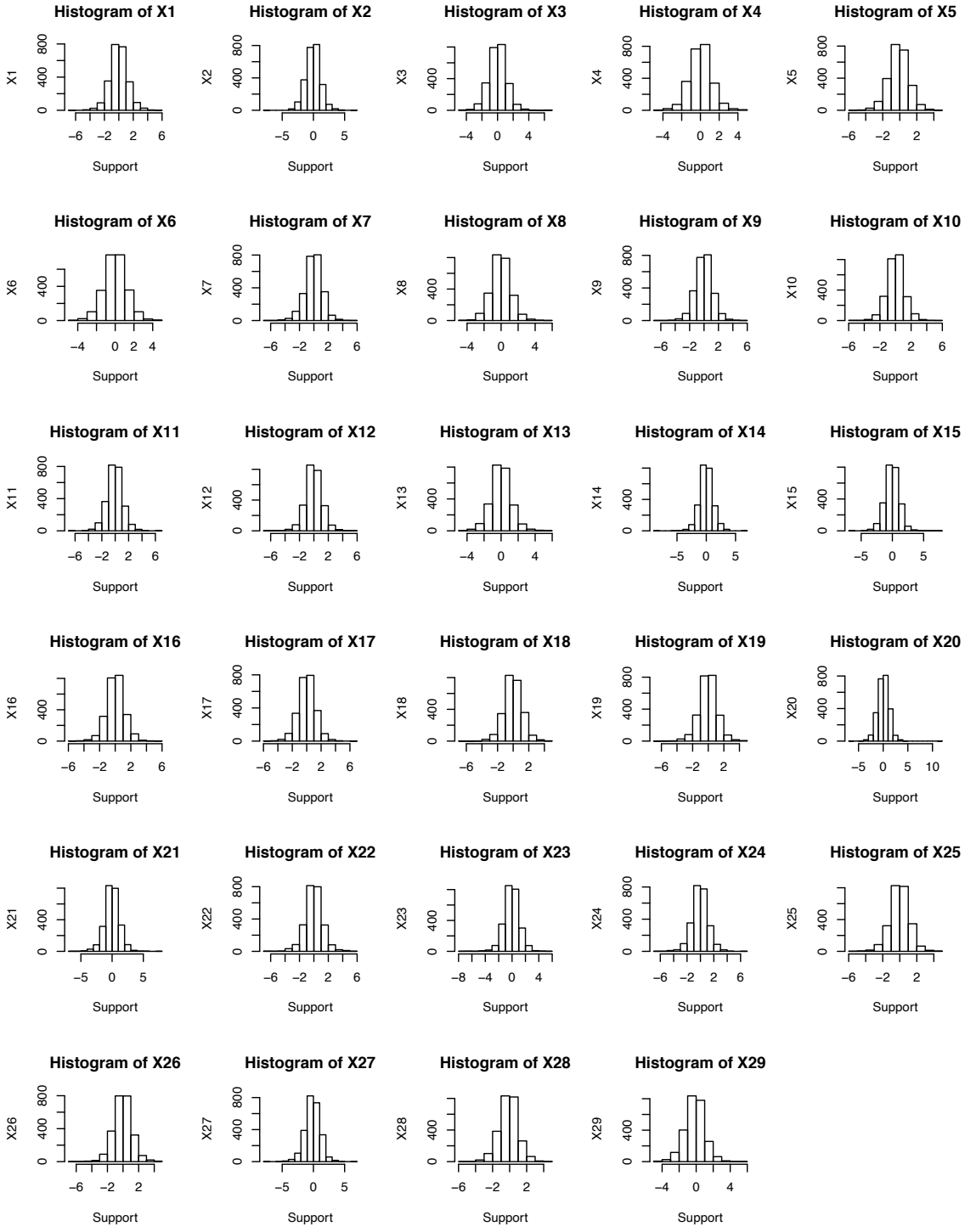


Figure 3.6: Unpenalized Histograms of Nonparametric Application to Exotic Particle Detection Data. Note: The first plot in the upper left corner represents the intercept (X1). All other plots are sequential from left to right as presented in Baldi et al. (2014).

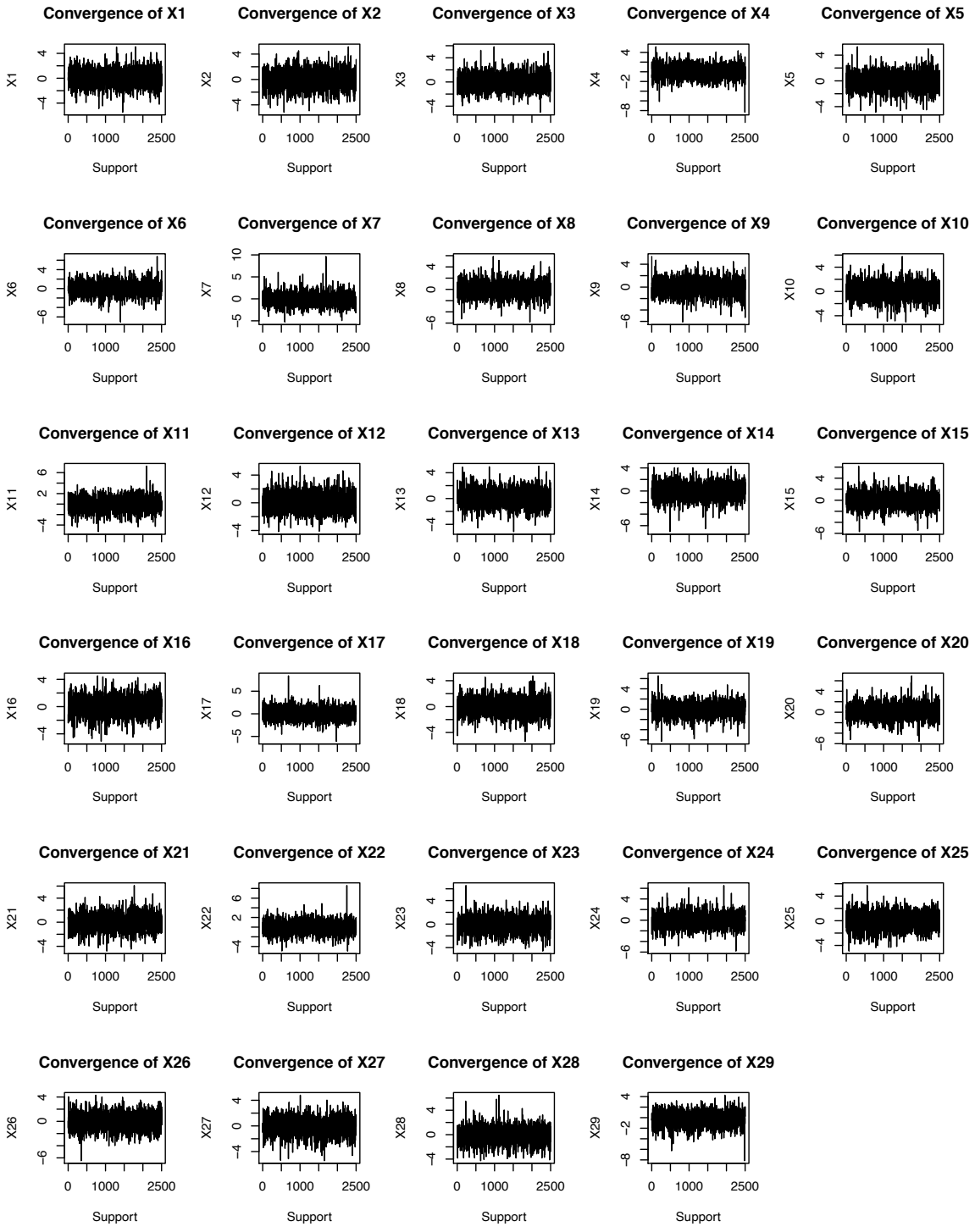


Figure 3.7: Penalized Convergence Plots of Nonparametric Application to Exotic Particle Detection Data. Note: The first plot in the upper left corner represents the intercept (X1). All other plots are sequential from left to right as presented in Baldi et al. (2014).

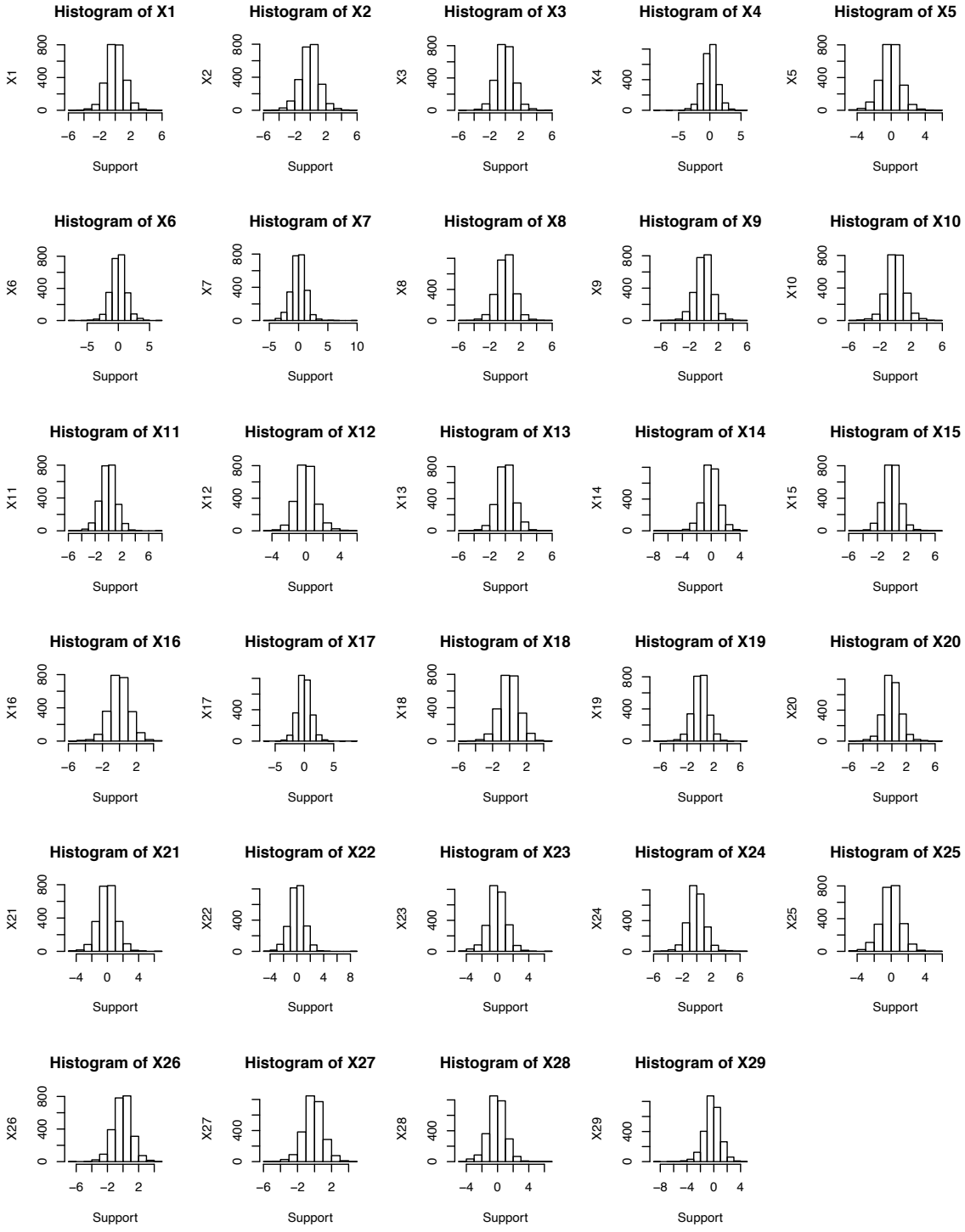


Figure 3.8: Penalized Histograms of Nonparametric Application to Exotic Particle Detection Data.
Note: The first plot in the upper left corner represents the intercept (X1). All other plots are sequential from left to right as presented in Baldi et al. (2014).

3.5 Discussion

The simulation results are extremely encouraging, because of the variety of datasets on which the different models are compared. Indeed the results validate the theorems and mathematical findings on which they are based. The Proposed Nonparametric method contained the true parameters nearly always with just 5,000 MCMC iterations, even without fixing the variance parameter as is done in existing widely used latent variable formulations. This level of coverage was attained with a smaller confidence interval than the Penalized Logistic regression. Furthermore, the methodology, even in a very simple formulation, easily outperformed ANN for classification, with the difference being statistically significant on average between the two models. Since the interpretation of the parameters remain tractable in the proposed model as opposed to ANN, it further highlights the usefulness of the methodology for myriads of scientific applications.

The application to the real-world datasets also gave extremely encouraging results, yielding deeper insights beyond just the efficiency of the methodology itself. The most apparent of these is that the methodology gives classification results which are 27.10% better than Random Forests for the biomedical dataset, with 94.00% accuracy for the high-energy application in TeD. In addition, the methodology also showed potential for performing model fit and model selection at the same time. The Bayesian Adaptive Lassso application gave results even better than the unpenalized version in the biomedical application, since its classification results were 28.03% better than the Random Forest application for the dataset. However, for the high-energy particle application this was not the case, with the unpenalized version outperforming the penalized application in both TeD and TrD. This may be explained by the small number of iterations performed, as the extra complexity of the penalized formulation usually requires more iterations.

This highlights the importance of convergence concepts as well as the underlying topological

spaces on which they are applied. Stronger convergence coupled with the ability to run it on a stronger topology such as L^1, means that even simple models can outperform, more complex models on weaker topologies. Further that this may be done without losing scientific interpretability of the parameter estimates. The ability to compare and contrast the suitability of model fits, for any of the infinitely many parametric distributional assumptions also adds another layer of applicability and usefulness to the methodology across the sciences.

The mathematical results also add to our continuing discussion on the importance of statistical "significance" as it relates to scientific significance. They point to the importance of methodologies that have strong convergence of the parameter estimates on stronger topological spaces over weaker convergence concepts (such as convergence in probability or convergence in distribution) on weaker topological spaces. As such, when inference is of interest, we may proceed using the methodology using simpler and more interpretable models. On the other hand, when classification and/or model fit are the goals, the methodology can be used in conjunction with the many excellent AI and ML models, on stronger topological spaces, for better results accordingly. Therefore, our analytic exercise becomes an attempt to find the best model, using the robust methodology, over finding the significant parameter per se. To be precise, since most models are wrong, but some are useful, the statistical goal can instead focus on robust methodologies, applied in sequentially more complex models, as needed, that rely on scientific interpretability of the model specification. If inference is not the primary goal, then we may improve on the many existing excellent AI and ML methods on stronger topological spaces, to get equivalent yet interpretable results, or in many cases better results as well. This approach gives us a more robust way to correlate scientific and statistical significance concepts to truly give the "Best of Both Worlds." Therefore, there are many possible extensions of the methodology to AI and ML applications across the sciences such as to Neural Networks and Support Vector Machines. However, these concepts require a deeper analysis of the connection between measure spaces and topological spaces, and as

such are left for future efforts.

As with any new methodology, however, its true usefulness to the sciences can only be ascertained with broad applications across the sciences, using datasets of varied characteristics. While the mathematical results give solid foundations and explanations for the excellent results, nevertheless, we must be vigilant in its application and estimation. That is, the methodology is extremely versatile in its ability to converge to the true parameter, but this does not preclude the other aspects of good data analysis such as checking for outliers or ensuring the predictors are not correlated with each other etc., especially if inference is the primary goal. However, the simulation results along with the real-world data application outcomes show much potential for the proposed methodology, and further verification is left as an open question to the greater scientific community to explore.

3.6 Conclusion

In conclusion, the mathematical foundations and simulation results show the proposed methodology makes notable contributions to widely used methodologies in the sciences. It retains parameter interpretability in a nonparametric setting, while reducing identifiability concerns with near perfect coverage probabilities with smaller confidence intervals than widely used methods. As such, it shows much potential for future real-world data applications. Accordingly, it represents a useful tool for mathematicians, statisticians and scientists to positvely contribute to our continuing conversation on the role of statistical significance and scientific significance and their interplay to answer scientific questions.

Chapter 4

An Unifying Framework

4.1 Introduction

The preceding results point to the need for a more general framework that can ensure identifiability of both the probability of success and the probability of failure uniquely. Evidently, in the Bayesian latent variable formulation pointwise convergence is not guaranteed even under very strong parametric assumptions on the probability of success across the two methodologies. Accordingly, let the probability of successes be F and F^* for binary regression and latent variable formulations respectively, which are necessarily unknown. Further note that since in the Bayesian implementation the error can take continuous values, we have the ability to also uniquely identify the probability of failure, and denote it here as F_0^* using the link constraint. Thus, the following discussion will make clear that the link constraint holding for each observation is an absolutely crucial component for almost sure convergence to hold. This assertion is true no matter the estimation technique involved such as Tanner and Wong (1987) or MLE since the characteristics of the underlying imposed topology are crucial for a properly defined linear operator on the relevant abelian group. These assertions are further elaborated below.

4.2 Mathematical Results for an Unifying Framework

In the following I present the importance of the link condition holding for each observation and state some general results as to how it extends the current GLM framework very broadly. To be precise, I first prove under what circumstance there is equivalency of the current binary and latent variable formulations. Then to present a more unified framework I present some minimal topological definitions which will be needed for the remainder of the Section. I then give results which illuminate why the current GLM framework cannot give convergence results which are almost sure. In the impossibility theorem I show the necessary

and sufficient conditions needed for almost sure convergence, and also tie in the results to the previous contributions across all chapters. Finally, I present the unified methodology in a mathematically rigorous manner.

4.2.1 Nonequivalency of Current Binomial and Latent Variable Methodologies

To see that the current latent variable and the binary formulations need not give equivalent results, we need to consider multiple criteria. The first of these have already been alluded to before. In particular, note that since asymmetric and symmetric DGPs induce different constraints on the latent probability of success, it can be used to give us our first result.

Proposition 4.1. *Let F be a distribution for the Bernoulli probability of success and F^* the distribution for the latent probability of success. Then a necessary condition for equivalence of the Binary Regression and Latent Variable specifications is that $F = F^*$.*

Proof. First note that from before in the symmetric case,

$$p_i = F(\boldsymbol{x_i'\beta_i}) = F^*[y_i^* > 0] = F^*[-\epsilon_i < \boldsymbol{x_i'\beta_i}] = F^*[\epsilon_i < \boldsymbol{x_i'\beta_i}] = F^*[\boldsymbol{x_i'\beta_i}]. \tag{4.1}$$

Assume the assumption of the proposition does not hold. Then,

$$F^*[-\epsilon_i < \boldsymbol{x_i'\beta_i}] \neq F^*[\epsilon < \boldsymbol{x_i'\beta_i}], \tag{4.2}$$

and we have by definition

$$p_i = F(\boldsymbol{x_i'\beta_i}). \tag{4.3}$$

Let $\boldsymbol{\beta_i}$ be given, further we know by construction $\boldsymbol{x_i}$ are considered fixed. Then,

$$p_i = F(\boldsymbol{x_i'}\boldsymbol{\beta_i}) \neq F^*[\epsilon < \boldsymbol{x_i}\boldsymbol{\beta_i}], \tag{4.4}$$

a contradiction to our hypothesis. The conclusion of the result then follows straightforwardly. □

The above result seems rather intuitive, since almost always the true distribution of success is unknown. However, it is more difficult to see that indeed the convergence is not even guaranteed pointwise between the two methodologies even if the true distribution of the probability of success is somehow known and assumed to be the same for both the binary and latent variable formulations. To see this first note that the probability of success in the binary case is given by $F(\boldsymbol{x_i}'\boldsymbol{\beta_i})$ and it is assumed that the probability of failure is given by $1 - F(\boldsymbol{x_i}'\boldsymbol{\beta_i})$, which appears to be a reasonable conclusion. Yet the above results and the uniqueness of the Jordan Decomposition implies that this relationship need not hold even pointwise for the latent formulation!

To see this perhaps an example would be the best tool at present. Consider the following sample points over $\lambda = \{-2, -1, 0, 1, 2\}$. WLOG assume the Hahn Decomposition exists such that the signed measure ν is finite. Further, that in the following latent specification we have the two unique measures give the following values,

$$\nu^+ = \{0, 0.25, 0.35\}, \tag{4.5}$$

$$\nu^- = \{0.25, 0.4\}. \tag{4.6}$$

Then if $\bar{F}_1$ and $\bar{F}_0$ are the unnormalized measures with subscripts indicating the relevant

Bernoulli outcomes we have,

$$|\bar{\nu}| = \frac{1}{2}\left(\bar{F}_1 + \bar{F}_0\right), \tag{4.7}$$

is a probability measure. Consider the sample points for {-0.25, 0.25}. We have,

$$|\bar{\nu}|(-0.25) = 0.19 \neq 1 - |\nu|(0.25) \approx 0.29, \tag{4.8}$$

A surprising result indeed! This leads us to our next result.

