# Nonparametric Functional Analysis of Generalized Linear Models Under Nonlinear Constraints

Kali Chowdhury, Johns Hopkins University

**Abstract**

This article introduces a novel nonparametric methodology for Generalized Linear Models which combines the strengths of the binary regression and latent variable formulations, while overcoming their disadvantages. The mathematical results are implemented through a novel Bayesian Hierarchical estimation methodology. Requiring minimal assumptions, it extends recently published parametric versions of the methodology and generalizes it. If the underlying data generating process is asymmetric, it gives uniformly better prediction and inference performance over the parametric formulation. Furthermore, it introduces a new classification statistic utilizing which I show that overall, it has better model fit, inference and classification performance than the parametric version, and the difference in performance is statistically significant especially if the data generating process is asymmetric. In addition, the methodology can be used to perform model diagnostics for any model specification. This is a highly useful result, and it extends existing work for categorical model diagnostics broadly across the sciences. The mathematical results also highlight important new findings regarding the interplay of statistical significance and scientific significance. Finally, the methodology is applied to various real-world datasets to show that it may outperform widely used existing models, including Random Forests and Deep Neural Networks with very few iterations.

*Keywords: Unbalanced Data; MCMC; Artificial Intelligence; Machine Learning; Nonparametric Regression and Categorical Data Analysis, Model Diagnostics.*

## 1 Introduction

Binary outcome models continue to be relevant for Artificial Intelligence (AI) and Machine Learning (ML) applications in the sciences (Hu et al., 2020; Chowdhury, 2021a), as they serve as the building blocks for various multinomial extensions (Murad et al., 2003). Thus, their importance for AI and ML applications is also widely recognized (Li et al., 2018). As such, improvements of these foundational methods remain important for the greater scientific community. In addition, given their ubiquity to the sciences, any proposed improvement over existing methods should retain and be equivalent to the exisitng frameworks in a logical and consistent way if the data support them. This article presents such a framework rooted in Real and Functional Analysis with numerous advatages in

regards to Model fit, Inference, and Prediction (MIP) over existing methods. However, before highlighting these contributions, a broad comparison to existing frameworks is needed given the popularity of such models.

In particular, existing methods whether applied in a latent variable (LV) or binary outcome (BO) specification remains field specific. For example, in Econometrics LV models have been used to understand behavior of the average individual within a population (Greene, 2003), for calculating propensity scores for causal interpretation and program evaluation (Imbens and Rubin, 2015), and also to understand heterogeneity through finite and infinite mixture distributions (Andrews et al., 2002). In Biomedical Sciences (Zhang et al., 2017), and in the Physical Sciences (Hattab et al., 2018) there is also an extensive history of each formulation, and this is true of many fields across the sciences.

Indeed, from Chowdhury (2021a), it is also apparent that the underlying assumptions of BO vs. LV models may be distinctly different. Thus, it may be difficult to reconcile divergence in MIP performances between their applications. Further complexities arise if the data are unbalanced as then the assumptions of popular models such as the Logistic (logit) or Probit (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 (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 logit formulation it gave 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 gave results better or equivalent to popular AI methods such as Artificial Neural Network (ANN) under a wide range of circumstances without 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 ML methods could be compared with better coverage probabilities than existing widely used methodologies such as the maximum likelihood (mle) logit 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 held 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 technicalities. As such, this paper adds to this parametric version presented in Chowdhury (2021a) using an entirely novel nonparametric application termed Latent Adaptive Hierarchical EM Like algorithm. Though the parametric version remains relevant for inference, especially if the underlying data generating process (DGP) is symmetric, the nonparametric application can 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 and ML models or GLMs.

Thus, this article presents several 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. 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 based on Chowdhury (2019). This large-sample test allows model performance comparison to understand if divergences are statistically different.

In addition, in one of the more useful applications of the methodology, a separate and novel large-sample test is introduced, which allows us to 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 it extends Liu and Zhang (2018) for a broader understanding of model diagnostics for data analysis. Thus, in what follows I first present the mathematical foundations of the methodology in Section 2, postponing all technical proofs to the supplementary materials. Extensive simulation studies are performed in Section 3 followed by multiple applications of the methodology to real-world datasets in Section 4. I then discuss the findings in Section 5 where I also discuss the implications of the mathematical results on our continuing discussion of statistical significance and scientific significance. Finally I end with some concluding thoughts in Section 6.