Proposition 4.2. *The binary regression and latent variable formulations are equivalent if and only if* $c_1\nu^+ = h(F^*) = F$ *and* $c_2\nu^- = 1 - F = (1 - h(F^*))$ *a.e. on the measureable space* (λ, Σ) *and* h *is a monotonic function of* F^* *with* $\{c_1, c_2\} \in \boldsymbol{R} \setminus \{-\infty, \infty\}$.

Proof. For the backward direction let, $c_1 = c_2 = 1$ and $F = F^*$ but $1 - F \neq 1 - F^*$. Then the statement clearly does not hold. Since then if $\nu^+ = 1 - \nu^-$, we have the binary regression assumptions may hold for the Jordan Decomposition of the signed measures, yet the latent variable formulation does not equal it even pointwise. Thus, the backward negation is immediate.

For the forward direction, now assume $F = h(F^*)$ and $1 - F = 1 - h(F^*)$ a.e. on the relevant measureable space. Then,

$$L(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_i) = \boldsymbol{F_i}(1 - \boldsymbol{F_i}), \tag{4.9}$$

$$L(y_i^*|\boldsymbol{x}_i, \boldsymbol{\beta_i}) = h(\boldsymbol{F_i^*}). \tag{4.10}$$

$$\implies F \propto h(F^*) \text{ and } 1 - F = 1 - h(F^*) \tag{4.11}$$

$$\implies L(y_i|\boldsymbol{x}_i, \boldsymbol{\beta_i}) \propto L(\boldsymbol{y}_i^*|\boldsymbol{x}_i, \boldsymbol{\beta}_i) \text{ since } h(F^*) \propto F \text{ pointwise.} \tag{4.12}$$

Since the signed measure ν can be decomposed into finite measures by proposition 3.3 and 3.4, using Birnbaum's Theorem and the likelihood principle, we would arrive at the same inference for each methodology. The statement then is verified. $\square$

The results are striking to this mathematician and gives several far reaching consequences. It states that the two methodologies need not be equivalent even under the assumption that $F = F^*$, even without assuming any measure theoretic applications. We must consider the probability of success and failure to be two separate measures for unique identifiability to hold. It further illuminates that no matter the estimation technique involved, simply assuming MLE results is not enough for congruence between the two models even in large samples, a finding readily validated in numerous empirical applications across the sciences (see for example Chowdhury (2021a) and Chapter 2 for a more detailed discussion).

In fact, the result gives rise to several other relevant questions as to when the assumptions on the existing frameworks can and cannot be supported. The results below highlight these considerations.

4.2.2 Topological Definitions

To facilitate this discussion the following definitions are asserted and may be found in almost any graduate level Topology book.

Definition 4.1. *A linear space $\mathcal{X}$ is an abelian group with group operation addition, such that for a real number α and $\sqcap \in \mathcal{X}$ and $\{\alpha, \beta\} \in \mathcal{R}$ a scalar product $\alpha.\sqcap \in \mathcal{X}$ and the following properties hold,*

- $(\alpha + \beta).\sqcap = \alpha.\sqcap + \beta.\sqcap.$
- $\alpha.(\sqcap + \sqsubseteq) = \alpha.\sqcap + \alpha.\sqsubseteq;\ \sqsubseteq \in X.$

- $(\alpha.\beta).\sqcap = \alpha.(\beta.\sqcap)$ *and* $1.\sqcap = \sqcap$.

The addition and scalar multiplications are defined pointwise on all of $\mathcal{X}$. On this space we may define a norm $||.||$ as if $\sqcap, \sqsubseteq \in \mathcal{X}$ and $\alpha \in \mathcal{R}$ then $||\sqcap|| = 0$, *if and only if* $\sqcap = 0$, $||\sqcap + \sqsubseteq|| \leq ||\sqcap|| + ||\sqsubseteq||$ and $||\alpha\sqcap|| = |\alpha|||\sqcap||$.

For two normed linear spaces we may define a linear operator as follows.

Definition 4.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be linear spaces. A mapping $T : \mathcal{X} \to \mathcal{Y}$ is called a linear operator if for each $\sqcap, \sqsubseteq \in \mathcal{X}$ and real numbers α and β,*

$$T(\alpha\sqcap + \beta\sqsubseteq) = \alpha T(\sqcap) + \beta T(\sqsubseteq). \tag{4.13}$$

In addition to the elementary definitions above I will work on a particular type of linear space the Banach spaces, and define it accordingly below.

Definition 4.3. *A Banach space is a complete normed linear space.*

To be complete we need one more topological concept, that of the Hausdorff Separation property.

Definition 4.4. *A topological space equipped with the Hausdorff Separation Property implies that any two points on the topological space can be separated by disjoint sets.*

These definitions now provide the tools needed to understand under what circumstances the existing latent and binary regression frameworks are equivalent. The result below highlights this specification.

Proposition 4.3. *Let $\mathcal{X}$ be a compact Hausdorff space, and $\mathcal{Y}$ a compact subspace on which the link condition holds. Then under the assumptions of Proposition 3.4, the existing Latent Variable and Binary Regression frameworks are equivalent if and only if the underlying probability of success (failure) is symmetric around the origin.*

Proof. Case I: Assume that the assumptions of Proposition 3.4 holds. Then using Theorem 3.1, Proposition 3.4 and Proposition 3.5 we have that ν^+ can be extended to all of $\mathcal{X}$ through a linear functional $\mathcal{L}$ on which

$$\mathcal{L}(\mathcal{X}) \leq \nu^+(\mathcal{X}). \tag{4.14}$$

Then from elementary functional analysis (Lax (2002)) we know that for $C(\mathcal{X})$ the space of continuous real-valued functions normed by the maximum norm every bounded linear functional $\mathcal{L}$ on $\mathcal{X}$ and $f \in C(\mathcal{X})$ we have that,

$$\mathcal{L}(f) = \int_{\mathcal{X}} f\nu^+(dx), \tag{4.15}$$

where ν^+ belongs to C', the space of all finite signed measures. Consider the measure space $(X, \Sigma, |\hat{\nu}|)$ as in Proposition 3.4. Then,

$$|\mathcal{L}| = 1. \tag{4.16}$$

But by Hahn-Banach we can extend this linear functional to all of $\mathcal{X}$, and therefore, define

$$\nu^- = F_0^* = 1 - \int_{\mathcal{X}} f\nu^+(dx), \tag{4.17}$$

to get the desired assertion.

Conversely, now assume that (4.17) holds. Then again by Hahn-Banach we have that there exists a linear funtional $\mathcal{L}_1 : \mathcal{X} \to \mathcal{Y}$ such that

$$|\mathcal{L}_1| = 1, \tag{4.18}$$

and,

$$\nu^+ = F^* = 1 - \int_{\mathcal{X}} f\nu^-(dx). \tag{4.19}$$

Thus, it remains to prove the necessity of symmetry around the origin.

But this specification makes clear the circumstances under which we can assume a symmetric distribution for the probability of success for the existing latent variable or binary regression framework, when,

$$\nu^- = F_0^* = 1 - \int_{\mathcal{X}} f\nu^+(dx) = \nu^+ = F^* = 1 - \int_{\mathcal{X}} f\nu^-(dx). \tag{4.20}$$

Thus, we are done.

Case II: Now assume that that the conditions of Proposition 3.3 holds. This scenario is considerably more convenient to deal with. Consider the case that

$$|\nu^+| = |\nu^-|. \tag{4.21}$$

Then again using Hahn-Banach we may extend a linear functional

$$\mathcal{L}_2 : \mathcal{X} \to \mathcal{Y}, \tag{4.22}$$

such that (4.20) holds. Thus, we are done.

□

Therefore, the circumstances under which such an assumption is justified would under most circumstances be considered extremely restrictive and unlikely to represent the underlying stochastic process! Yet this mathematical formulation and line of reasoning provides even further striking results regarding the impossibility of almost sure convergence in any GLM framework for the existing specification. These results are summarized below.

4.2.3 The Impossibility of Almost Sure Convergence in the Current Framework

To show that the existing GLM framework does not guarantee almost sure convergence in general for any estimation technique it is necessary for us to consider linear operators between linear spaces relevant to both the underlying systematic component and for the link function. The following lemma puts this into more concrete terms.

Lemma 4.1. *Let $\mathcal{X}$ be a finite dimensional linear space and consider $\mathcal{Y}$ as the linear subspace on $\eta = g(\mu) = \lambda$, as defined before as the link condition that ties the systematic component to the mean. Then there exists no unbounded linear operator T such that*

$$T : \mathcal{X} \to \mathcal{Y} \tag{4.23}$$

is continuous.

Proof. First note that, by construction $\mathcal{X}$ is a finite dimensional linear space. Then for $\boldsymbol{\beta} \in \mathcal{R}^n, n < \infty$, $\lambda(\mathcal{X}, \boldsymbol{\beta}) \in \mathcal{Y} \subset \mathcal{X}$ is a linear subspace of $\mathcal{X}$. Further by construction of a GLM we know that $T : \mathcal{X} \to \mathcal{Y}$ must exist. Furthermore, by construction this linear

operator must be unique for any given $\boldsymbol{\beta}$. Assume, T is unbounded. Then, for any $\sqcap \in \mathcal{X}$,

$$||T(\sqcap)|| \geq M||\sqcap||, \forall \sqcap \in \mathcal{X}. \tag{4.24}$$

By definition, $\sqcap_n \to \sqcap$ then pointwise convergence implies $\{T(\sqcap_n)\} \to T(\sqcap)$. Suppose T is continuous. WLOG consider $\epsilon = 1$ at $\sqcap = 0$. Then we should be able to pick a δ such that $||T(\sqcap) - T(0)|| < 1$ with $||\sqcap|| < \delta$. But by assumption T is unbounded. Thus, there exists no $M \geq 0$ such that

$$||T(\sqcap)|| < M||\sqcap||, \forall \sqcap \in \mathcal{X}. \tag{4.25}$$

Therefore, no such δ exists since no such M exists with $\delta = (M||\sqcap||)^{-1}$. Therefore, T is not continuous as needed.□

The above results are instructive. It is not possible to find a continuous linear operator between the sample space and the link function if the link function can be either infinite dimensional with respect to the strong topology or it takes nonfinite values. Even if we assume the observed explanatory variables are a finite sample from a finite dimensional space, perhaps we may disregard the infinite dimensional case, but we cannot disregard any undefined values taken by such a linear operator, as this implies the operator is not continuous. If the operator is not continuous, many well known convergence results fail to hold regardless of whether Bayesian or Frequentist estimation methodology is used. Indeed this further implies that the results of Tanner and Wong (1987) need not be continuous as there may exist a linear operator which need not be bounded. Consequently, in Albert and Chib (1993) we need not have unique convergence to the mean regardless of the MCMC method used to identify the posterior even if the observed data coincide with the strong assumptions mentioned previously.

Further note that in Chapter 3 I discussed that equivalency between the binomial and latent

variable regression specification relies on the likelihood principle. That is, the two likelihoods must be proportional to each other for the almost sure convergence to hold between them. Accordingly, consider the Logistic link function,

$$ln\left(\frac{p_i}{1-p_i}\right) = \lambda_i. \tag{4.26}$$

As discussed in Chapter 2 the link function is indeed not bounded as

$$\{p_i\} \to 1 \implies ln\left(\frac{p_i}{1-p_i}\right) \to \infty. \tag{4.27}$$

Therefore, there can be no continuous linear operator between the sample space and the link field equipped with either the normal or hausdorff topology. The statement of the result above is actually rather more innocuous than perhaps its implications may initially indicate. Consider the following corollary as a direct implication of the results above.

Corollary 4.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be as in Lemma 4.1. Then the Logistic and Probit formulations are equivalent in the sense of Birnbaum for any continuous or discrete GLM formulation.*

Proof. To see this rather surprising result first note the existence of a latent variable formulation is guaranteed by Proposition 3.3 and Proposition 3.4. Further note that in Proposition 3.5 I showed that a monotonic transformation of ν^+ and ν^- such that

$$\nu^+ = h(F^*) = F \text{ and} \tag{4.28}$$

$$\nu^- = 1 - h(F^*) = 1 - F, \tag{4.29}$$

would result in the same inference using Birnbaum's Theorem. Consider a simple application

using the density functions of the Logistic,

$$\boldsymbol{f_1}(\boldsymbol{y}, X, \boldsymbol{\beta_1}, \sigma_1) = \frac{exp(-[\boldsymbol{y} - X.\boldsymbol{\beta}_1]/\sigma_1^2)}{(1 + exp(-[\boldsymbol{y} - X.\boldsymbol{\beta}_1]/\sigma_1^2))^2}, \tag{4.30}$$

and that of the normal density,

$$\boldsymbol{f_2}(\boldsymbol{y}, X, \boldsymbol{\beta}, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-[\boldsymbol{y} - X.\boldsymbol{\beta}]/\sigma^2). \tag{4.31}$$

Thus, if $\sigma = \sigma_1$ the two exponential densities are proportional to each other, with an adequate monotonic transformation and without the imposition of a link constraint should give similar inference results in large samples under i.i.d. assumption even if the model fit and prediction results differ. Since the current Latent Variable framework imposes fixing the variance of the latent distribution for identification purposes, we may readily apply the constraint that $\sigma_1 = \sigma$. But the Bayesian and Frequentist formulations are identical under prior restrictions using existing latent variable framework. Therefore, the statement of the corollary must hold in either formulation. Thus, the assertion of the corollary then readily follows. □

Indeed the result is validated across the sciences where (see for example Albert and Chib (1993), Cameron and Trivedi (2010)) they state each model's parameters appear to be a constant multiple of the other. Specifically, $\boldsymbol{\beta}_{Logit} \approx 1.6\boldsymbol{\beta}_{Probit}$ and seem to apply quite well in empirical applications. This then gives one of the more poignant results of this work.

Corollary 4.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be as in Lemma 4.1. Then pointwise convergence is not guaranteed for any $F^* = F$, and the linear functional*

$$T : \mathcal{X} \to \mathcal{Y}, \tag{4.32}$$

and the statement holds whether we use a Bayesian or Frequentist formulation.

Proof. Under the conditions of Lemma 4.1 we have that no continuous linear functional exists between the sample space $\mathcal{X}$ and $\mathcal{Y}$, for any sequence $\{c(X)\boldsymbol{\beta}\} \to \lambda \in \mathcal{Y}$, if for some n, $\nu^{+}(\lambda) = \infty$ or $\nu^{-}(\lambda) = -\infty$. Therefore, by Proposition 3.3 and Proposition 3.4 using Corollary 4.1 we have that no continuous functional exists such that pointwise $T : \mathcal{X} \to \mathcal{Y}$ holds.

Furthermore, by Proposition 3.4 we know that a latent variable formulation in the discrete case can be extended to the continuous case. Therefore, the statement of the result holds for any GLM.

Observe, that this result is independent of the linear functional used and therefore it is independent of whether a Bayesian or Frequentist operator is used in the estimation process. □

This result seems rather strong, since continuous regressions are run everyday in virtually all scientific fields without identifiability concerns always apparent. However, note that the issue is more pernicious than may appear. The issue becomes apparent when the probability of success or failure is exactly equal to 1. This necessitates the $\boldsymbol{\beta}$'s to be numerically large for the link condition to hold explicitly, even if not considered in the estimation process. This in turn can result in non-convergence. This is often dealt with in practice, by throwing away a particular observation that may be causing estimation issues. While such an approach can lead to convergence to some parameter, it is not clear why such a parameter should be equivalent to the true likelihood under Birnbaum's theorem. In fact, the mathematical results state that they should not be. Another approach often taken is to start from multiple starting points to get the best model fit results. However, even this approach does not guarante convergence to the true parameter even pointwise, since there are infinitely many models and starting points that can be considered. Therefore, the need for a more rigorous functional analysis approach that is applicable across the sciences seems clear. Accordingly, in Chapter 2 I ensured that almost sure convergence can be achieved without facing the issues

of an unbounded, discontinuous linear functional pointwise for some functional specification of $\lambda(X, \boldsymbol{\beta})$. In Chapter 3, I presented a methodology which ensured that if the latent and binary formulations are equivalent, then we may assure almost sure convergence of the parameters. In the current formulation, using some of the results of the previous chapters, and some of the insights presented above, I present a methodology which ensures convergence to the unique measure using Jordan Decomposition as in Chapter 3 but in a far more general framework, for any continuous, bounded link specfication subject to the link constraint holding for each observation.

4.2.4 An Unified Almost Sure Convergence Methodology

The preceding chapters have through rigorous mathematical arguments laid the foundations for almost sure convergence to the true parameters of interest under the binary regression framework. They have done so by harnessing the advantages of both the latent variable and binomial regression case to overcome their respective disadvantages. Specifically, we know that the binary error can only take one of two values, either 0 or 1. On the other hand, for the latent variable formulation we know the error can be continuous. In the nonparametric section I set the mathematical foundations for the completely new robust methodology using the Jordan Decomposition Theorem for a signed measure. Those results showed it to be superior to existing methods with improvements to the parametric version under various settings. However, it still ensured for equivalency to the binomial regression framework that $\nu^- = 1 - F^* = 1 - F$. However, in the examples above I argued that this condition need not hold at all even if $F^* = F$. Therefore, in this section I outline a methodology where we may relax this restrictive constraint.

Accordingly, consider a likelihood function as follows,

$$L(X|\boldsymbol{\beta}) = \boldsymbol{c}(\boldsymbol{\nu}^{+}(\boldsymbol{\lambda}))^{(k)}(\boldsymbol{\nu}^{-}(\boldsymbol{\lambda}))^{(n-k)}, \boldsymbol{c} \in \mathcal{R},\ k = \sum_{i=1}^{n} y_i. \tag{4.33}$$

Note that such a likelihood function can be supported anytime we have independence between the two sample spaces over the cutoff point of 0. The framework of Kass and Steffey (1989) ensures that this can be done as an extension of the methodology presented in Chpater 3. As in the nonparametric case, the current formulation also allows for this cutoff to be based on a normalized posterior probability such as the median. Using the ability to run the estimation algorithm over the proportional posterior then allows us to extend the nonparametric methodology in a more robust formulation.

In particular, note that the above formulation in (4.33) is justifiable anytime the formulations discussed in Chapter 2 and Chapter 3 are valid. We know by Kass and Steffey (1989), these formulations can be supported anytime there are unobserved variables that may impact the outcome of interest. Since by necessity observed $\mathcal{X}$ are not infinite dimensional, we see that the formulation is valid and viable, in addition to the discussions of Chapter 2 and Chapter 3. Therefore the link condition holding pointwise now takes the following formulation,

$$(\boldsymbol{\nu}^{+})^{\alpha_1^*} = \boldsymbol{\lambda}_{|S^{+}}(X, \boldsymbol{\beta}), \tag{4.34}$$

$$(\boldsymbol{\nu}^{-})^{\alpha_2^*} = \boldsymbol{\lambda}_{|S^{-}}(X, \boldsymbol{\beta}). \tag{4.35}$$

This general framework then requires a substantially more intricate proof to guarantee almost sure convergence of LAHEML extended to all measure spaces whether finite or $\sigma-$finite. To see this the theorems below make this formulation clear in a mathematically rigorous way. In doing so it adds to some well-known results in Real Analysis and Pure Mathematics.

Theorem 4.1. *Let (X, Σ, ν) be a finite measure space, with ν finite a.e. Then under Latent Adaptive Hierarchical EM Like Formulations in Frequentist or Bayesian framework,*

$$\hat{\beta} \xrightarrow{a.s.} \beta \tag{4.36}$$

in $L^p(X, \nu)$, where $1 \leq p \leq \infty$ and $p = \infty$ represents the essentially bounded case.