# 2 Methodology

Following the notation of Agresti (2003) note that a GLM expands ordinary regressions to atypical 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 $E(\boldsymbol{y})$ (please note that for notational simplicity I suppress the dependence on $X$ unless otherwise states) relates to the systematic component. The random components of $\boldsymbol{y}$ are considered independent and identically distributed (i.i.d.)[1]. 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.

To retain the current models should the data support them, I follow the more general framework as in Chowdhury (2021a), and for completeness, restate the latent variable formulation as considered in Albert and Chib (1993) here as well. Let $\boldsymbol{y}^*$ be a latent or unobserved continuous random variable. Then an index function model for binary outcome gives the GLM,

$$\boldsymbol{y}^* = c(X)\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where

$$y_i = \begin{cases} 1 \; if \; y_i^* \geq 0, \\ 0 \; if \; y_i^* < 0, \end{cases} \tag{2}$$

[1]Henceforth, uppercase letters indicate matrices and bolded letters indicate vectors.

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 currently. It first seeks of a way to incoporate the strengths of both the latent variable and the binary regressions in a mathematically rigorous way. 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. Thus, below first I outline the mathematical results for the latent variable formulation using signed measures which has distinct advantages over the current latent variable formulation. For a nontechnical discussion of this framework I refer the reader to Section 2.3 and Section 2.2[2]. Then I discuss the large sample tests for model diagnostic and then briefly highlight the asymptotic properties of ARS.

## 2.1 Mathematical Results

Below, I first present the mathematical foundations for the discussions above[3]. For a discussion on when the latent variable and binomial regression formulations are equivalent, I refer the reader to Chowdhury (2021a)[4]. Thus, I take the results of Chowdhury [Ibid] and the insights from the discussions above to present a more coherent framework that unifies the methodologies in a mathematically rigorous way.

To that end note that 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, i.e., $\eta_i = \lambda(\boldsymbol{x}_i, \boldsymbol{\beta})$. The novelty of this methodology is in considering a signed measure over the $\sigma - algebra$ defined on the support of the link function. Below first this formulation is illustrated in a nontechnical way and then the mathematical results are given.

[2]The proofs of the results are also postponed to the supplementary materials.

[3]Some immediately relevant definitions may be found in the supplementary materials for this paper. I kindly refer the reader to any graduate level Analysis book for the remainder of the definitions.

[4]For the current formulation I assume that the formulations are equivalent.

## 2.2 Equivalency of Binomial Regression and Latent Variable Formulations

To see how the equivalency of the models can be ensured, first note that by construction of the binary regression,

$$E(\boldsymbol{y}|X) = F(\boldsymbol{y} = \mathbf{1}|X) = \boldsymbol{p}(X), \tag{3}$$

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,

$$f(y_i|\boldsymbol{x}_i) = p_i^{y_i}(1-p_i)^{(1-y_i)}. \tag{4}$$

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,

$$\begin{aligned} \boldsymbol{p}_i = F(\boldsymbol{x'}_i\boldsymbol{\beta}_i) = F^*[y_i^* > 0] \\ = F^*[-\epsilon_i < \boldsymbol{x'}_i\boldsymbol{\beta}_i] = F^*[\boldsymbol{\epsilon}_i < \boldsymbol{x'}_i\boldsymbol{\beta}_i] = F^*[\boldsymbol{x'}_i\boldsymbol{\beta}_i]. \end{aligned} \tag{5}$$

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,

$$L(X|\boldsymbol{y}, \boldsymbol{\beta}) = \prod_{i=1}^{n} F(y_i, \boldsymbol{x}_i, \boldsymbol{\beta})^{y_i}(1 - F(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}))^{1-y_i}, \tag{6}$$

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,

$$p_i = F(\boldsymbol{x'}_i\boldsymbol{\beta}_i) \neq F^*[-\epsilon_i < \boldsymbol{x'}_i\boldsymbol{\beta}_i], \tag{7}$$

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,

$$E(y_i) = F(\boldsymbol{x'}_i\boldsymbol{\beta}_i), \tag{8}$$

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 (7). 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 this 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 known data likelihood, namely binomial under i.i.d. assumptions.

Accordingly, note that if due to reasons above (7) occurs we may consider a pointwise transformation such as,

$$(F^*[-\epsilon_i < \boldsymbol{x'}_i\boldsymbol{\beta}_i])^{\boldsymbol{\alpha}^*} = E(y_i) = F(\boldsymbol{x'}_i\boldsymbol{\beta}_i). \tag{9}$$

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 itself 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 (9) 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,

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \boldsymbol{c}(\boldsymbol{y}, \boldsymbol{y^*})L(\boldsymbol{\beta}|\boldsymbol{y^*}), \tag{10}$$

for some constant $\boldsymbol{c}(\boldsymbol{y}, \boldsymbol{y}^*)$. While the mathematical results in the main text provide the necessary foundations for this construct, 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. This is detailed in an intuitive way in Section 2.3.

## 2.3 Existence and Uniqueness of Signed Measure

Since probability spaces necessarily deal with unsigned measures, the importance of such a construction based on signed measures may not immediately be apparent. Therefore, to highlight its importance this section gives a nontechnical discussion of the signed measure construction without overly complex mathematical notations. As such, to begin our discussion consider the usual Logistic regression for the binary specification when we specify

$$F(\mathbf{x}_i, \boldsymbol{\beta}) = (1 + exp(\mathbf{x}_i\boldsymbol{\beta})^{-1})^{-1}. \tag{11}$$

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

$$log\left(\frac{\boldsymbol{F}(\boldsymbol{y}, X, \boldsymbol{\beta})}{\mathbf{1} - \boldsymbol{F}(\boldsymbol{y}, X, \boldsymbol{\beta})}\right) = \boldsymbol{\lambda}(X, \boldsymbol{\beta}) = c(X)\boldsymbol{\beta}, \tag{12}$$

where following Chowdhury (2021a) $\lambda$ is a continuous, bounded function of the explanatory variables in the Generalized Linear Model (GLM) framework. In the nonparametric setting this condition need not hold if F, the probability of success is not the Logistic distribution. Furthermore, the underlying probability of success and its distribution is something that is inherently unknown and as always 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. Indeed, simply assuming a nonparametric distribution is also not enough to ensure equivalency between a binary regression framework and its latent variable specification (Proposition 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 $(X, \Sigma, \nu)$ is also a measure space where $\nu$ is the restriction of $\nu_0$ to $\Sigma$. In particular, note that by Skohorod we may define a random variable $\boldsymbol{\lambda}(X, \boldsymbol{\beta})$, such that $(\boldsymbol{\lambda}(X, \boldsymbol{\beta}), \Sigma)$ 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 $\zeta$, 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,

$$\boldsymbol{\nu} = \boldsymbol{\nu}^+ - \boldsymbol{\nu}^-, \tag{13}$$

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

$$\nu(\boldsymbol{x}_i, \boldsymbol{\beta})^{\tau^*} = \lambda(\boldsymbol{x}_i, \boldsymbol{\beta}) \; and \; \tau \in \mathbf{R} \setminus \{-\infty, \infty\}, \tag{14}$$

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 (9) 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 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 $\boldsymbol{\lambda}(\boldsymbol{x}_i, \boldsymbol{\beta}) \geq \mathbf{0}$,

$$log\left(\frac{\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x}_i, \boldsymbol{\beta})}{\mathbf{1} - \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x}_i, \boldsymbol{\beta})}\right) = \boldsymbol{\lambda}(\boldsymbol{x}_i, \boldsymbol{\beta}) = \hat{\boldsymbol{F}}(\boldsymbol{x}_i, \boldsymbol{\beta})^{\boldsymbol{\alpha}^*}, \tag{15}$$

holds for some $\alpha^*$ 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 (15) 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.