Proof. Consider Theorem 3.2, where almost sure convergence was asserted for ν finite or finite a.e. on X. Thus, the case for $1 \leq p < \infty$, is immediate. It remains then to show the case for $p = \infty$. Thus, as in Theorem 3.3, let

$$f_n(\lambda(\hat{\beta}^{(j)})|y^{(j)}, \alpha^{(j)^*}) := f_n^j. \tag{4.37}$$

Let $\{f_n\}$ be the sequence of functions on X for all j. Then $\{f_n\}$ is bounded and finite by construction for each $i \in \{1, 2, ..., n_j\}$, where $\cup_{i=1} E_i^j$ are the respective disjoint covering sets. Let $\epsilon > 0$, then there exists a $\delta > 0$ for $E \in \Sigma$ such that,

$$if\ \nu(E) < \delta\ then\ \int_E |f_n| < \epsilon, \tag{4.38}$$

follows straightforwardly from finiteness over E. Therefore, by Dunford-Pettis Theorem (Royden and Fitzpatrick (2010)), f_n is weakly compact. Thus, by the Kantorovich Representation Theorem, there exists a linear functional,

$$T_\nu : L^\infty(X, \nu) = \int_X f\ d\nu \to \mathcal{R}, \tag{4.39}$$

where T is an isometric isomorphism of (X, Σ, ν) on to the dual of $L^\infty(X, \nu)$ (Ibid). Thus, we are done. □

Theorem 4.2. *Let X be a locally compact and Hausdorff topological space such that (X, Σ, ν) is a $\sigma-$finite measure space. Then under Latent Adaptive Hierarchical EM Like Formulations in Frequentist or Bayesian framework,*

$$\hat{\beta} \xrightarrow{a.s.} \beta \tag{4.40}$$

in $L^p(X, \nu)$, where $p \in \{1, ..., \infty\}$, where $p = \infty$ represents the essentially bounded case.

Proof. Consider the space of functions on X, $\mathcal{L}$ which are essentially bounded such that, there exists some $M \geq 0$ with

$$|f| \leq M \ a.e. \ on \ X. \tag{4.41}$$

Then there exists a linear functional such that $0 \leq |f| \leq 1$. Following existing set up (Royden and Fitzpatrick (2010)) for L^P spaces, define $L^P(X, \nu)$ to be the collection of $[L] \in \mathbf{L}$, as the collection of extended real valued functions on X, which are finite a.e. on X. Thus, integrability implies measureability for all $f \in \mathcal{L}$. That L^P is a banach space is well extablished and it is stated without proof going forward.

Accordingly, consider the dual of X, X^*. By Alaoglu's Theorem the unit ball on X^* is weak-* compact. Let

$$\bar{B}^*(1) = \{\psi_k \in X^* : |\psi_k - \psi| \leq 1\}. \tag{4.42}$$

Fix $\delta_k > 0$, then by Alaoglu there are finitely many $\psi_i's \ s.t. \ 0 \leq i \leq 1$,

$$|\psi_i - \psi_k| < \delta_k, \tag{4.43}$$

for some $x \in X$ such that we may define sets of the form

$$X_k = \{x, x_{n_k} \in X : |\psi_k(x_{n_k}) - \psi(x)| < \frac{1}{2^{n_k}}\}. \tag{4.44}$$

Further note that X is $\sigma - finite$ such that there exists disjoint open sets with $X = \cup_{k=1}^{n} E_k$, where $E_1 = \{\omega\} \cup X \sim \cup_{k=2}^{n} E_k$, with ω the one point Alexandroff compactification of E_1. Now each E_k is endowed with the subspace topology from X. Let Λ_k be a dense subset of $(-1, 1)$, and define on E_k a normally ascending collection of open subsets $O^{o}_{\lambda_k}$ of E_k with $\lambda_k \in \Lambda_k$. Let $f_k : E_k \to \mathcal{R}$, with $f_k = 1$ on $E_k \sim \cup_{\lambda \in \Lambda} O^{o}_{\lambda_k}$, and otherwise setting

$$f_k(x) = inf\{\lambda_k \in \Lambda : x_n \in O^{o}_{\lambda_k}\}. \tag{4.45}$$

Then

$$f_k : E_k \to [-1, 1], \tag{4.46}$$

is continuous.

Thus, for each E_k we may define a normally ascending collection of diadic rationals such that,

$$E_k \subseteq O^{o}_{\lambda_1} \cup O^{o}_{\lambda_2} ... O^{o}_{\lambda_{n_k}} \subseteq \bar{O^{o}_{\lambda_1}} \cup \bar{O^{o}_{\lambda_2}} ... \bar{O^{o}_{\lambda_{n_k}}}, \tag{4.47}$$

for some $n_k \in \mathcal{N}$. Consider the collection of functions on

$$\mathcal{L}_{|O^{o}_{\lambda_1} \cup O^{o}_{\lambda_2} ... O^{o}_{\lambda_{n_k}}}, \tag{4.48}$$

and consider the product topology on it. Then by the Tychenoff Product Theorem, it is also compact.

Take a sequence $\{x_{n_k}\} \in E_k$ such that $\{x_{n_k}\} \to x \in E_k \subseteq X$. Then $f_k(\{x_{n_k}\})$ is discontinuous at most at countably many points. Thus, consider a sequence of functions f_{n_k} on (4.48) and diagonalize it such that for some n_k, the neighborhood for (4.44) for $\frac{1}{2^{n_k}}$ is not continuous. Take $n_{k+1} > n_k$. Continue this process untill it terminates to get an unique sequence of possibly disjoint normally ascending collection of open sets $O^{*\mathrm{o}}{}_{\lambda_k \in \Lambda_k}$. If the construction is disjoint, we have identified a collection of disjoint open sets that cover E_k. If not use the fact that X is Hausdorff and locally compact such that,

$$E_{n_k=j} = E_k \sim \cup_{n_k=1}^{j-1} E_{n_k} \implies E_k = \cup_{n_k=1}^{j} E_{n_k}, \tag{4.49}$$

is a disjoint collection of open sets. Since f is finite a.e. on X, it is also finite a.e. on E_k, and therefore the diagonalization process is valid for each E_k.

Therefore, by construction the restriction of the measure space (X, Σ, ν) to

$$(X_{|\cap_{\lambda_k} O^{*\mathrm{o}}_{\lambda_k \in \Lambda_k}}, \Sigma_{|\cap_{\lambda_k} O^{*\mathrm{o}}_{\lambda_k \in \Lambda_k}}, \nu_{|\cap_{\lambda_k} O^{*\mathrm{o}}_{\lambda_k \in \Lambda_k}}), \tag{4.50}$$

is also a measure space, since by construction $\nu(O_1^{\mathrm{o}}) < \infty$. Moreover it is finite and by compactness, a finitely additive measure space. Since by the continuity of measure,

$$\nu(E_k) \leq \sum_{\lambda_k \in \Lambda_k} \nu(\bar{O}^{\mathrm{o}}{}_{\lambda_k \in \Lambda_k}) < \infty. \tag{4.51}$$

Thus, using Theorem 4.1 we have that E_k may be covered by a countable collection of open sets $O^{\mathrm{o}}_{\lambda_k \in \Lambda_k}$, whose closure $\bar{O}^{\mathrm{o}}{}_{\lambda_k \in \Lambda_k}$ contains E_k, with the closure being compact and Hausdorff with respect to the weak-* topology on it, and the restricted measure space on it is bounded and finitely additive. Thus, we may define a linear operator,

$$T_{\nu_{|\cup_{\lambda_k} O^{\mathrm{o}}_{\lambda_k \in \Lambda_k}}}(f_k) \to \int_{\cup_{\lambda_k} O^{\mathrm{o}}_{\lambda_k \in \Lambda_k}} f_k \, d\nu_{O^{\mathrm{o}}_{\lambda_k \in \Lambda_k}} \text{ for all } f \in \mathcal{L}^{\infty}(E_k, \nu_{|\cup_{\lambda_k} O^{\mathrm{o}}_{\lambda_k \in \Lambda_k}}). \tag{4.52}$$

Then, T is an isometric isomorphism of the normed linear space $(X_{|\cup_{\lambda_k} O^{*^{\circ}}_{\lambda_k \in \Lambda_k}}, \Sigma_{|\cup_{\lambda_k} O^{*^{\circ}}_{\lambda_k \in \Lambda_k}}, \nu_{|\cup_{\lambda_k} O^{*^{\circ}}_{\lambda_k \in \Lambda_k}})$, on to $\mathcal{L}^{\infty}(E_k, \nu_{|\cup_{\lambda_k} O^{\circ}_{\lambda_k \in \Lambda_k}})$, by the Kantorovich Representation Theorem. Using Theorem 3.3 and Dunford-Petis for each $\delta_k > 0$ we may define an ϵ_k such that there exists an $M_k > 0$ with $n_k > N_k \in \mathcal{N}$,

$$\int_{\{x \in E_k : f_k(x) \geq M_k\}} |f_k| < \epsilon_k, \ for \ all \ n_k > N_k \in \mathcal{N}. \tag{4.53}$$

But,

$$\nu(E_k) \leq \nu(\cup_{n_k=1}^{j(n_k)} O^{\circ}_{\lambda_k \in \Lambda_k}) < \infty, \tag{4.54}$$

thus by Borel-Cantelli note that for each n_k, by the countable monotonicity of $\nu_{|O^{\circ}_{\lambda_k \in \Lambda_k}}$,

$$\lim_{n_k \to \infty} \nu \left(\cup_{\lambda_k = n_k}^{\infty} O^{\circ}_{\lambda_k \in \Lambda_k} \right) < 1, \tag{4.55}$$

thus all but finitely many of the $x's \in \cup_{\lambda_k = n_k}^{\infty} O^{\circ}_{\lambda_k \in \Lambda_k}$ belong to finitely many of the $O^{\circ}_{\lambda_k}{}'s$. Then by Egoroff we may choose an $n_k > N_k$, such that there exists a collection of subsets E_{n_k} of E_k such that $f_{n_k} \to f_k$ uniformly on E_{n_k} but $\nu(E_k \sim E_{n_k}) < \epsilon_k$.

Thus we may write,

$$\Big| \int_X f d\nu - \int_{\cup_{k=1}^{\infty} E_k} f_k \Big| < \Big| \int_X f d\nu - \int_{\cup_{\lambda_k} O^{\circ}_{\lambda_k \in \Lambda_k}} f_k \Big| +$$
$$\Big| \int_{\cup E_{n_k}} f_k - \sum_{E_{n_k}} T_{\nu_{|\cup_{\lambda_k} O^{\circ}_{\lambda_k \in \Lambda_k}}}(f_k) \Big| + \Big| \sum_{E_{n_k}} T_{\nu_{|\cup_{\lambda_k} O^{\circ}_{\lambda_k \in \Lambda_k}}}(f_k) - \int_{\cup_{k=1}^{\infty} E_k} f_k \Big|$$

Taking the limit of $E_k(\lambda_k) \to \infty$ as $\lambda_k \to \infty$ gives us then,

$$\Big| \int_X f d\nu - \int_{\cup_{k=1}^{\infty} E_k} f_k \Big| < 3\epsilon_k, \tag{4.56}$$

take δ_k such that $\epsilon_k < 1/3$, answers the ϵ challenge. Thus strong convergence immediately implies that we may define a sequence of random variables X_{n_k} from E_{n_k} onto $[0,1)$ such that,

$$\begin{aligned}
\nu\left(\lim_{n\to\infty} \cup_{n_k=1}^{n} \left|\nu^{-1} X_{n_k} - \nu^{-1} \int_{X_{|E_k}} f \; d\nu\right|\right) \leq \\
\nu\left(\lim_{n\to\infty} \cup_{n_k=1}^{n} \nu^{-1} \left|X_{n_k} - \int_{X_{|E_k}} f \; d\nu\right|\right) \leq \\
\nu\left(\lim_{n\to\infty} \cup_{n_k=1}^{n} \nu^{-1} \left|X_{n_k} - (X_{|E_k})\right|\right) \leq \\
\left(\lim_{n\to\infty} \sum_{n_k=1}^{n} \nu_{|E_k} \left|E_{n_k} - E_{n_k} \cap O^{\circ}_{\lambda_k \in \Lambda_k}\right|\right). \qquad (4.57)
\end{aligned}$$

Therefore, take $N_{\bar{K}} = max(N_k)$ over all k such that if $n_k = n(\lambda_k) > N_{\bar{K}}$ we get

$$\nu\left(\lim_{n\to\infty} \cup_{n_k=1}^{n} \left|\nu_k^{-1} X_{n_k} - \nu^{-1} \int_{X_{|E_k}} f \; d\nu\right|\right) = \lim_{n(\lambda_k)\to\infty} \sum_{n_k} \frac{1}{2^{n_k}} \to 1. \qquad (4.58)$$

Thus, we are done. □

Remark 4.1. *The result has some important consequences for probabilistic models, since an application of LAHEML ensures almost sure convergence for all integrable functions over the sample space.*

4.3 Monte Carlo Simulations

In order to validate the robustness of the proposed unified methodology a Bayesian framework is used for extensive simulation studies, penalized and unpenalized, on various DGP's, both symmetric (Logit and Probit) and asymmetric (Complementary Log-Log). For this purpose, datasets were generated from the standard normal distribution as before, for different sample sizes ($n = \{100, 500, 1000, 2000\}$) and models,

$$\boldsymbol{y} = Intercept + X_1 + X_2, \tag{4.59}$$

$$\boldsymbol{y} = Intercept + X_1 + exp(X_2), \tag{4.60}$$

$$\boldsymbol{y} = Intercept + exp(X_1) + sin(X_2). \tag{4.61}$$

All datasets as before, had 3 parameters to estimate, for the intercept (β_1) and for two explanatory or independent variables drawn from the standard normal ($\{\beta_2, \beta_3\}$) with the appropriate transformations indicated above. Then for known β values, a Probit, Logit or a Complementary Log-Log DGP was used to generate outcomes (dependent variable $\boldsymbol{y}$), that varied in the number of 1's that were present.

In particular, the known $\{X, \boldsymbol{\beta}\}$ values along with each functional form above can be used to calculate the probability of each observation for each specific model. Thus, we can consider the calculated $\boldsymbol{y}$ values along with the generated X's as the data on which we can fit our chosen statistical models for each DGP. Finally, another step was done to create datasets which had different numbers of successes as opposed to failures. Thus, the unbalancedness of the data were varied between $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Here, 0.5 indicates equal number of successes and failures (balanced), 0.4 indicates 10% fewer successes than failures and so forth. Accordingly, for each sample size there are five different unbalanced datasets, each of

which has three parameters or β's to estimate for each of the three DGPs for each of the models specified (linear, non-linear or mixed). As such, for each model, there are 60 different datasets, each with 3 parameters to estimate, for a total of $180 \times 3 = 540$ parameters to estimate, compare and contrast. The penalized results are summarized below.

Table 4.1: Simulation Coverage in Percentage Summary All DGPs (at 1% Significance Level, Reported in Percentage)

	Prop. N	Prop. NU	Prop. NP	Prop. NPU	Prop. Log.	Bayesian Probit	Pen. Logit
LGR Covr. (NL)	97.37	100.00	98.68	98.68	97.37	77.63	68.42
PR Covr. (NL)	98.33	98.33	96.67	98.33	95.00	70.00	53.33
Comp. Lg. Covr. (NL)	98.75	100.00	98.75	100.00	100.00	88.75	72.50
LGR Covr. (Mx.)	100.00	100.00	100.00	100.00	100.00	83.33	75.00
PR Covr. (Mx.)	100.00	100.00	100.00	98.68	100.00	89.47	69.74
Comp. Lg. Covr. (Mx.)	98.75	100.00	98.75	98.75	98.75	90.00	69.74
LGR Covr. (L)	98.68	98.68	100.00	98.68	100.00	81.58	73.68
PR Covr. (L)	97.50	97.50	98.75	96.25	95.00	88.75	75.00
Comp. Lg. Covr. (L)	100.00	100.00	100.00	98.75	100.00	86.25	73.75

Note: Prop. N., indicates Proposed Nonpenalized, Prop. NU., indicates Proposed Nonpenalized Unified method, Prop. NP., indicates Proposed Penalized, Prop. NPU., indicates Proposed Penalized Unified method, Prop. Log., indicates Parametric Logistic of Chowdhury (2021a), Bayesian Probit indicates the Bayesian Latent Probit, and Pen. Logit indicates the maximum likelihood Penalized Logistic regression. This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000, 2000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 20 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 60 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 180 parameters per DGP for a total of 540 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets.

Table 4.2: Simulation Confidence Interval Range for All DGPs (at 1% Significance Level)

	Prop. N	Prop. NU	Prop. NP	Prop. NPU	Prop. Log.	Bayesian Probit	Pen. Logit
LGR Covr. (NL)	6.01	5.76	5.85	5.43	5.58	1.79	5.40
PR Covr. (NL)	5.92	6.10	5.46	5.60	5.15	2.07	5.22
Comp. Lg. Covr. (NL)	5.64	5.69	5.53	5.29	5.58	1.90	5.58
LGR Covr. (Mx.)	6.47	5.67	5.37	6.19	5.67	2.20	6.52
PR Covr. (Mx.)	5.95	5.75	5.29	5.46	5.40	2.13	6.01
Comp. Lg. Covr. (Mx.)	6.1	5.46	5.74	5.74	5.34	2.03	5.24
LGR Covr. (L)	5.71	5.65	5.67	5.73	5.61	1.86	6.13
PR Covr. (L)	5.74	5.43	5.49	5.60	5.42	2.01	6.45
Comp. Lg. Covr. (L)	5.92	5.58	5.60	5.49	5.36	1.87	6.25

Note: Prop. N., indicates Proposed Nonpenalized, Prop. NU., indicates Proposed Nonpenalized Unified method, Prop. NP., indicates Proposed Penalized, Prop. NPU., indicates Proposed Penalized Unified method, Prop. Log., indicates Parametric Logistic of Chowdhury (2021a), Bayesian Probit indicates the Bayesian Latent Probit, and Pen. Logit indicates the maximum likelihood Penalized Logistic regression. This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000, 2000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 20 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 60 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 180 parameters per DGP for a total of 540 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets.

Table 4.3: Simulation Summary of ARS for All DGPs

	Prop. N	Prop. NU	Prop. NP	Prop. NPU	Prop. Log.	Bayesian Probit	Neural Net
Non-Linear	0.07	0.10	0.07	0.05	0.23	0.25	0.16
Mixed	0.17	0.05	0.07	0.07	0.23	0.27	0.21
Linear	0.08	0.07	0.05	0.05	0.19	0.23	0.21

Note: Prop. N., indicates Proposed Nonpenalized, Prop. NU., indicates Proposed Nonpenalized Unified method, Prop. NP., indicates Proposed Penalized, Prop. NPU., indicates Proposed Penalized Unified method, Prop. Log., indicates Proposed Logistic, Bayesian Probit indicates the Bayesian Latent Probit, and Pen. Logit indicates the maximum likelihood Penalized Logistic regression. This is a summary over all three DGPs (Logistic, Probit and Complementary Log-Log), run over sample sizes of n = {100, 500, 1000, 2000} and unbalancedness of {0.1, 0.2, 0.3, 0.4, 0.5} for all linear, non-linear and mixed models fitted (here 0.5 indicates equal number of 1's and 0's (balanced), 0.4 indicates 10% fewer 1's than 0's and so forth). For each DGP there are 20 different datasets to consider for each of the linear, mixed and non-linear models considered for a total of 60 different datasets per DGP. For each dataset there are three parameters of interest or β's. In total there are 180 parameters per DGP for a total of 540 parameters to be estimated over the entire simulation study. The results are summarized by average over all simulated datasets.

4.4 Empirical Application

To compare and contrast the methodology with those presented in the previous chapters, I apply it to various datasets below. The datasets include the Intoxication dataset as well as the Higgs dataset used in Chapter 3, in addition to the Challenger space shuttle disaster dataset of 1986. Below more detailed explanations are given for these datasets avoiding duplication where possible.

4.4.1 Detecting Heavy Drinking Events Using Smartphone Data

To detect heavy drinking events using smartphone accelerometer data in Killian et al. (2019) as in Chapter 3, I run the unified methodology in a penalized and unpenalized setting as in Chapter 3. For completeness note that the authors identified heavy drinking events within a four second window of their measured variable of Transdermal Alcohol Content (TAC) on smartphone accelerometer data. Their best classifier was a Random Forest with about 77.50% accuracy. A similar analysis was done on a far simpler model of TAC readings against the accelerometer readings as predictors, for all subject's phone placement in 3D space, for the x, y and z axes,

$$TAC = Intercept + x - axis\ reading + y - axis\ reading + z - axis\ reading. \quad (4.62)$$

TAC here was set to 1 if the measurement was over 0.08 and 0 otherwise as in Chapter 3. The same four second time window of accelerometer readings were used in the analysis with the assumption that the TAC readings were unlikely to change in such a small time interval. Please recall the results were extremely encouraging for the application in Chapter 3, with perfect TeD (20% of the data) ARS classification accuracy, with $1,000$ iterations and 500 burn-in period. The penalized application also had perfect classification accuracy in TeD,

which seemed to directly contradict the conventional wishdom of the need to perform model selection and inference separately. These seemingly excellent results would suggest that the data align well with the nonunified methodology. Nevertheless, an application of the current methodology showed it to be equally effective again with only $1,000$ iterations.