### 2.3.1 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 $(X, \Sigma_0, \nu_0)$ 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 1.** *For the measureable space $(\lambda(X, \boldsymbol{\beta}), \Sigma)$ There exists a signed measure $\nu$ such that,*

$$y^* = \begin{cases} 1 \ if \ \lambda \geq 0 \\ 0 \ if \ \lambda < 0, \end{cases} \tag{16}$$

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

Above and for the remainder of this manuscript for notational simplicity I employ $\lambda$ 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 2.** *For the measureable space $(\lambda, \Sigma)$ there exists an unique decompostion of the signed, finite measure $\nu$ as a function of two positive measures $\nu^+$ and $\nu^-$ such that,*

$$\nu = \nu^+ - \nu^- \ where, \tag{17}$$

$$y_i^* = \begin{cases} 1 \ if \ \lambda \geq 0 \\ 0 \ if \ \lambda < 0, \end{cases} \tag{18}$$

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

**Proposition 3.** *Let $(\lambda, \Sigma, \nu)$ be a measure space as above and $\nu$ a finite signed measure on it. Then,*

$$|\nu|(\Sigma) = \bar{\nu}^+(S^+) + \bar{\nu}^-(S^-) \tag{19}$$

*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 $\nu$.*

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

**Theorem 1.** *Let $(X, \Sigma, \mu)$ 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 $\mu(\cup_j E_j) < \infty$.*

The existence of this result is important for dealing with signed measures, which can take one of the nonfinite values which is not $\sigma - finite$. Therefore, utilizing this result we can now show one of the more important results and corollaries of Theorem 1, which will be important for the construction of finite or $\sigma - finite$ measure spaces for all GLMs. It is detailed below.

**Proposition 4.** *Let $(\lambda, \Sigma)$ be a measureble space with $\mu$ 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 $\nu$ 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.*

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 $E \in \Sigma$,

$$|\bar{\nu}| = |\nu|\delta_{E\cap S^+}, \ where \ \delta_{E\cap S^+} = \begin{cases} 1 \ if \ E \in S^+ \\ 0 \ o.w. \end{cases} \tag{20}$$

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 1 and Proposition 4, can be overcome to address this information loss concern, and the following proposition addresses this[5]. Thus, the Hahn-Banach Theorem can be used to define a linear functional over all integrable functions in $L^p(\lambda, \Sigma, |\bar{\nu}|)$ with $1 \leq p < \infty$.

**Proposition 5.** *Consider a Hahn-Decomposition of the measure space $(\lambda, \Sigma, \nu)$ into $\{S^+, S^-\}$ as defined before, where the signed-measure $\nu$ WLOG takes the value of $-\infty$ but is not $\sigma-$finite. Then there exists a linear functional $\mathcal{L}$ which extends any measure $\nu^+$ over $S^+$ to all of $L^p(Q, |\bar{\nu}|)$, with $|\bar{\nu}|$ as in Proposition 4, and $Q$ is the measureable space $(\lambda, \Sigma)$ with $1 \leq p < \infty$.*

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

**Corollary 1.** *Let $(\lambda, \Sigma, |\bar{\nu}|)$ be a $\sigma-$finite measure space, where $|\bar{\nu}|$ is defined as in proposition 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,*

$$\{f_n\} \to f, \tag{21}$$

*in measure.*

**Corollary 2.** *Let $(\lambda, \Sigma, |\nu|)$ be a $\sigma-$finite measure space, where $|\nu|$ is defined as in proposition 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,*

$$\{f_n\} \to f, \tag{22}$$

*in measure.*

These results point to convergence in probability of the parameters in the current formulation. While convergence in probability is useful, below a stronger result is given in Theorem 3. Furthermore, these results also have several nonintuitive applications to non-binary analysis and the remarks below highlight some of them.

[5]The relevant definitions can be found in the supplementary materials

**Remark 1.** *First, note that the decomposition above is unique, and as such the existence of a latent variable implies the existence of an unique pair of positive measures that can represent it.*

**Remark 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.** *That a* $\sigma-$*finite signed measure can be extended to a complete measure space* $(\lambda, \Sigma, \nu)$ *with* $\nu$ *a restriction of the outermeasure on* $\Sigma$ *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* $\lambda$*. 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 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* $\Sigma$ *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* $\lambda$*. 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 the methodology. However, first I give some foundational results for the uniqueness of the link constraint.

**Theorem 2.** *Let* $(\lambda, \Sigma)$ *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* $j \in \{1, ..., (k+1)\}$*.*

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

**Theorem 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,*

$$\hat{\beta} \xrightarrow{a.s.} \beta. \tag{23}$$

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

**Corollary 3.** *Under the conditions of Theorem 3 we have that for* $1 \leq p < \infty$*,*

$$\{g_n\} \xrightarrow{a.s.} p(\beta|y). \tag{24}$$

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 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.*

These results are fundamental to the existence and uniqueness of the methodology based on signed measures. However, in order to apply it, an equally important aspect is the estimation procedure. Next this is elaborated.

## 2.4 Nonparametric Latent Adaptive Hierarchical EM Like (LAHEML) Algorithm

The mathematical foundations above may be used to implement an extremely versatile almost sure convergent methodology for the parameters of interest, using a Bayesian Hierarchical MCMC framework through data augmentation. I refer to the estimation methodology as Latent Adaptive Hierarchical EM Like Algorithm (LAHEML), which is nevertheless more general than the EM algorithm. This is because following Chowdhury (2021a), only ergodicity and aperiodicity are required for the convergence results to hold. In fact, the mathematical foundations ensure that the adaptive algorithm may also allow for both model fit and model selection to be performed at the same time. While specifc applications of the algorithm can be found in the supplemetary materials in both penalized and unpenalized formulations, here I present the general algorithm.

**Algorithm 1.** ***LAHEML Algorithm***

***Require:*** *Starting values for parameters $\boldsymbol{\beta}^{(j)}$. Functional specification $f(\boldsymbol{y}, X)$. Uninformative or Diffuse Prior specification.*

1. *Subject to the link constraint holding for each observation, compute the truncated signed measures as a function of $\kappa$.*
2. *Compute the link constraint parameter distribution for $\alpha^{(j+1)*}$.*
3. *Perform an MH step to get $\beta^{(j+1)}$ as a function of $\alpha^{(j+1)*}$ for the relevant likelihood through data augmentation such that, $\beta^{(j)} \leftarrow \beta^{(j+1)}$.*
4. *Iterate to completion.*

The actual implementation of the algorithm is rather involved because of the need to consider the likelihood principle in the various steps. As such, specific applications in unpenalized and penalized formulations may be found in the supplementary materials. In both applications the mathematical construct ensures that $\{\boldsymbol{\beta}^{(j)}\}$ converges to its true distribution given the data $\boldsymbol{y}$. Since $\boldsymbol{\alpha}^*$ is a function of $\boldsymbol{\beta}$, the convergence results hold for its distribution as well. In particular, the bias of the nonparametric density estimates are corrected by ensuring the link condition holds for each observation. In the implementation

of the LAHEML algorithm, it is also worthwhile considering that the cutoff points of $\kappa \in (0,1)$ a probability, may 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 Section 3.

A penalized application was also considered for LAHEML using the Bayesian Adaptive Lasso as in Leng et al. (2014). This penalized version of the methodology is contingent on a different prior specification than in the unpenalized case. Specifically, in this case the prior is given by,

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{k=1}^{p} \frac{\lambda_k}{2\sqrt{\sigma^2}} e^{\lambda_k|\beta_k|/\sqrt{\sigma^2}}. \tag{25}$$

For the highlighted application we may 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 hyperparameters on the shrinkage parameters, which can no longer be considered uninformative. Leng et al. (2014) consider a gamma prior on the $\lambda_k$'s and I also follow this same formulation. Thus, the prior on the shrinkage parameters can be given by

$$\pi(\lambda_k^2) = \frac{\delta^r}{\Gamma(r)} (\lambda_k^2)^{r-1} e^{-\delta\lambda_k^2}. \tag{26}$$

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 all present applications I set the $\delta$ and $r$ hyperparameters both equal to some small number such as 0.1. I defer further discussions to the simulation and empirical applications and now elaborate on the asymptotic model diagnostic and Adjusted ROC-Statistic below.