In particular, in the current application all methodologies were run for $1,000$ iterations as in Chapter 3, and below I present the convergence and histogram plots in Figure 4.1, Figure 4.2, Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7, and Figure 4.8 for both the unified and nonunified methodologies. The results tell a fascinating story regarding

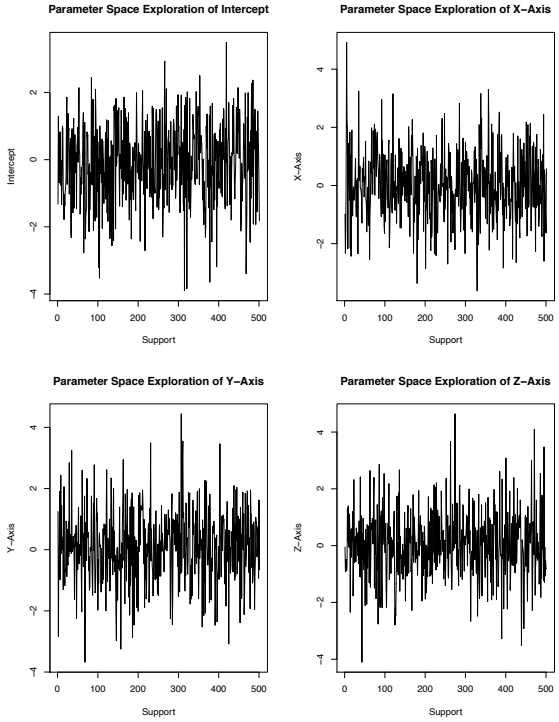


Figure 4.1: Heavy Drinking Event Data Sample Space Exploration Plot for Nonparametric Unified Methodology.

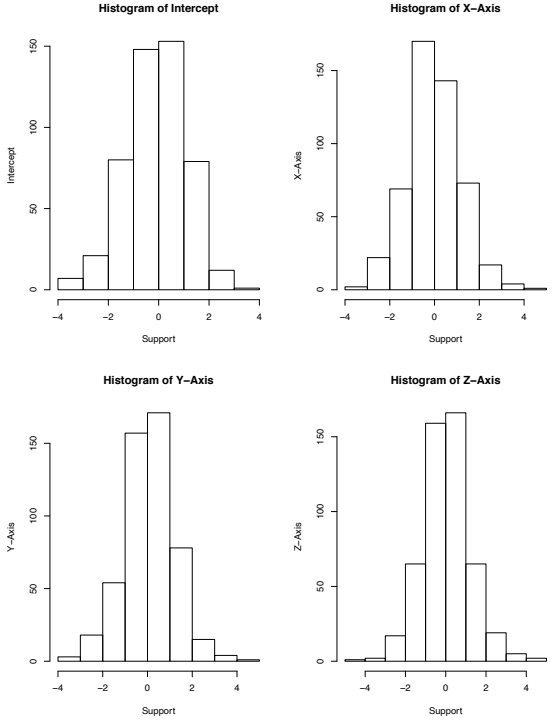


Figure 4.2: Heavy Drinking Event Data Histogram of Parameters for Nonparametric Unified Methodology.

MIPs. The TeD classification results for the nonunified methodology were excellent, with perfect identification. For the TrD, however, the results were less effective with around 70.00% accuracy. This trend though was largely consistent with the results from the unified methodology also giving perfect classification in TeD with the TrD classification results

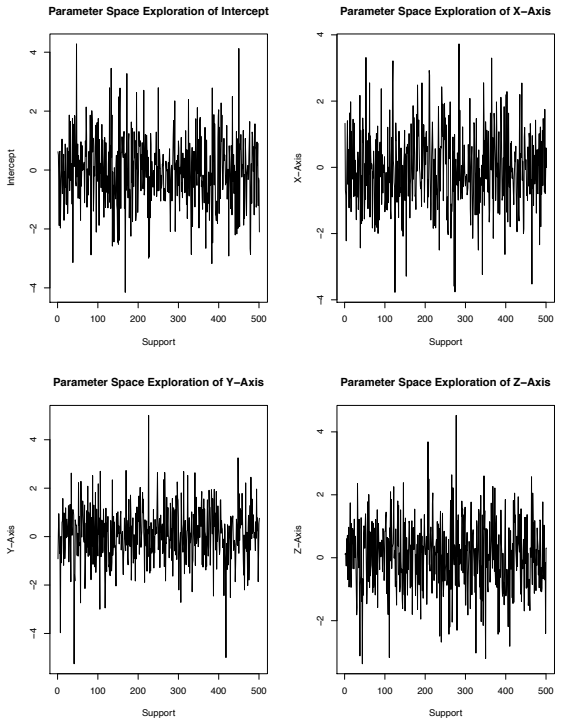


Figure 4.3: Heavy Drinking Event Data Sample Space Exploration Plot for Nonparametric Penalized Unified Methodology.

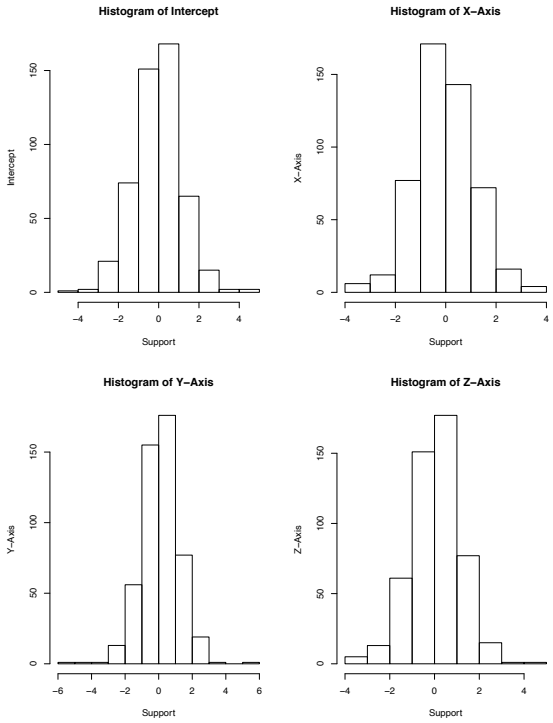


Figure 4.4: Heavy Drinking Event Data Histogram of Parameters for Nonparametric Penalized Unified Methodology.

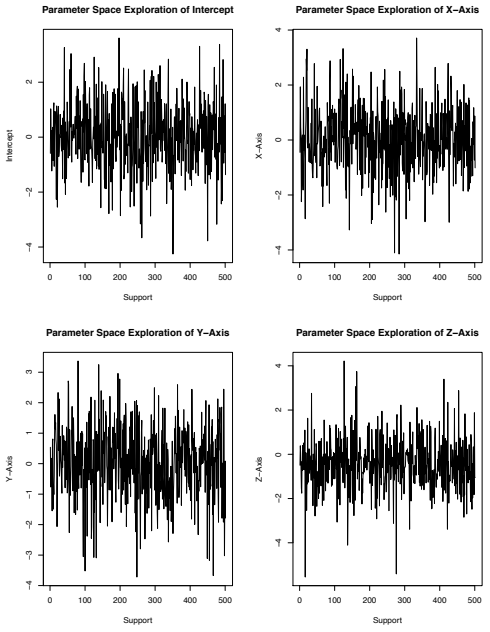


Figure 4.5: Heavy Drinking Event Data Sample Space Exploration Plot for Nonparametric Methodology.

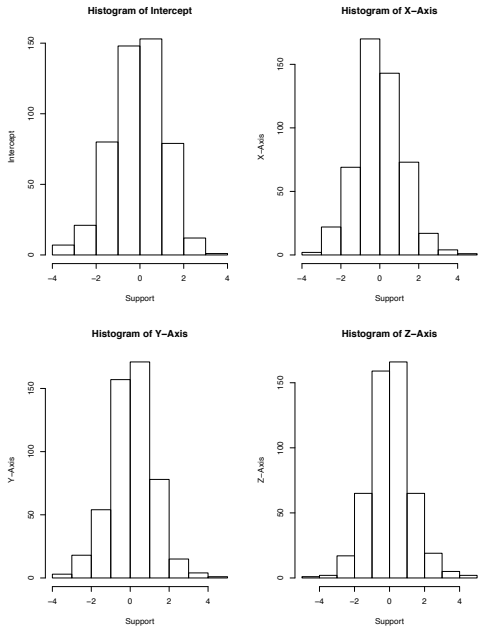


Figure 4.6: Heavy Drinking Event Data Histogram of Parameters for Nonparametric Methodology.

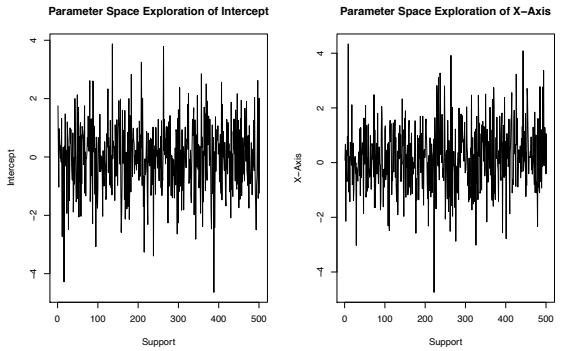


Figure 4.7: Heavy Drinking Event Data Sample Space Exploration Plot for Nonparametric Penalized Methodology.

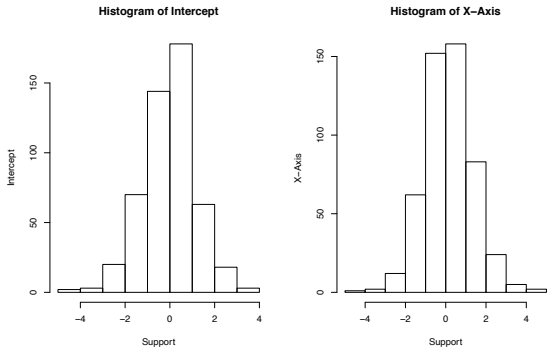


Figure 4.8: Heavy Drinking Event Data Histogram of Parameters for Nonparametric Penalized Methodology.

slightly worse, though not significantly so. In regards to inference, the results were largely

Table 4.4: Intoxication Dataset Summary of ARS for All Relevant Methodologies

Methodology	TrD	TeD
Unified Penalized	0.21	0.06
Penalized Nonparametric	0.34	0.00
Nonparametric	0.30	0.00
Unified Nonparametric	0.24	0.00
Parametric	0.68	0.76
Existing Bayes	0.77	0.66
MLE Logistic	0.91	0.90
Penalized Logistic	0.91	0.90

Table 4.5: Intoxication Dataset Summary of AIC for All Relevant Methodologies

Methodology	TrD	TeD
Unified Penalized	1.32	0.22
Penalized Nonparametric	3.16	0.95
Nonparametric	3.13	0.94
Unified Nonparametric	1.77	1.07
Parametric	1.31	0.97
Existing Bayes	1.83	1.18
MLE Logistic	1.21	1.00
Penalized Logistic	1.21	1.00

identical, thus providing further verification for the applicability of the Likelihood Principle, for both the unified and nonunified nonparametric cases. This is a direct result of the linear operators in both methodologies being continuous, as defined from the sample space

Table 4.6: Intoxication Dataset Parameter Summary for All Relevant Methodologies

Predictor	Estimates	CI-Low	CI-High	Methodology
Intercept	0.24**	0.01	0.47	(1)
X-axis	0.02	-0.22	0.25	(1)
Y-axis	0.03	-0.19	0.25	(1)
Z-axis	-0.54**	-0.83	-0.26	(1)
Intercept	-0.06**	-0.11	-0.01	(2)
X-axis	-0.01	-0.05	0.04	(2)
Y-axis	0.17**	0.13	0.21	(2)
Z-axis	-0.08**	-0.13	-0.02	(2)
Intercept	0.22	-0.03	0.48	(3)
X-axis	-0.07	-0.31	0.17	(3)
Y-axis	0.21	-0.04	0.46	(3)
Z-axis	-0.82**	-1.05	-0.59	(3)
Intercept	0.17	-0.07	0.42	(4)
X-axis	-0.02	-0.27	0.23	(4)
Y-axis	0.16	-0.06	0.38	(4)
Z-axis	0.02	-0.2	0.24	(4)
Intercept	-0.13	-0.3	0.05	(5)
X-axis	0.01	-0.19	0.2	(5)
Y-axis	0.07	-0.13	0.27	(5)
Z-axis	-0.21**	-0.37	-0.05	(5)
Intercept	-0.01	-0.14	0.11	(6)
X-axis	-0.12	-0.32	0.08	(6)
Y-axis	0.24**	0.06	0.43	(6)
Z-axis	-0.02	-0.15	0.1	(6)
Intercept	-0.87***	-0.9	-0.85	(7)
X-axis	-0.04*	-0.09	0	(7)
Y-axis	0.17***	0.11	0.23	(7)
Z-axis	0.00***	0.00	0.00	(7)
Intercept	-0.87***	-0.9	-0.84	(8)
X-axis	-0.04*	-0.11	0.02	(8)
Y-axis	0.17***	0.09	0.25	(8)
Z-axis	0.00***	0.00	0.00	(8)

Note: (1) Nonprametric, (2) Unified Nonparametric, (3) Penalized Nonparametric, (4) Unified Penalized Nonparametric, (5) Parametric, (6) Existing Bayesian, (7) MLE Logistic, (8) Penalized Logistic.

to the link subspace. However, the proposed methodology has significant advantages over the nonunified case in terms of model fit. That is considering the median $\boldsymbol{\beta}s$, with the Binomial Likelihood (LAHEML integrates out the probability of success so the original

data likelihood is used here), the unified methodology was nearly 2.4 times better in the TrD and almost 4.3 times as good in the TrD for the Proposed Unified Nonparametric methodology over the Proposed Nonparametric methodology. The results are also consistent for the penalized applications. These results are entirely consistent with the underlying mathematical foundations, and I discuss them more in the forthcoming Section 4.5. Please further note that the goal here is to compare the unified and nonunifed applications and not the other models per se[1].

[1]The Bayesian Latent Probit was run for $5,000$ iterations whereas the Proposed Parametric Logistic was run for only $1,000$ iterations here. Please refer to Chowdhury (2021a) for a more in depth comparison of those models.

4.4.2 Exotic Particle Detection Using Particle Accelerator Data

The second application of the methodology was for identifying high-energy particles in Physics (Baldi et al. (2014)) as in Chapter 3. Recall that there are 28 feature sets in the paper, with the first 21 features the kinematic properties measured by detectors in particle accelerators. The last 7 high-level features were derived from the first 21 to discriminate between the two classes. Therefore, inference is not the primary purpose for this application. The classes to be identified is either 0 and 1, and refer to noise and signal respectively before. For completeness, the model specification is restated below.

$$Signal/Noise = Intercept + \sum_{i=1}^{28} Feature_i. \tag{4.63}$$

As before, for more information on the actual feature sets I refer the reader to the original paper. Given the large datasize, for these applications, LAHEML was run for only $1,000$ iterations with 500 burn-in period. The convergence plots, along with the histograms of each parameter may be found in Figure 4.9, Figure 4.10, Figure 4.11, Figure 4.12, Figure 4.13, Figure 4.14, Figure 4.15 and Figure 4.16. The penalized and unpenalized estimation formulations were identical to that for the Intoxication application for Biostatistics. As such, the classification outcomes were extremely encouraging, and can be found in Table 4.7. The AICs can be found in Table 4.8.

Table 4.7: Higgs Dataset Parameter Summary for All Relevant Methodologies

Methodology	TrD	TeD
Unified Penalized	0.41	0.18
Penalized Nonparametric	0.48	0.19
Nonparametric	0.51	0.19
Unified Nonparametric	0.53	0.16
Parametric	0.77	0.77
Existing Bayes	0.67	0.67
MLE Logistic	0.58	0.58

Table 4.8: Higgs Dataset Summary of AIC for All Relevant Methodologies

Methodology	TrD	TeD
Unified Penalized	3.68	1.48
Penalized Nonparametric	4.11	1.65
Nonparametric	3.83	1.57
Unified Nonparametric	3.51	1.36
Parametric	1.36	1.36
Existing Bayes	0.83	0.75
MLE Logistic	1.28	1.28

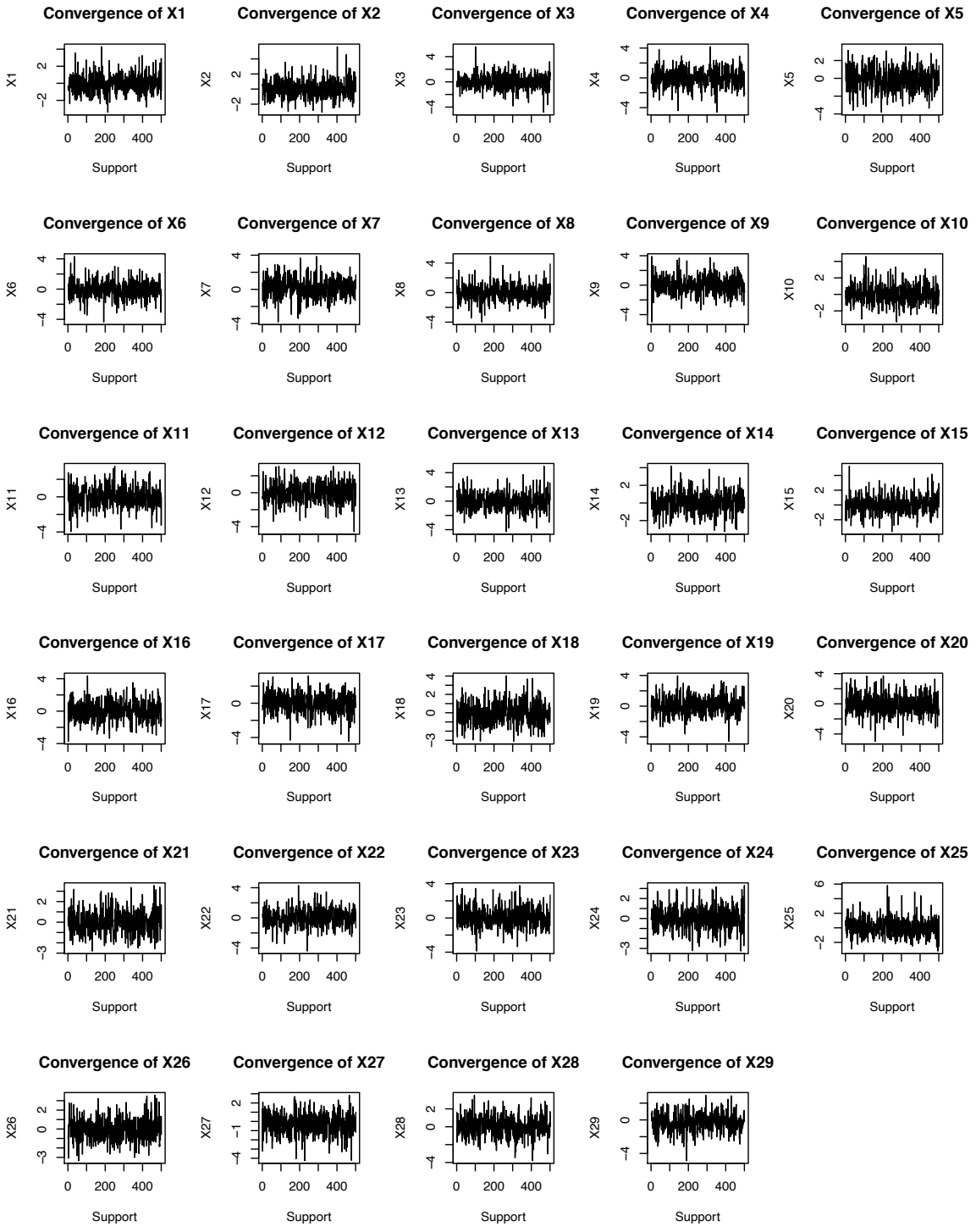


Figure 4.9: Exotic Particle Detection Data Sample Space Exploration Plot for Nonparametric Unified Methodology.

Given the small number of iterations one would expect that these MIP results can be further improved for the unified applications for penalized or unpenalized cases. However, even in

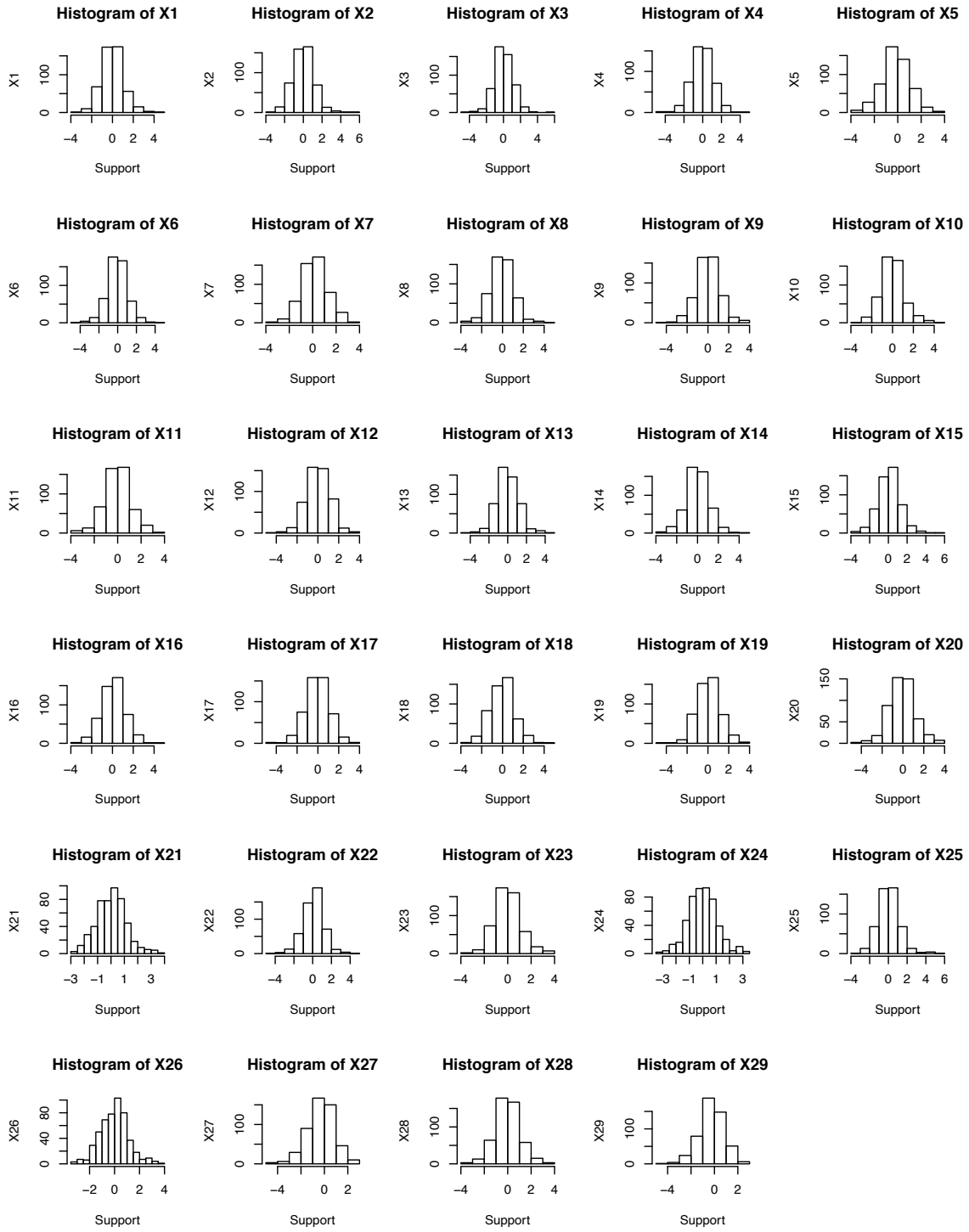


Figure 4.10: Exotic Particle Detection Data Histogram of Parameters for Nonparametric Unified Methodology.

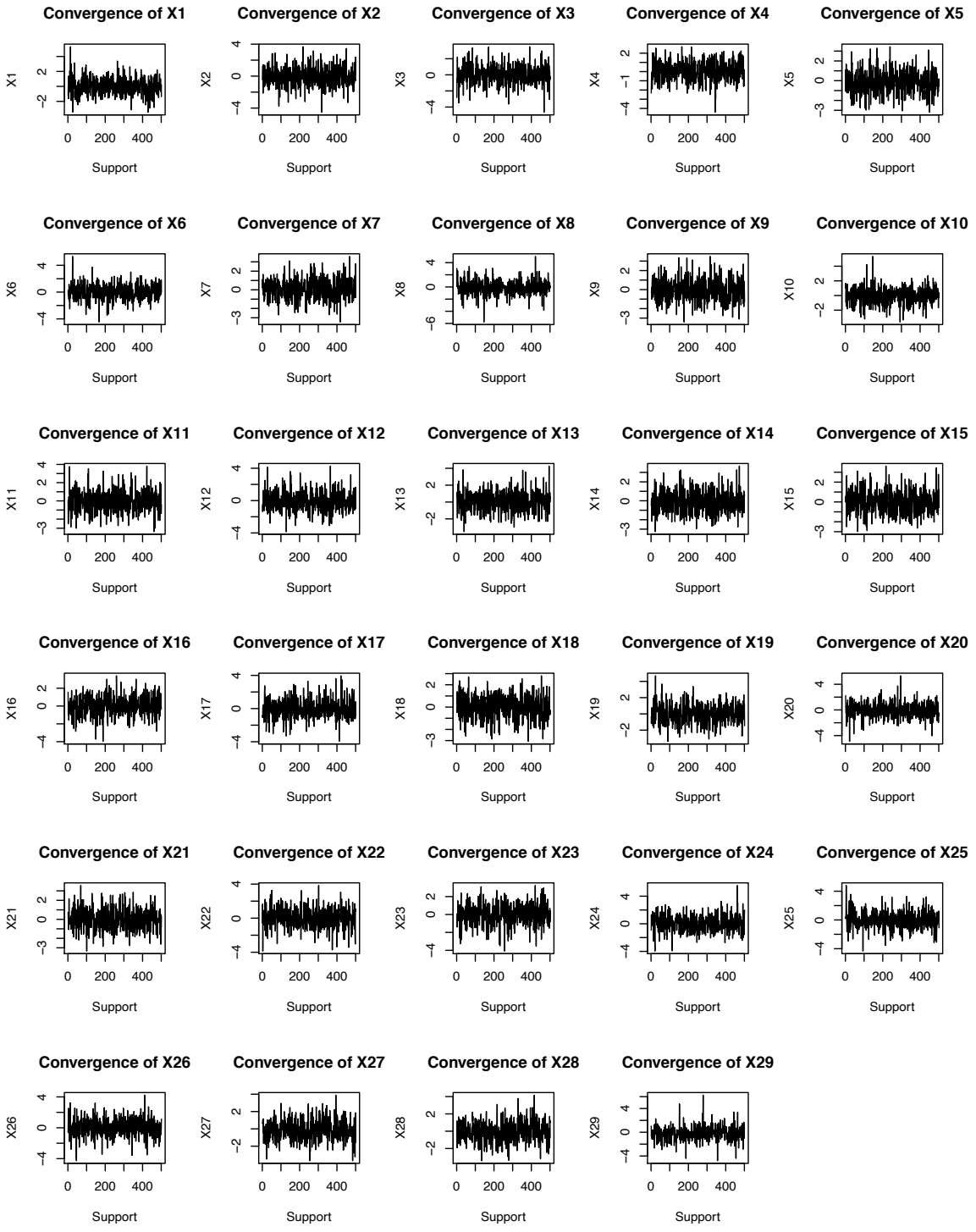


Figure 4.11: Exotic Particle Detection Data Sample Space Exploration Plot for Nonparametric Penalized Unified Methodology.

this small number of iterations, the Unified Penalized and Unified Nonparametric methodologies outperformed their nounified counterparts in TeDs for both ARS and AIC. Thus, it

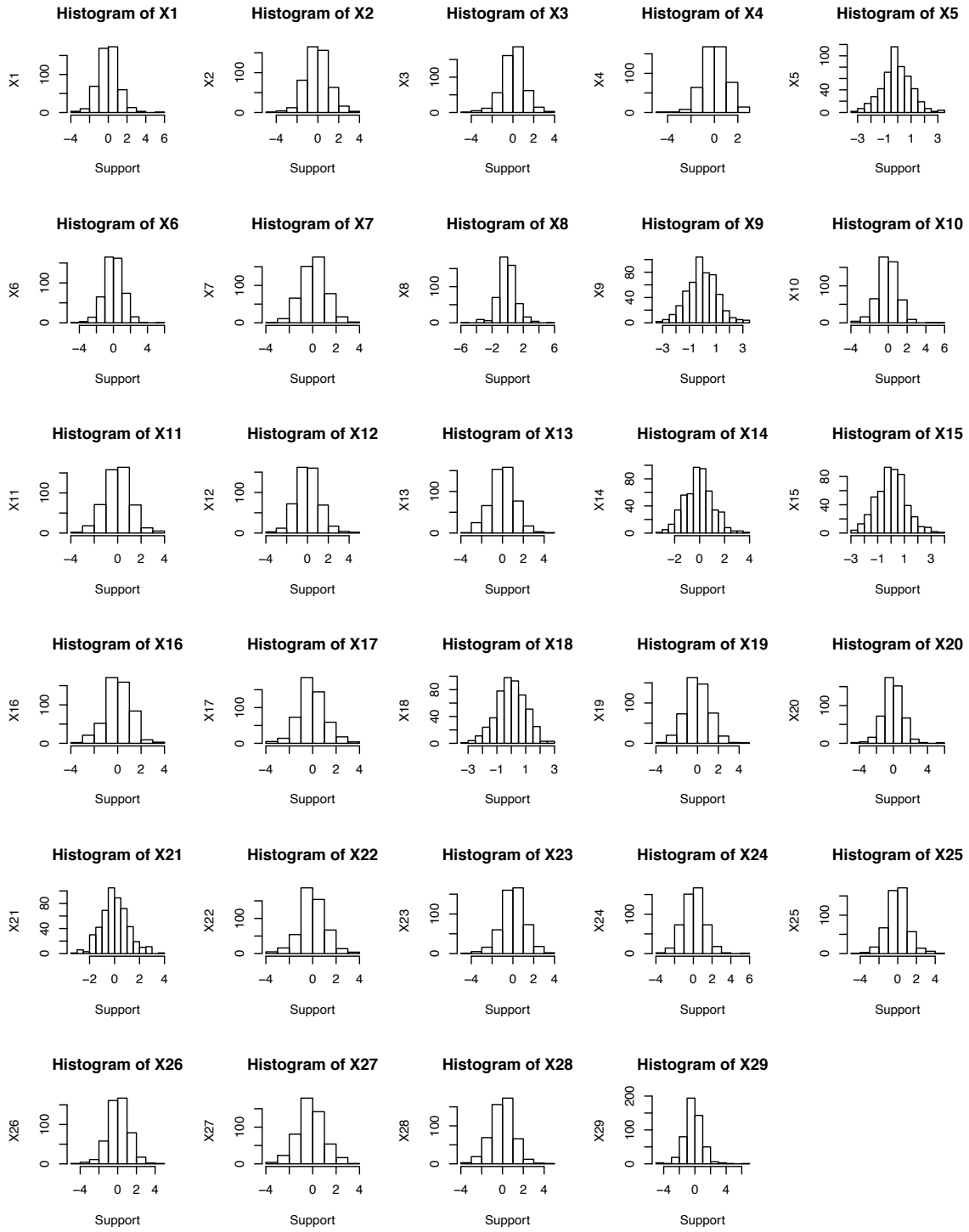


Figure 4.12: Exotic Particle Detection Data Histogram of Parameters for Nonparametric Penalized Unified Methodology.

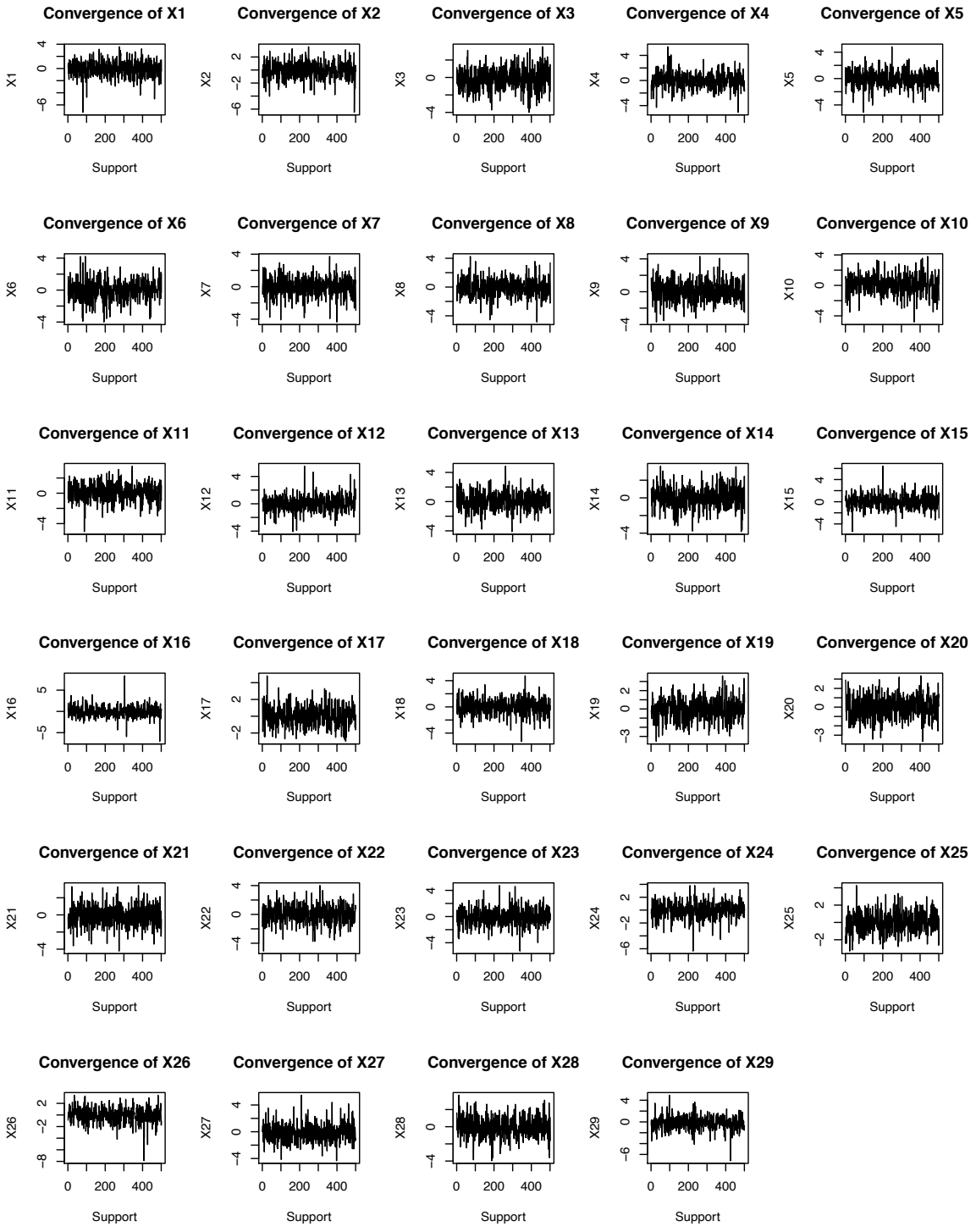


Figure 4.13: Exotic Particle Detection Data Sample Space Exploration Plot for Nonparametric Methodology.

seems reasonable to surmize that with the same number of iterations the unified methodolo-

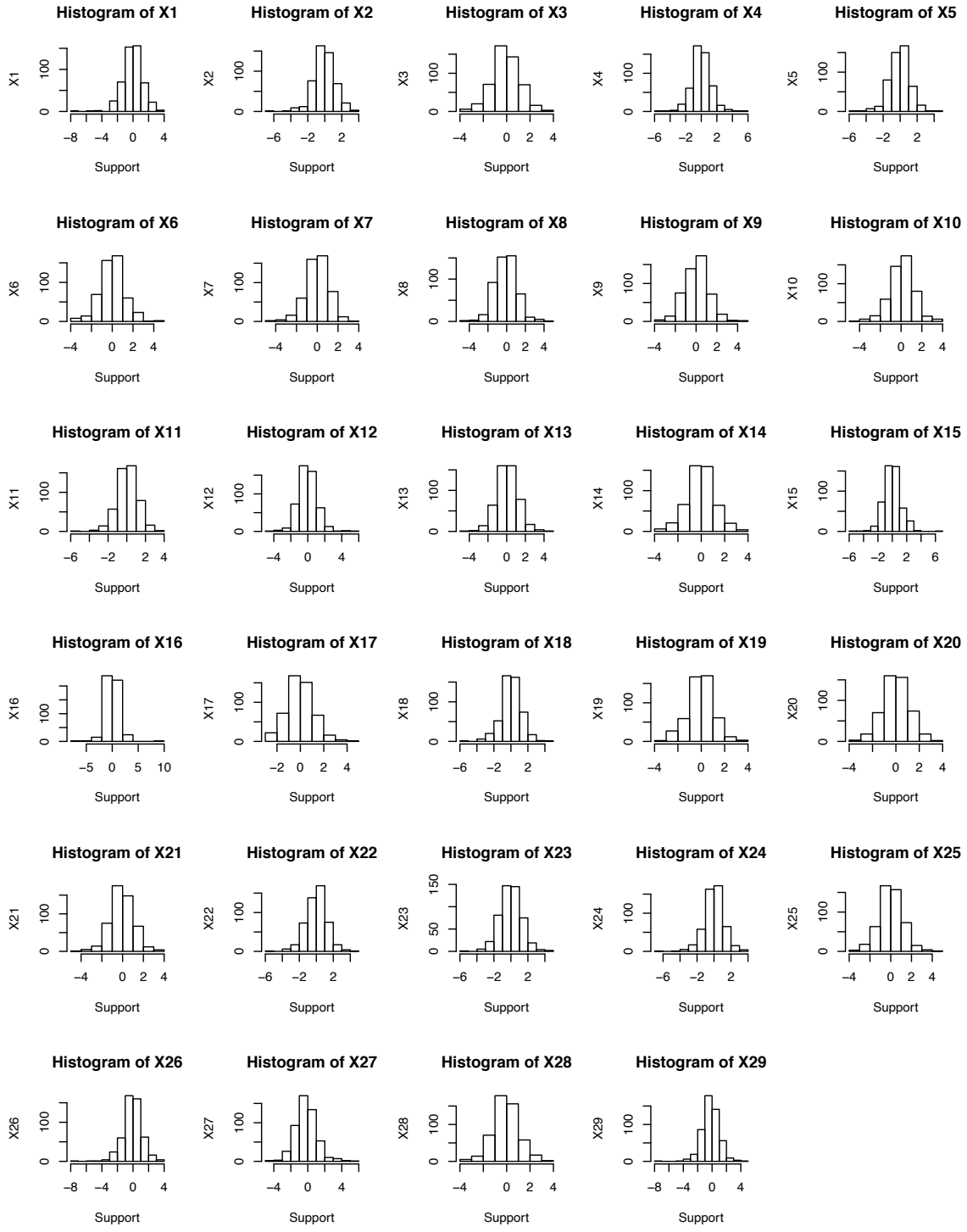


Figure 4.14: Exotic Particle Detection Data Histogram of Parameters for Nonparametric Methodology.

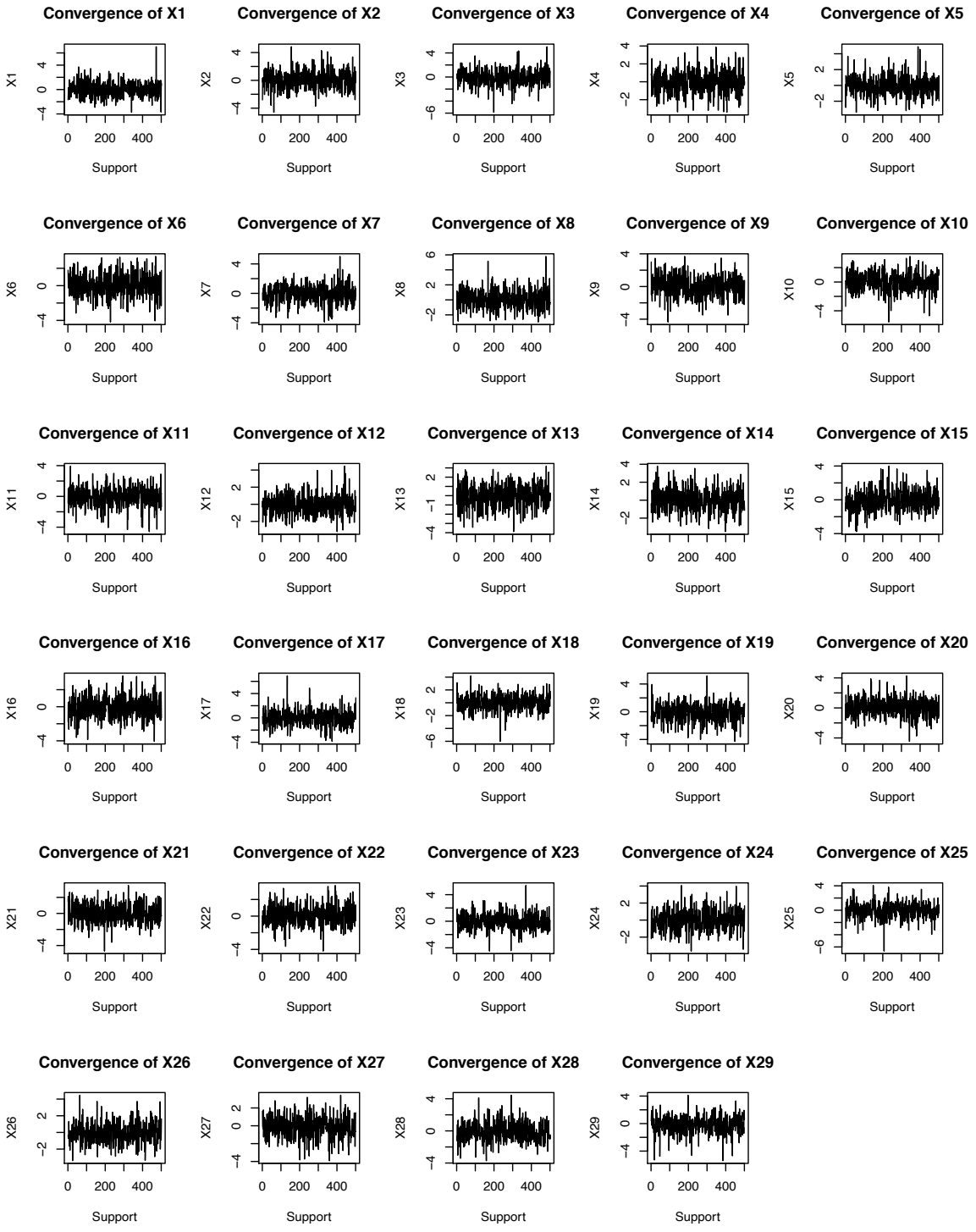


Figure 4.15: Exotic Particle Detection Data Sample Space Exploration Plot for Nonparametric Penalized Methodology.

gies would outperform the results of the nonunified applications in Chapter 3[2]. As before,

[2]Again the Bayesian Latent Probit was run for $5,000$ iterations, while the Proposed Parametric application

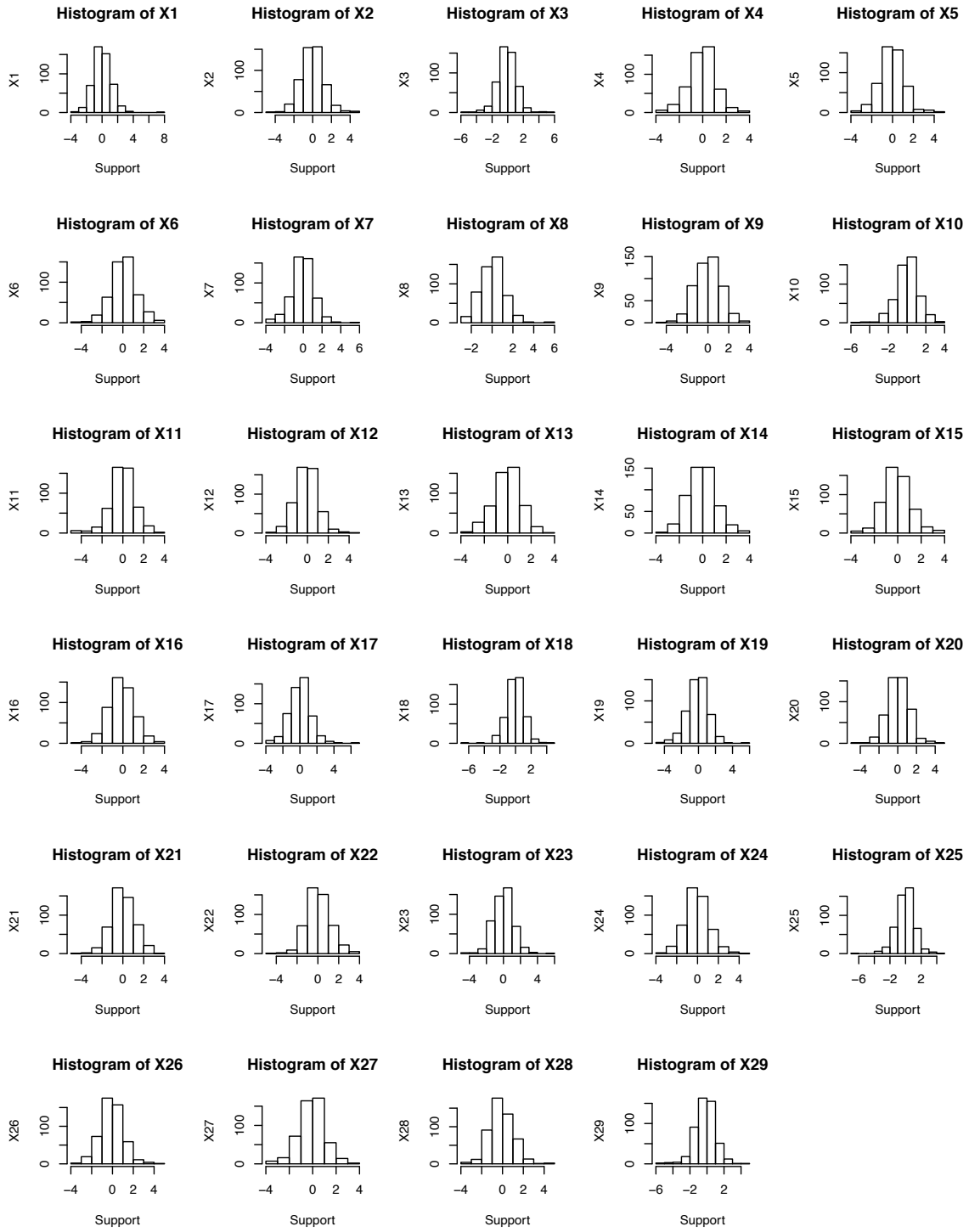


Figure 4.16: Exotic Particle Detection Data Histogram of Parameters for Nonparametric Penalized Methodology.

the goal here is not to compare the performances of the published parametric versions, but rather the Proposed Unified Nonparametric and Proposed Nonparametric versions in both the penalized and unpenalized applications. As such, I discuss these findings along with their mathematical implications more broadly in Section 4.5.

4.4.3 Challenger Disaster

The final application of the methodology is for the Challenger disaster. This dataset holds a special place in my memory, as it was the first dataset on which I saw categorical models, applied. In fact, that model was the Logistic, which the Chapters preceding this, and of course this one as well, improves in regads to MIPs. Thus, it seems fitting to end with an application on this dataset that started my initial curiosity into such models. To give some background information, note that the Challenger Space Shuttle explosion occured in 1986 due to the failure of an O-ring, a component on the rocket. It is now widely recognized that the material this component is made of, is susceptible to stress especially when the outside temperature is low. There are certain engineering reasons for this, which are not important for the current discussion presently, and I refer the reader to Draper (1995) for further discussions on this. In addition, unlike in other papers that use the same dataset (see for example Draper (1995)) I change the task of interest slightly here, from understanding the number of O-rings under thermal stress to understanding the probability of an O-ring experiencing thermal stress as a function of the outside temperature and a variable called leak-check pressure. Thus, the model may be given as,

$$O-ring\,Under\,Stress\;=\;Intercept+Temp.+Log(\sqrt{Leak-Check\;Pressure}). \quad (4.64)$$

To understand the probability that an O-ring will experience thermal distress at a temper-

was run for only $1,000$ iterations. Thus, the results here are not directly comparable to those in Chapter 2.

ature of 31 degrees Farenhite, an analysis was done on the O-rings experiencing thermal distress for the 23 shuttle launches prior to the Challenger disaster[3]. Thus, I seek to extrapolate the probability of stress, and therefore failure of an O-ring as a function of this temperature. Though the dataset is extremely small, the TrD consisted of the first 18 observations and TeD comprised of the rest.

The results are interesting in that all models compared gave perfect TeD classification. However, the unified methods outperformed the Nonparametric Penalized application in TrD and matched the Nonparametric application in TrD. In regards to AIC, however, the Nonparametric Penalized application had the best TeD result, but the unified methodologies again outperformed the nonunified methods on average over both TrD and TeD combined. Once again, as for the other dataset applications, the main goal here is to compare the unified and nonunified methodologies, and as such the comparisons for other methodologies are not explicitly considered, though the results are consistent with previous findings for these applications as well. Note that as before the Parametric Logistic is run for only $1,000$ iterations but the Bayesian Latent Probit is run for $5,000$ iterations, as such, these results are not directly comparable. I refer the reader to Chowdhury (2021a) for more discussions on this. Finally, in regards to inference, each of the unified methodologies were consistent in finding both temperature and pressure as significant, but not the intercepts. In contrast, the Proposed Nonparametric methodology found the intercept as significant as well as temperature and pressure, but the Proposed Penalized Nonparametric methodology found all except the intercept as significant. Accordingly, the results of the unified methodologies are again more consistent than the other models compared. More discussions on this and its implications are given in the next section 4.5.

[3]The launch temperature on the day of the Challenger disaster was 31 degrees Farenhite.

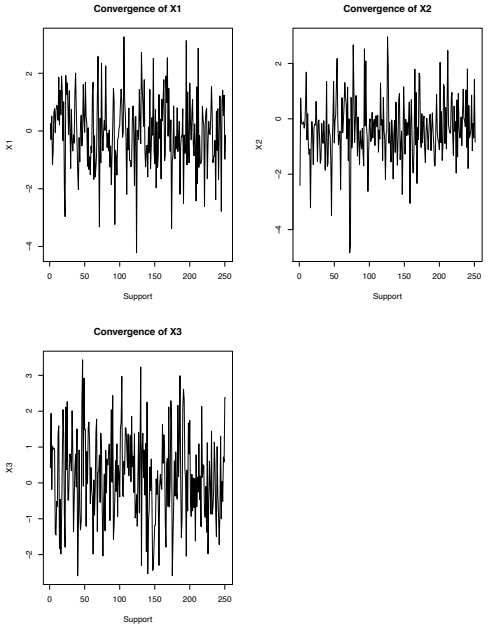


Figure 4.17: Sample Challenger Sample Space Exploration Plot for Nonparametric Methodology.

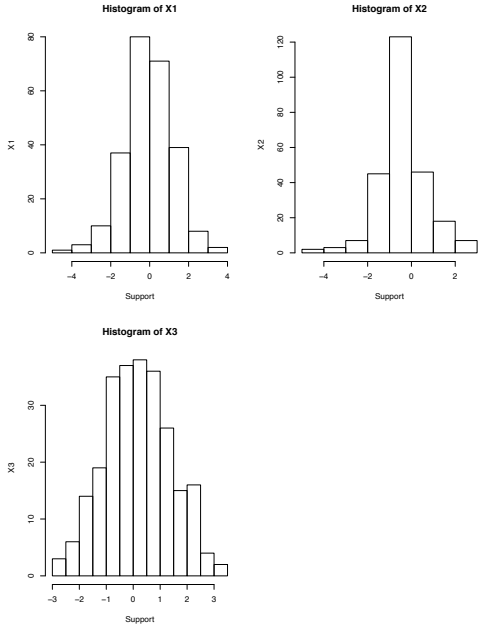


Figure 4.18: Sample Challenger Histogram of Parameters for Nonparametric Methodology.

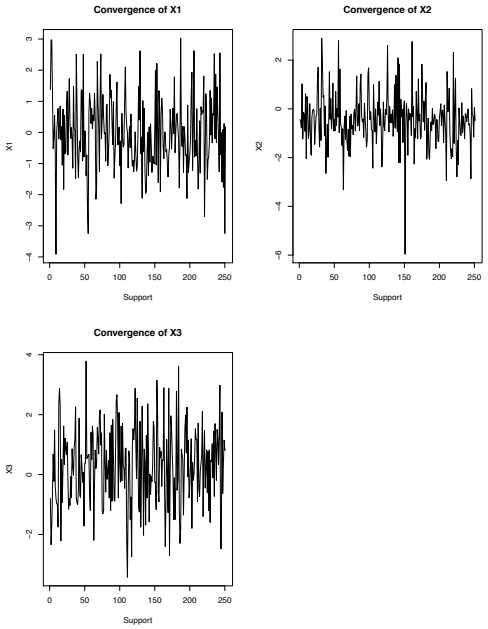


Figure 4.19: Sample Challenger Sample Space Exploration Plot for Nonparametric Penalized Methodology.

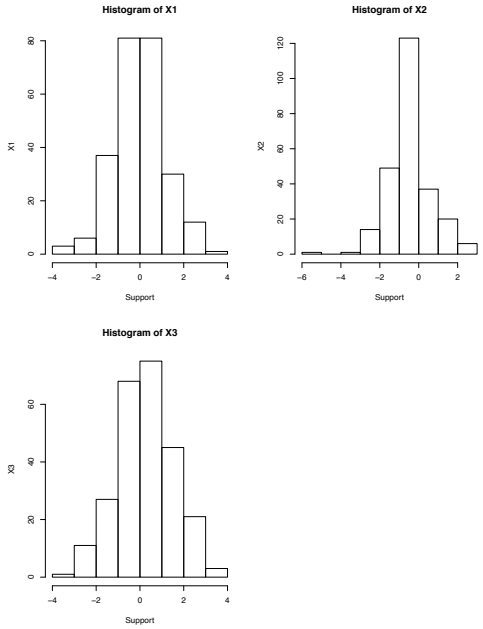


Figure 4.20: Sample Challenger Histogram of Parameters for Nonparametric Penalized Methodology.

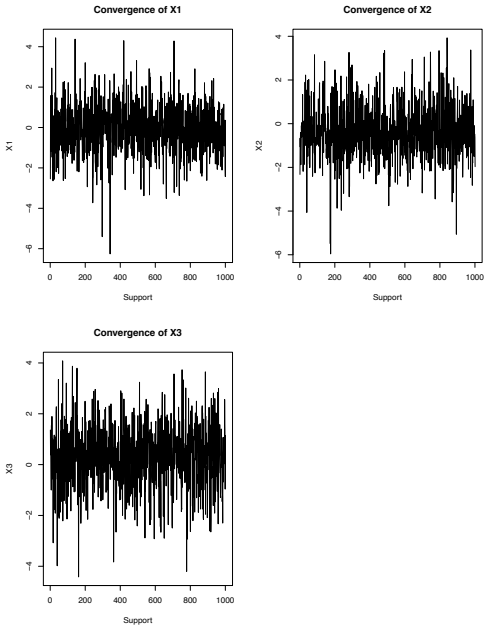


Figure 4.21: Challenger Sample Space Exploration Plot for Nonparametric Unified Methodology.

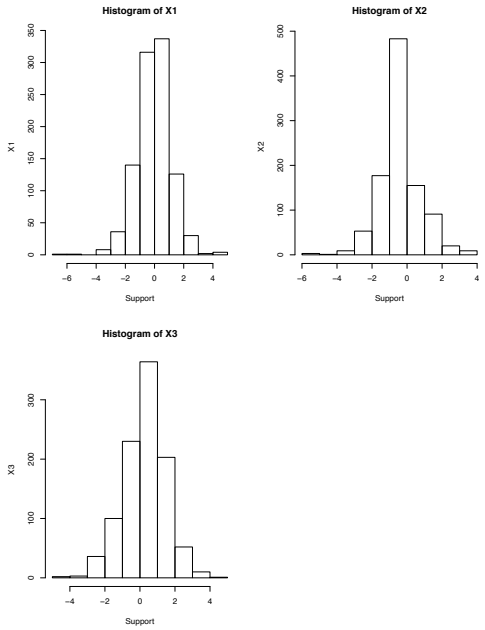


Figure 4.22: Challenger Histogram of Parameters for Nonparametric Unified Methodology.

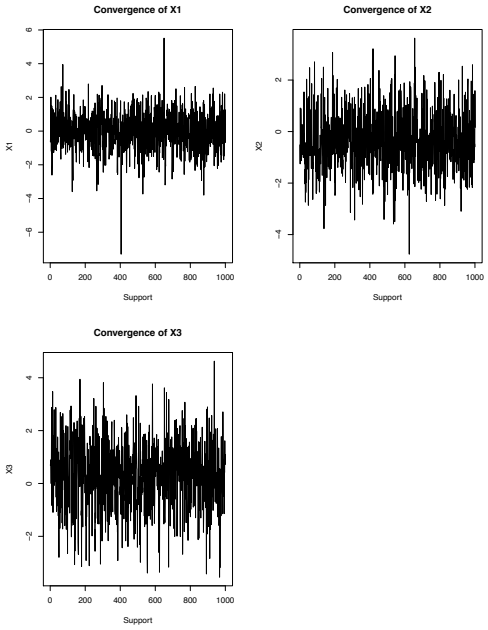


Figure 4.23: Challenger Sample Space Exploration Plot for Nonparametric Penalized Unified Methodology.

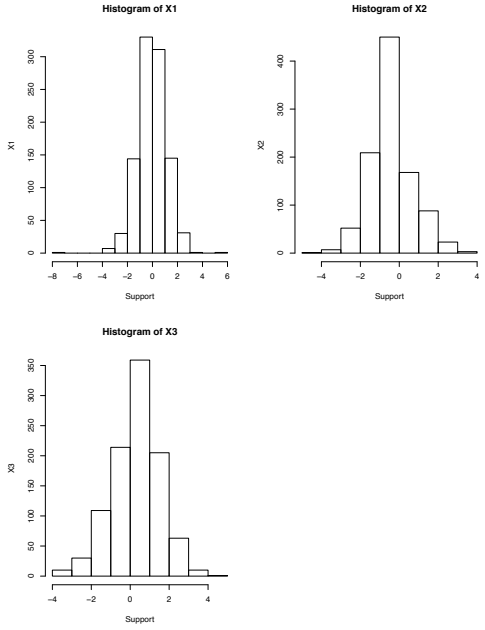


Figure 4.24: Challenger Histogram of Parameters for Nonparametric Penalized Unified Methodology.

Table 4.9: Challenger Dataset Summary of ARS for All Relevant Methodologies

Methodology	TrD	TeD
Unified Penalized	0.00	0.00
Penalized Nonparametric	0.33	0.00
Nonparametric	0.00	0.00
Unified Nonparametric	0.00	0.00
Parametric	0.34	0.00
Existing Bayes	0.41	0.00
MLE Logistic	0.41	0.00
Penalized Logistic	0.41	0.00

Table 4.10: Challenger Dataset Summary of AIC for All Relevant Methodologies

Methodology	TrD	TeD
Unified Penalized	0.96	1.25
Penalized Nonparametric	3.16	0.95
Nonparametric	2.20	1.42
Unified Nonparametric	1.02	1.25
Parametric	1.43	2.41
Existing Bayes	2.63	2.57
MLE Logistic	1.09	2.06
Penalized Logistic	1.24	2.50

Table 4.11: Challenger Dataset Parameter Summary for All Relevant Methodologies

Predictor	Estimates	CI-Low	CI-High	Methodology
Intercept	0.11	-0.3	0.52	(1)
Temperature	-0.4**	-0.73	-0.08	(1)
Pressure	-0.49**	-0.81	-0.18	(1)
Intercept	-0.06**	-0.12	-0.01	(2)
Temperature	-0.54**	-0.59	-0.5	(2)
Pressure	0.34**	0.28	0.40	(2)
Intercept	-0.15	-0.48	0.18	(3)
Temperature	-0.55**	-0.86	-0.24	(3)
Pressure	0.42**	0.05	0.79	(3)
Intercept	0.06	-0.09	0.21	(4)
Temperature	-0.45 **	-0.60	-0.31	(4)
Pressure	0.48**	0.3	0.65	(4)
Intercept	-0.27**	-0.42	-0.12	(5)
Temperature	-0.45**	-0.62	-0.28	(5)
Pressure	-0.35**	-0.49	-0.21	(5)
Intercept	-1.52**	-1.64	-1.4	(6)
Temperature	-0.63**	-0.66	-0.59	(6)
Pressure	0.67**	0.60	0.74	(6)
Intercept	-4.5	-14.57	5.56	(7)
Temperature	-1.97*	-4.31	0.38	(7)
Pressure	1.35	-2.86	5.56	(7)
Intercept	-3.65	-17.89	7.2	(8)
Temperature	-1.24**	-5.09	0.2	(8)
Pressure	1.11	-3.71	6.76	(8)

Note: (1) Nonprametric, (2) Unified Nonparametric, (3) Penalized Nonparametric, (4) Unified Penalized Nonparametric, (5) Parametric, (6) Existing Bayesian (7) MLE Logistic (8) Penalized Logistic.

4.5 Discussion

The results above are noteworthy in several contexts, including for Model Fit, Inference, and Prediction. Thus in the forthwith I will expand on the importance of these findings for our understanding of how statistical significance is related to scientific significance. However, before discussing these topics I would first like to highlight some of the mathematical results that allow us to make these conclusions that have broad impact across the sciences.

One of the most important results of this formulation is that the formulation ensures that mathematically a linear operator exists between the Hausdorff space and the link field. The claim is significant in that as sample size increases, in the current framework we cannot guarantee that the estimation process, whether Bayesian or Frequentist, will be bounded for all observations. Thus, though this occurs at the unbounded points in the Hausdorff space, we cannot assert that this occurrence will have measure 0. However, the formulation presented here ensures, under measure theoretic foundations, using novel analytic results, that the estimation process will be continuous. Moreover, it assures that such a linear operator will strongly converge to the true parameters. LAHEML therefore ensures that as a function of the model specification and the observed Xs, convergence of the parameters of interest will be almost sure. This is an expansive result which many existing convergence methodologies cannot claim, because in their estimation the assertion may be violated. This has some rather strong implications for our continuing discussion on statistical significance and its relation to scienctific significance, and I will expand on this more below.

Another surprising result is that though the formulation may seem categorcal in nature, the results of Chapter 3 ensure that as the number of observations increase it can also be applicable to continuous outcome models. The mathematical results of Chapter 4 reinforce those results. Thus, as almost sure convergence results are for general functional forms, they are also valid and applicable for continuous outcome data. As such, the methodologies pre-

sented above overcome one of the principle shortcomings of mathematical model estimations present in the sciences for many decades.

Indeed, in many ways the formulation is not intuitive at first glance. Since, it would seem counterintuitive to use a signed measure to estimate functional forms over $\sigma - algebras$ that are a function of probability distributions. Yet the locally compact Hausdorff formulation ensures the most general topological circumstances under which such a construction remains valid. In fact, it is easy to see that in existing latent variable formulations, in the absence of continuity of the linear operator, the conditions for convergence do not hold in Tanner and Wong (1987). This is because, in such a case the underlying functional specifications are not equicontinuous. Thus, it is nested within the present formulation and its implementation through LAHEML.

Evidently, the use of a measure theoretic approach rooted in functional and real analysis thus have demonstrable advantages over the existing formulations. This is because, through it, we may consider the $L^p(X, \nu)$ spaces for $1 \leq p \leq \infty$, where as before $p = \infty$ is the essentially bounded case. As we saw in the proof of almost sure convergence above, using the locally compact Hausdorff property we are able to show that there exists a linear operator on the L^p spaces, which are also Banach spaces which are an isometric isomorphism to the space of bounded finitely additive signed measures. As such, we are able to apply the methodology on stronger Banach Spaces such as $L^{p=1}$ which contains the set of functions in the other L^p spaces. Consequently, it readily allows more flexibility in the potential functions that we may use to optimize over. On the other hand, positive finiteness almost everywhere, in conjunction with a locally compact Hausdorff property, implies that we may identify an unique signed measure using Riesz-Moarkov for every linear functional. Thus, LAHEML uses these existence and uniqueness properties to identify the unique linear operator that is a function of this unique signed-measure, thus explaining the excellent results above. Accordingly, above I extend Riesz-Markov Theorem to the general L^p spaces for all $1 \leq$

$p \leq \infty$, a result which appears to be new in this formulation in analysis, and it certainly is unique to Mathematical Statistics in a latent variable formulation especially in the Bayesian formulation.

Furthermore, this formulation has another remarkable property which goes beyond the independent and identically distributed model formulations which are the cornerstone of the sciences. To see this first note that through the use of Kass and Steffey (1989) we know that observations in LAHEML can be thought of as conditionally independent. Further note that in the unified formulation we showed that the sample space may be separated into countable unions of disjoint sets, over which we may define a subspace, such that a bounded finitely additive measure space exists over it. But a finitely additive measure may easily be converted to a distribution, and thus, for each disjoint set we may and do define a separate distribution. Consequently, these distributions need not be identical at all! Whether or not they are will depend on the extension of measure spaces through Hahn-Banach on the entire $\sigma - algebra$. Yet the proofs of the existence, uniqueness, and almost sure convergence results are not dependent on any particular distributional assumptions, identical or otherwise! Thus, this most general of formulations require our data to be neither independent nor identical!

Of course, as with any model specification, our characterization of the underlying phenomenon is a function of the observed Xs. In addition, there may be circumstances under which the assumptions in Kass and Steffey (1989) may not hold either. In such a case, there are a multitude of other methodologies which can be used in conjunction with the proposed methodologies accordingly. Regardless, the usefulness of the methodology and its general construction shows much potential for broad applicability across the science, including for MIPs.

Hopefully, it is clear that one of the chief contributions of this research is the realization that just because a model outperforms another in regards one of the MIPs, it does not imply that it will outperform it in another. Indeed, there are numerous examples above where

the opposite is true. What then can we glean from these findings, in light of the discussion on the virtues of the methodologies above? Firstly, we saw that it need not be the case that we perform model fit and model selection separately in all cases. For example, we saw in the Intoxication dataset that the Unified Penalized methodology had the best AIC of all the methods compared. In the Challenger dataset it also had the best overall model fit. Furthermore, this level of performance was achieved without any noticeable drop in prediction performance, since it was close to the best methods in this regard. In regards to inference as well, the results for the Unified Penalized methodology remained consistent especially for the Challenger dataset, where it found both temperature and pressure to be significant.

In considering the other methodologies, we can also see that the unified versions uniformly outperformed the nonunified versions, especially in Prediction, and did so again without sacrificing interpretability of the parameter estimates. The mathematical results above provide solid foundations for these results. However, broadly we may think of the nonunified versions as a specific versions of the unified methodology. Thus, the methodology is able to identify the correct parameters, whether the underlying DGP is symmetric or asymmetric, and in either case does not impose the link function approaching 1 or 0 at the same rate. Furthermore, it has many of the virtues of the nonunified version of Chapter 3, such as not needing to hold the variance constant and having continuous errors as well. Many of the same large-sample tests discussed there can also be applied to this formulation. However, now we may choose to apply the tests separately over the Hahn decomposition sets or together to see which methodology gives the best desired result.

In fact, the findings regarding significance of the nonparametric methodology is also relevant here and demand some further discussion. That is, the ability to perform MIPs using stronger topological spaces imply that we no longer need to sacrifice one mathematical goal for another. This is because a larger space of functions allow us a broader range of possible

candidates against which we may optimize our parameters of interest as a function of the Xs. In so doing a scientifically interpretable linear operator may outperform complex learning algorithms without sacrifing interpretability of parameter estimates. Since the parameters converge almost surely as a function of the Xs and the model specification, significance then no longer needs to be a 0-1 answer (ironically). Thus, we may use stronger convergence properties, on stronger topologies to sequentially make a model specification more complex as needed. If, on the other, hand inference is not an immediate goal, we may use these same robust properties of the methodologies with existing excellent AI and ML methodologies to give equivalent or better results. While these extensions are not pursued currently here, I allude to some of them in the forthcoming chapter.

One of the principle contributions of this research is in the insights it provides on the interplay of statistical significance and scientific significance. Focusing solely on large sample results with finite sample sizes can lead to bias and inconsistency of the parameter estimates in general. However, the results highlight that even in large samples the convergence strength of the methodology is crucial to scientfically rely on statistical inferential results in a robust way. In particular, almost sure convergence in concert with the underlying measure space on which the inferential results are considered is crucial for scientific significance beyond just statistical significance. To be precise, many Machine Learning (ML) and Artificial Intelligence (AI) algorithms and computational packages which can implement them often, though not always, sacrifice interpretability for Model Fit and Prediction. Occam's Razor is perhaps the most well known addage that comes to mind in such contexts, since the more complex we make our model, the harder it is for us to interpret its results. Consider for example, AI and ML methods such as Neural Networks (NN) or Support Vector Machines (SVM). The former may contain multiple hidden layers based on basis expansions of functions that defy any scientific foundation for its existence and the latter suffers from the same predicament depending on the various assumptions used in the model specification. In particular, say we want to predict the heights of certain individuals, and implement a model with sin(log(abs(temperature of

a certain lake in antarctica))) such that it perfectly classifies each individual in the Training Data (TrD). Then inspite of such good Model Fit or Prediction (MP) results in the TrD, it is hard to find a reasonable scientific explanation as to why in the general population such a model would predict height well.

Thus, it is reasonable to expect that model variables should be correlated in some scientific manner, which goes beyond just prediction in TrDs, as TeD results and interpretability may be equally important. Therefore, solely relying on one of the many information criteria for TrDs, such as Akaike Information Criteria (AIC) or Bayesian Information Criteria (BIC), does not necessarily guarantee better prediction results for TeDs. This is a fact relevant even if the underlying assumptions of the model specification on the TrD remains true for the TeD, since the presence of unknown latent relationships may become apparent only for the TeD, but not the TrD. Existing learning algorithms, Supervised or Unsupervised, can overcome some of these weaknesses, yet they may suffer from interpretability and overfitting beyond MP results, since the underlying calculations may be Blackboxes without scientifically relevant functional interpretations.

On the other hand, using scientifically interpretable models by itself does not always guarantee the best MP results for TrDs or TeDs either. This is because the existing explanatory variables in the sample may be insufficient to capture the functional specification at all sample points leading to interpretable parameter estimates, yet poor MP results for TeDs. A fact which is also relevant for the methodologies presented here. This is because all such models are a function of the observed Xs.

Therefore, conventional wisdom generally recognizes that the goal of the analysis should decide the type of methodology that should be used. So if one cares about classification, then we may use one of the many excellent existing AI or ML algorithms separately if interpretability is the goal, we may instead use a simpler functional specification. Yet the connection between MIP for any model considered should provide a more coherent framework

for comparison across diffent model and estimation processes considered. Therefore, it would be ideal to go beyond just the Likelihood Principle or information specific criteria, relying on asymptotic results, which themselves are reliant on large samples, to have methodologies that optimize MIP results on all dimensions.

The issues regarding scientific and statistical significance is of course multifaceted depending invariably on the explanatory variables present, measurement issues, scientfic question of interest, in addition to the mathematical assumptions made in the model specification. Yet an often overlooked criteria is the convergence properties of the methodologies used in model estimations. Since clearly we cannot have an infinite number of sample points, large sample results may be susceptible to both the sample size as well as the estimation procedure (please see Chowdhury (2021a) for further discussion on the parameter bias that may result from this exclusion). Thus, it should not be surprising that violations of the assumptions on which our model is built may give poor MIP results. Unfortunately, Blackbox learning algorithms can do the job for us only partially without prespecified restrictions on the model due to lack of interpretability.

The proposed methodologies through the use of LAHEML, overcomes these existing shortcomings and gives consistent results subject to the model specified and converges to the true parameters under general circumstances. This way the model parameters will converge to the best possible value given the data and the particular model specified. Thus, if the parameters converge to the true values for the model and the model is a priori known, interpretability can be asserted in a rigorous mathematical manner. In so doing, we may also maximize the MP criteria, as a function of the particular model specified and the observed data, without needing to learn it explicitly in some mysterious unobservable way sacrificing interpretability in the process. If the MP results are not deemed to be adequate, the Mathematician can then consider more complex models, perhaps one with interaction terms as opposed to only polynomials of lower orders. As the data change or are updated the model would remain

interpretable, being updated to be more complex, but in a known and interpretable way.

As an example, many scientific models are based on convergence in distribution or convergence in probability, on underlying topological spaces which are "weaker." Thus, it suffices to state that conclusions drawn from weaker topological spaces, even if strong convergence is asserted, are only partially informative in comparison to a methodology that asserts almost sure convergence on a stronger topological space. Therefore, the scientific conclusions and inference that we may draw from a methodology that relies on such properties should accordingly be better as well.

Indeed, this work shows that the conventional wisdom of treating Classification and Inference as separate tasks may not always be necessary. This is because an interpretable model using the right methodology with stronger convergence properties applied to stronger topological spaces can give equivalent or better results than Blackbox AI and ML methods in many circumstances. However, that is not to say that existing methods cannot be used, especially when the scientific question does not require Inference or Prediction at the same time per se. However, if Inference and interpretability are goals, especially when applications of them in the methodologies presented may give similar results to existing AI and ML models, good modeling philosophy should require that they be used first.

Such a philosophy has a long and illustrious history from Aristotle to Occam's Razor in the sciences. Afterall, if we we believe in the saying that all models are wrong, but some are useful, must we not then rely on robust methodologies with strong mathematical foundations over reliance on overly complex models? In essence, it highlights that relying on MP criteria to improve Inference and Prediction (IP) may be jointly achievable at the same time under specific mathematical preliminaries, in the constructions presented here. That is, a scientifically interpretable model with robust topological foundations with strong convergence properties can be extremely useful for classification without losing inferential characteristics for the proposed methodologies. On the other hand, a simple model applied without these

robust methodologies can lead to the wrong conclusions over existing AI and ML methods, though possibly at the cost of interpretability.

So what then should be the modeling philosophy to go beyond relying on p-values for the sciences? Well unfortunately, there is still much we have to learn. However, we may still be able to say some interesting things given the methodological contributions here. Chief among these is that relying solely on AIC or BIC or other Model Fit criteria might give an incomplete picture, if the data do not conform to the subtle mathematical assumptions of a model. Afterall, there are an infinite number of models one can run on a dataset, therefore, relying solely on minimizing model fit criteria can be just a little time consuming. Relying solely on Prediction criteria can also lead us down the "Rabbit Hole" since the performance on TrD and TeD need to be considered carefully. In either case, relying on performance criteria for any one category of modeling objectives does not guarantee that the results will be scientifically interpretable, even if the p-values are small for a predictor or the confidence intervals are amenable to claiming significance.

Therefore, the mathematical results here suggest that a scientifically interpretable model may have excellent predictive capabilities without sacrificing model fit or inference. However, in order to apply such a model, one must consider the convergence properties of the estimation procedure and certain subtle connections between topological spaces and measure spaces. What is important is to note that almost sure estimation methodologies such as LAHEML used in conjunction with stronger topological spaces may give excellent predictive results without sacrificing interpretability of parameter estimates. These results highlight a modeling exercise to be tied to the model specification, and not necessarily entirely dependent on large-sample results or Blackbox learning processes. Therefore, use of LAHEML in scientifically interpretable models can be a first step, which may be sequentially made more complex as necessary irrespective of the statistical goals. Furthermore, if MP is the desired goal, existing AI and ML models may accordingly be improved with these more ro-

bust methodologies that have strong convergence properties on stronger topological spaces, since intepretability would no longer be a constraint. In all such cases, methodologies that ensure almost sure convergence is to be preferred over methodologies with other convergence properties. In addition, all such models should be preferred when applied to stronger topological spaces in conjunction with almost sure convergence, to give truly "The Best of Both Worlds!"

4.6 Conclusion

In summary, this chapter presents the most generalized form of the methodologies. It has all the advantages of the previous methodologies and also expands on others. It therefore provides the ideal foundation on which to build any number of supervised or unsupervised methodologies in either the Frequentist or Bayesian formulation. As such, it provides further insights into our continuing discussion on the interplay of scientific significance and statistical significance broadly across scientific fields.

Chapter 5

Artificial Intelligence and Machine Learning Applications: An Overview

5.1 Applications to Artificial Intelligence and Machine Learning

The above methodologies have provided the complete mathematical foundations for extending many existing AI and ML methods without difficulty. Since it will not be possible to give a complete discussion on every single method possible, I broadly discuss some of the more well known applications in the sciences. Accordingly, consider when inference is not a focus of our mathematical modeling. Then we may extend the Latent Adaptive Hierarchical EM Like (LAHEML) methodology in learning algorithms in both supervised and unsupervised applications. Since the methodologies can be extended in either the Parametric, Nonparametric, or the Unified Framework, unless otherwise stated it should be understood that the extensions can be applied in either formulations. Accordingly, I first consider Supervised Learning models below and then discuss how they may be extended to various unsupervised applications.

5.1.1 Supervised Learning

At present I use the terminology Supervised Learning more broadly than perhaps generally recognized. In particular, the predictor variables are referred to as the independent variables (IV) as customary, however, supervised in the current sense means that the functional specification is known a priori for the relationship of the IVs to the Dependent Variables (DVs) of interest. Thus supervised in the present context implies we know specifically what the functional relationship is between the DV and the IV for at least one stage of the model specification. The application to NN in Section 5.1.1.3 will make this distinction clearer. However, I begin with discussion on a few other distinctive models.

5.1.1.1 Regression Splines

Regression splines can be seen as a special application of the methodologies above when we (pre)specify a particular number of cutoff points. Since any locally compact Hausdorff space may be separated into locally compact subspaces, any regression spline design can be approximated by the methodologies above.

5.1.1.2 Regression Trees

Since the methodology can be applied to any regression framework, it can also be easily applied in the construction of Regression Trees at the baseline level. All other parts of existing formulations may remain the same.

5.1.1.3 Artificial Neural Networks

Consider a K-class classification problem with one hidden-layer between input and classification layers. Then a parametric version of the results from previous chapters and using the notation from Hastie et al. (2009) can be given in the following formulation,

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, ..., M, \tag{5.1}$$

$$T_k = \beta_k^T Z, k = 1, ..., K, \tag{5.2}$$

$$f_k(X) = g_k(T). \tag{5.3}$$

For the parametric version we may assume the σ is the Logistic formulation with $g_k(T)$ being the multilogit formulation. Where the present model differs of course is the link condition needing to hold for each observation as before from Chapter [2], with the linear functional

being continuous according to the discussion in Chapter 5].

The nonparametric formulation can be given by using the results of Chapter [3] and Chapter [4]. A particular model can be given as,

$$Z_m = \nu(\alpha_{0m} + \alpha_m^T X), m = 1, ..., M, \tag{5.4}$$

$$T_k = \beta_k^T Z, k = 1, ..., K, \tag{5.5}$$

$$f_k(X) = g_k(T) = \frac{\nu^{T_k}}{\sum_{l=1}^{K} \nu^{T_l}}. \tag{5.6}$$

Of course, other layers can be added similarly as needed, therefore, Deep-Broad Neural Networks can be formulated accordingly without much difficulty as well. A version of this work can be found in Chowdhury (2021c). In fact, the work shows that while arbitrarily deep networks may approximate any function of the X's arbitrarily closely, that broad networks can estimate any function arbitrarily closely as well. This then gives us our next theoretical result.

Theorem 5.1. *Under the conditions of Theorem 4.2, a Broad Neural Network (BNN) can approximate any function of the observed variables arbitrarily closely as the sample size increases.*

Proof. This is a direct consequence of the results of Theorem 4.2. Note that Theorem 4.2 shows that there exists an unique linear operator from the sample space to the link field such that it is a function of an unique signed measure ν. Accordingly, given a linear operator which is itself a function of this unique signed measure, almost sure convergence of the operator is guarateed by the results of Theorem 4.2. The statement of the Theorem then readily follows. □

I call such a framework the Natural Neural Network (NNN or N^3). This is because most scientific applications are rooted in the real numbers, which are locally compact. Thus, the results above may be applied generally to all such models. Accordingly, even a single layered neural network can identify any function arbitrarily closely given enough derived features. As such, this formulation has ready extensions to many unsupervised learning applications, some which are discussed below in Section 5.1.2.

5.1.2 Unsupervised Learning

Here also there are many excellent existing methodologies in the literature. The foundational nature of the methodologies presented above, however, ensures that they can be easily applied to them as well. Accordingly, I give a brief discussion of it below.

5.1.2.1 Regression Forests

First note that even the most basic application of the methodology was able to outperform Random Forests as we saw in Chapter 3. Therefore, it seems a reasonable conclusion that when a Random Forest is applied using this methodology we will get improved results over existing methodologies.

5.1.2.2 Support Vector Machines (SVM)

For SVM, the application is again straightforward since the Kernel application in SVM may be adapted to the Neural Network example given above in Section 5.1.1.3.

5.2 Discussion

The literature on AI and ML applications of methodologies on Hilbert spaces is extensive. The contribution of the current methodologies and the foundational mathematical results allow us to apply them to general Banach spaces. However, since LAHEML may be applied on stronger topological spaces through Theorem 4.2 and Theorem 5.1, these excellent existing methodologies can therefore also be used in this framework to potentially improve MIP results accordingly.

The existence of an unique signed measure which may be used in conjunction with Theorem 4.2, to find unique linear operators that estimate our parameters of interest almost surely guarantees these assertions. As such, the potential applications and extension of LAHEML through these results to the numerous AI and ML methodologies are extensive. Accordingly, it is not possible to discuss all such extensions presently. However, the discussion above, though brief on some well known and popular methods, provides a baseline on which many such extensions can be based. I now conclude with some further discussions on these extensions in the final chapter.

Chapter 6

Future Research Direction and Concluding Thoughts

6.1 Future Research Directions

This thesis has provided a general framework under which mathematical models may be estimated with almost sure convergence of the parameter estimates on stronger topological spaces. Given its foundational nature, it has many possible applications across the sciences beyond those discussed in the previous chapters. This includes but is not limited to Causal Inference, Genetic Epidemiology, and Instrumental Variables to name just a few. In essence, anywhere a linear operator may be applied, LAHEML can be adapted accordingly in a particular model formulation. Therefore, the work presented here shows the potential to be widely applicable across many scientific fields and mathematical or statistical exercises.

6.2 Concluding Thoughts

In conclusion, the work highlights the importance of the link condition holding for all observations, generally for mathematical and statistical models. In particular, through LAHEML such pointwise convergence can be used to ensure almost sure convergence of the parameter estimates. A fact which may further be used to show convergence on stronger topological spaces. Accordingly, these findings show how such estimation processes are subtly connected to particular measure spaces through the use of signed measures. This in turn ensures that the likelihood principle holds generally for our models of choice. As such, this has profound consequences for our understanding of the interplay of "statistical significance" and "scientific significance," with deep mathematical connections.

In particular, it highlights the importance of the topological spaces on which we consider "statistical significance." In addition, it also points out the importance of the strength of convergence concepts on our ability to make inferential conclusions which are mathematically, logically, and scientifically accurate. Therefore, overall the methodologies presented and their

foundations based on Real Analysis, Functional Analysis, and Mathematical Statistics have broad implications for mathematical models across varied domains.

As such, it is hoped that the work here will help future scientists make scientific conclusions based on statistical models which are more aligned with mathematical realities. Which as a direct consequence should continue to push our understanding of the world around us in an interpretable and more scientific way as well. Therefore, it is my hope that the work presented here will help answer scientific questions of relevance through LAHEML and its extensions in a more mathematically precise way than has been possible before.

Bibliography

Abramson, C., Andrews, R. L., Currim, I. S., and Jones, M. (2000). Parameter bias from unobserved effects in the multinomial logit model of consumer choice. *Journal of Marketing Research*, 37(4):410–426.

Abrishami, A. and Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, 13(2):485–499.

Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

Agresti, A. and Kateri, M. (2011). *Categorical data analysis*. Springer.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Allenby, G. M. and Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of econometrics*, 89(1-2):57–78.

Andrews, R. L., Ansari, A., and Currim, I. S. (2002). Hierarchical bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *Journal of Marketing Research*, 39(1):87–98.

Bai, X., Zhang, F., and Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics*, 13(1):407–418.

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):1–9.

Beitia-Antero, L., Yáñez, J., and de Castro, A. I. G. (2018). On the use of logistic regression for stellar classification. *Experimental Astronomy*, 45(3):379–395.

Bland, M. J. and Altman, D. G. (2000, Last retrieved 10-17-2020). The odds ratio (electronic source). *<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1127651>*.

Bornmann, L., Leydesdorff, L., and Wang, J. (2014). How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics*, 8(1):175–180.

Cameron, A. C. and Trivedi, P. K. (2010). *Microeconometrics using stata*, volume 2. Stata press College Station, TX.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Choi, J., Yi, S., and Lee, K. C. (2011). Analysis of keyword networks in mis research and implications for predicting knowledge evolution. *Information & Management*, 48(8):371–381.

Chowdhury, K. (2019). Supervised machine learning and heuristic algorithms for outlier detection in irregular spatiotemporal datasets. *Journal of Environmental Informatics*, 33(1).

Chowdhury, K. (2021a). Functional analysis of generalized linear models under non-linear constraints with applications to identifying highly-cited papers. *Journal of Informetrics*, 15(1):101112.

Chowdhury, K. P. (2021b). *Functional analysis of generalized linear models under non-linear constraints with Artificial Intelligence and Machine Learning Applications to the Sciences*. PhD thesis, University of California, Irvine.

Chowdhury, K.P., S. W. (2021c). Bayesian latent adaptive deep neural networks on locally compact hausdorff spaces with applications to support vector machines. In *2021 Conference of International Society for Bayesian Inference*.

Chowdhury, K.P., S. W. (2021d). Nonparametric application of functional analysis of generalized linear models under nonlinear constraints. In *Symposium on Data Science and Statistics*. American Statistical Association.

Davison, N., Warren, R., Mason, K., McElhone, K., Kirby, B., Burden, A., Smith, C., Payne, K., and Griffiths, C. (2017). Identification of factors that may influence the selection of first-line biological therapy for people with psoriasis: a prospective, multicentre cohort study. *British Journal of Dermatology*, 177(3):828–836.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):45–70.

Edelman, B., Luca, M., and Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2):1–22.

Fouque, J.-P., Garnier, J., Papanicolaou, G., and Solna, K. (2007). *Wave propagation and time reversal in randomly layered media*, volume 56. Springer Science & Business Media.

Greene, W. (2003). *Econometric analysis Pearson Education India*.

Guerrero, V. M. and Johnson, R. A. (1982). Use of the box-cox transformation with binary response models. *Biometrika*, 69(2):309–314.

Hallsworth, M., List, J. A., Metcalfe, R. D., and Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148:14–31.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hattab, M., de Souza, R., Ciardi, B., Paardekooper, J., Khochfar, S., and Dalla Vecchia, C. (2018). A case study of hurdle and generalized additive models in astronomy: the escape of ionizing radiation. *Monthly Notices of the Royal Astronomical Society*, 483(3):3307–3321.

Hofmans, J. (2017). Modeling psychological contract violation using dual regime models: An event-based approach. *Frontiers in psychology*, 8:1948.

Hu, Y.-H., Tai, C.-T., Liu, K. E., and Cai, C.-F. (2020). Identification of highly-cited papers using topic-model-based and bibliometric features: the consideration of keyword popularity. *Journal of Informetrics*, 14(1):101004.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kass, R. E. and Steffey, D. (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association*, 84(407):717–726.

Killian, J. A., Passino, K. M., Nandi, A., Madden, D. R., Clapp, J. D., Wiratunga, N., Coenen, F., and Sani, S. (2019). Learning to detect heavy drinking episodes using smartphone accelerometer data. In *KHD@ IJCAI*, pages 35–42.

King, M. L. and Goh, K.-L. (2002). Improvements to the wald test. In *Handbook of Applied Econometrics and Statistical Inference*, pages 251–275. Marcel Dekker.

Lax, P. D. (2002). *Functional Analysis*. John Wiley and Sons.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Leng, C., Tran, M.-N., and Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 66(2):221–244.

Li, K., Mai, F., Shen, R., and Yan, X. (2018). Measuring corporate culture using machine learning. *Available at SSRN 3256608*.

Liu, D. and Zhang, H. (2018). Residuals and diagnostics for ordinal regression models: A surrogate approach. *Journal of the American Statistical Association*, 113(522):845–854.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological review*, 66(2):81.

Maity, A. K., Pradhan, V., and Das, U. (2018). Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. *The American Statistician*, pages 1–10.

Mandal, S. K. (2017). Performance analysis of data mining algorithms for breast cancer cell detection using naïve bayes, logistic regression and decision tree. *International Journal Of Engineering And Computer Science*, 6(2):20388–20391.

Murad, H., Fleischman, A., Sadetzki, S., Geyer, O., and Freedman, L. S. (2003). Small samples and ordered logistic regression: Does it help to collapse categories of outcome? *The American Statistician*, 57(3):155–160.

Phan, X.-H. and Nguyen, C.-T. (2008). A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference. *JGibbLDA. DOI: http://jgibblda. sourceforge. net*.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(1):15–24.

Royden, H. L. and Fitzpatrick, P. (2010). *Real analysis*. Pearson.

Simonoff, J. S. (1998). Logistic regression, categorical predictors, and goodness-of-fit: It depends on who you ask. *The American Statistician*, 52(1):10–14.

Sohrabi, B. and Iraj, H. (2017). The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. *Scientometrics*, 110(1):243–251.

Talukdar, D. (2008). Cost of being poor: retail price and consumer price search differences across inner-city and suburban neighborhoods. *Journal of Consumer Research*, 35(3):457–471.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, 34(4):273.

Tsai, C.-F. (2014). Citation impact analysis of top ranked computer science journals and their rankings. *Journal of Informetrics*, 8(2):318–328.

Uddin, S. and Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics*, 10(4):1166–1177.

Wang, M., Wang, Z., and Chen, G. (2019). Which can better predict the future success of articles? bibliometric indices or alternative metrics. *Scientometrics*, 119(3):1575–1595.

Zhang, W., Chen, Y., Liu, F., Luo, F., Tian, G., and Li, X. (2017). Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC bioinformatics*, 18(1):18.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., and Zhang, G. (2018). Does deep learning help topic extraction? a kernel k-means clustering method with word embedding. *Journal of Informetrics*, 12(4):1099–1117.

.1 Appendix

.1.1 Appendix A: Technical Proofs of Theorems, Propositions and Corollaries

Theorem 2: The Generalized Logistic Link function, $log(\mathbf{P}^{\alpha^*}(1-\mathbf{P})^{-1})$ is Analytic.

Proof Let $P_i = \mathbf{P}(\beta \mathbf{x_i})$. I first proceed to show that $P_i^{\alpha_i}$ is analytic.

Consider $P_i^{\alpha_i}$ and fix P_i. Note by definition $P_i \in [0,1]$, and let us restrict $P_i \in (0,1)$ for the remainder of this proof, excluding $P_i \in \{0,1\}$, sets of measure 0. Consider further a compact set $k = [a_*, a^*] \in \{0\} \cup R^+ = K$ and a Taylor series approximation at $\bar{\alpha_i} \in (a_*, a^*)$ is given by

$$f(P_i, \alpha_i) = f(\alpha_i) = P_i^{\alpha_i} = P_i^{\bar{\alpha_i}} + ln(P_i) P_i^{\bar{\alpha_i}} (\alpha_i - \bar{\alpha_i}) + \frac{ln(P_i)^2 \times P_i^{\bar{\alpha_i}}}{2!} (\alpha_i - \bar{\alpha_i})^2 + ... \tag{1}$$

Then expand lnP_i in another taylor series expansion around 1 to get

$$ln\ P_i = (\gamma - 1) - \frac{1}{2}(\gamma - 1)^2 + \frac{1}{3}(\gamma - 1)^3 - \frac{1}{4}(\gamma - 1)^4 + \text{Error}, \tag{2}$$

where γ belongs to some neighborhood of 1. Let γ^* be the optimized value for (2). Define $\eta(\gamma^*)$ as this functional value. Then (1) becomes

$$\begin{aligned} f(\alpha_i) = P_i^{\bar{\alpha_i}} + \eta(\gamma^*) P_i^{\bar{\alpha_i}} (\alpha - \bar{\alpha_i}) + \eta(\gamma^*)^2 P^{\bar{\alpha_i}} (i - \bar{\alpha_i})^2 + ... \\ + \eta(\gamma^*)^{n-1} P^{\bar{\alpha_i}} (\alpha_i - \bar{\alpha_i})^{n-1} + \eta(\gamma^*)^n P^{\bar{\alpha_i}} (\alpha - \bar{\alpha_i})^n, \end{aligned} \tag{3}$$

Let

$$a_n = |\frac{d^n f(P_i, \alpha_i)}{d\alpha_i^n}| = |\frac{\eta(\gamma^*)^n P_i^{\bar{\alpha}_i}(\alpha_i - \bar{\alpha}_i)^n}{n!}|, \tag{4}$$

and consider the series

$$\sum_{n=0}^{\infty} a_n. \tag{5}$$

In particular, consider the ratio test such that

$$n \to \infty \frac{a_{n+1}}{a_n} = |n \to \infty \frac{\eta(\gamma^*) P_i^{\bar{\alpha}_i}(\alpha_i - \bar{\alpha}_i)}{n}|, \tag{6}$$

Note that for any fixed $P_i \in (0, 1)$ and $\forall \bar{\alpha}_i \in [0, \infty)$, and $\bar{\alpha}_i \neq \infty$,

$$P_i^{\bar{\alpha}_i} \in [0, 1]. \tag{7}$$

Thus,

$$\eta(\gamma^*) P_i^{\bar{\alpha}_i} \leq \eta(\gamma), \tag{8}$$

$$\implies n \to \infty \frac{\eta(\gamma^*) P_i^{\bar{\alpha}_i}(\alpha_i - \bar{\alpha}_i)}{n} \leq n \to \infty \frac{\eta(\gamma^*)(\alpha_i - \bar{\alpha}_i)}{n}. \tag{9}$$

Consider, $\forall \bar{\alpha_i} \in (-\infty, 0]$, then we have

$$\implies n \to \infty \frac{\eta(\gamma^*) P_i^{\bar{\alpha_i}} (\alpha_i - \bar{\alpha}_i)}{n} \geq n \to \infty \frac{\eta(\gamma^*)(\alpha_i - \bar{\alpha}_i)}{n}. \tag{10}$$

Therefore, $\exists\ \alpha_i^* \in R$, and $\bar{\alpha_i} \neq -\infty$, such that,

$$\implies n \to \infty \frac{\eta(\gamma^*) P_i^{\bar{\alpha_i}} (\alpha_i - \alpha_i^*)}{n} = n \to \infty \frac{\eta(\gamma^*)(\alpha_i - \alpha_i^*)}{n}. \tag{11}$$

But $\eta(\gamma^*)(\alpha_i - \alpha_i^*)$ is fixed. Therefore,

$$n \to \infty \frac{\eta(\gamma^*)(\alpha_i - \alpha_i^*)}{n} \to 0. \tag{12}$$

Therefore, $P_i^{\bar{\alpha_i}}$ is analytic for every $\bar{\alpha_i} \in R \backslash \{-\infty, \infty\}$ and in particular for every $\bar{\alpha_i} \in [a_*, a^*]$. Thus, a_n is bounded and in particular $a_n \to 0$. It then readily follows that $P_i^{\bar{\alpha_i}}$ is analytic and its Taylor Series approximation exists for each $\bar{\alpha_i} \in R$ since R is the union of a family of open sets from elementary analysis and we do not consider the extended real number line. This leads to an extension to $(1 - P_i)^{\delta}$ as follows, $(1 - P_i)^{\delta_i}$ is also analytic. The result can be achieved by letting $\tilde{P}_i = (1 - P_i)$. Then, $\forall P_i \in (0, 1)$ proceeding as before the statement follows. It remains to show then that

$$\frac{P_i^{\alpha}{}_i}{(1 - P_i)^{\delta_i}} \tag{13}$$

is also analytic. However, this is established, since the ratio of two analytic functions is itself analytic. See for example, any introductory complex analysis book. Consequently, the proposed function, is real, analytic and therefore, continuous on the proposed domain for every observation $i \in \{1, ..., n\}$. Further since $\log(.)$, is a monotonic function, the monotonic

transformation of an analytic function, the odds-ratio, is also analytic, and the result is established for each i. Further, by the independence assumption of GLM the n-dimensional result readily follows.

Theorem 3 There is an unique solution to the link modification problem for the Generalized Logistic GLM formulation where the link constraint is binding for some $\alpha^* \in R^n \setminus \{-\infty, \infty\}$, given $\mathbf{P_i} \notin \{0,1\}$, $\mathbf{x_i} \notin \{0, \infty, -\infty\}$ for each $i \in \{1, ..., n\}$ and $\beta_j \notin \{\infty, -\infty\}$ with $j \in \{(1, ...(k+1))\}$.

Proof Let $P_i = \mathbf{P}(\beta \mathbf{x_i})$ and consider as before that $P_i \neq \{0,1\}$ and finite β_j for each $j \in \{(1, ...(k+1))\}$, then

$$log\left\{\frac{P_i^{\alpha_i^*}}{(1-P_i)}\right\} = \beta' \mathbf{c}(\mathbf{x_i}) \iff log\left\{\frac{P_i^{\alpha_i^*}}{(1-P_i)}\right\} - \beta' \mathbf{c}(\mathbf{x_i}) = 0, \tag{14}$$

$\forall P_i \in (0,1)$, the left hand side is finite, and the function itself is analytic on the reals. Therefore, $\exists M_i > \beta' \mathbf{x_i}$ such that

$$log\left\{\frac{P_i^{\alpha_i^*}}{(1-P_i)}\right\} - M_i < 0, \tag{15}$$

and $\exists M_i' < \beta' \mathbf{x_i}$ such that

$$log\left\{\frac{P_i^{\alpha_i^*}}{(1-P_i)}\right\} - M_i' > 0. \tag{16}$$

Therefore, given P_i and β finite, by the Intermediate Value Theorem $\exists\ \{\alpha_i^*, \delta_i^* = 1\}$ such that

$$log\left\{\frac{P_i^{\alpha_i^*}}{(1-P_i)}\right\} - \beta'\mathbf{c}(\mathbf{x_i}) = 0, \tag{17}$$

and α_i^* is unique.

Let $M1 = sup_i\ M_i$ and $M2 = inf_i\ M_i'$ taken over all i. We know this exists since the functional specifications are bounded and real valued.

Then,

$$log\left\{\frac{P_i^{\alpha_i^*}}{(1-P_i)}\right\} - M1_i < 0, \tag{18}$$

and $\exists M_i' < \beta'\mathbf{x_i}$ such that

$$log\left\{\frac{P_i^{\alpha_i^*}}{(1-P_i)}\right\} - M2_i > 0. \tag{19}$$

Therefore, by the intermediate value theorem for each i the Generalized Logistic GLM formulation holds with ready extension to the n-dimensional case as needed.

Proposition 1: There exists a family of link functions given by a monotonic transformation of the Generalized Odds function, $\mathbf{P}^{\alpha^*}(1-\mathbf{P})^{-1}$ such that for $(\alpha^* = \mathbf{1}, \delta^* = \mathbf{1})$ it represents the Generalized Logistic Link function for each observation.

Proof Let $\lambda(\beta'\mathbf{c}(\mathbf{x_i})) = \frac{e^{\beta'\mathbf{c}(\mathbf{x_i})}}{1+e^{\beta'\mathbf{c}(\mathbf{x_i})}} = Pr(y_i = 1|\mathbf{c}(\mathbf{x_i}), \beta) = P_i$. Then consider,

$$\frac{P_i}{1-P_i} = \frac{\frac{e^{\beta'\mathbf{c}(\mathbf{x_i})}}{1+e^{\beta'\mathbf{c}(\mathbf{x_i})}}}{1-\frac{e^{\beta'\mathbf{c}(\mathbf{x_i})}}{1+e^{\beta'\mathbf{c}(\mathbf{x_i})}}} = e^{\beta'\mathbf{c}(\mathbf{x_i})} \implies log\left\{\frac{P_i}{1-P_i}\right\} = \beta'\mathbf{c}(\mathbf{x_i}), \tag{20}$$

Thus, let $\alpha_i = \delta_i = 1$ *in* (13),

$$\implies g_0(\mu, \alpha_i = 1, \delta_i = 1) = \frac{P_i}{1-P_i} \implies log\left\{\frac{P_i}{1-P_i}\right\} = log(g_0(P_i, 1, 1)), \tag{21}$$

The n-dimensional result easily follows from independence.

.1.2 Frequentist Estimation Algorithm for Proposed Logistic Regression

With the regularity assumptions either in BOM or LVOM, let us consider an estimation process for equation (25) of the text. As this is now a constrained optimization problem, the estimation can proceed as follows.

---

1. For each observation estimate β using the relevant first order condition using a suitable hill climbing algorithm through Maximum Likelihood Estimation (MLE) subject to the link condition holding for each observation.

2. If the link condition does not exist due to analytical or numerical inconsistency either perform a Taylor approximation for $log(P_i|\mathbf{x_i}, \beta, \alpha_i^*)$ to solve for each α_i^* or solve for each α_i^* on a grid.

3. Get estimates of $(\bar{\beta^*},\bar{\alpha^*})$ under i.i.d. assumption.

It is clear that in the BOM or LVOM specification the estimation procedure above is valid for any GLM.