## 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, $F(\boldsymbol{\beta}|y) = \lambda(X, \boldsymbol{\beta})$. 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 we can calculate the divergence of $\hat{\boldsymbol{F}}(\hat{\boldsymbol{\beta}}|\boldsymbol{y})$ from (2.5). For example, we may compute the value of $\alpha^*$, say $\bar{\alpha^*}$ that minimizes,

$$\left(\hat{\boldsymbol{F}}(\boldsymbol{\beta}|\boldsymbol{y})^{\boldsymbol{\alpha^*}} - \boldsymbol{\lambda}(X, \boldsymbol{\beta})\right). \tag{27}$$

In particular, we know for GLM, if the convergence has occured to the true distribution then $\alpha^*$ 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 $\alpha^*$ as well to get,

$$\begin{aligned} \alpha^* \sim N(1, E(E(\alpha^*|\boldsymbol{\beta}^*) - 1|\boldsymbol{\beta}^*)^2), \implies \\ \hat{\bar{\alpha}}^* \overset{asymp.}{\sim} N\left(1, \frac{\sum_{i=1}^{n}(\alpha_i - 1)^2}{n-1}\right). \end{aligned} \tag{28}$$

$\boldsymbol{\beta}^*$ 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.

**Algorithm 2.** ***Large Sample Test for Model Diagnostic.***

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*[6].

This framework can similarly be extended to the likelihood ratio test, under the appropriate null values. Since this algorithm can be directly applied no matter the parametric distribution assumed, without the need to fix any variance parameter of the latent distribution, it extends Liu and Zhang (2018) accordingly. As such, it is applicable to any of the infinitely many distributional assumptions that are possible, without the need to actually make such an assumption in the model specification at all. As such, it is one of the more important contributions of this work.

## 2.6 Asymptotic Distribution of Adjusted ROC-Statistic

In order to analyze adequacy of classification performance, there are many existing tools such as the Receiver-Operating Curve. Here I consider the 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. In addition, the mathematician can employ bootstrap or other methods as well to compare the performance difference between models.

[6]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.

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

$$True\ Positive = \frac{|S(t) \bigcap G|}{|G|}, \tag{29}$$

$$True\ Negative = \frac{|\neg S(t) \bigcap \neg G|}{|\neg G|}, \tag{30}$$

$$False\ Positive = \frac{|S(t) - G|}{|D - G|}, \tag{31}$$

$$False\ Negative = \frac{|\neg S(t) - G|}{|G|}. \tag{32}$$

Then,

$$ARS = \frac{\frac{|S(t)-G|}{|D-G|} + \frac{|\neg S(t)-G|}{|G|}}{\frac{|S(t) \bigcap G|}{|G|} + \frac{|\neg S(t) \bigcap \neg G|}{|\neg G|}} \tag{33}$$

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

Then ARS can be shown to have an asymptotic distribution (Bland and Altman, 2020) given by

$$\begin{aligned} log(ARS) &\sim N(log(Oddsratio), \sigma^2), \\ where\ \sigma &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}} \end{aligned} \tag{34}$$

where $n_{ij}$ are the counts within each cell in the confusion matrix. A derivation of this is given in the supplementary materials for this paper. Currently, note 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.

### 2.6.1 Inference

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

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

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

$$\frac{\sqrt{n}\ log(ARS)}{\sigma}) \sim N(0, 1). \tag{35}$$

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,

$$\text{Test Statistic} = \frac{\bar{\kappa_1} - \bar{\kappa_2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}, \tag{36}$$

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. As such, the methodology can be readily extended through a Likelihood Ratio based formulation. Semiparametric estimation of ARS can also proceed straightforwardly and both of these applications are discussed in the supplementary materials.

# 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 following Chowdhury (2021a). Various DGPs both symmetric (Logit and Probit) and asymmetric (Complementary Log-Log) were considered, for different sample sizes ($n = \{100, 500, 1000, 2000\}$) and models,

$$\boldsymbol{y} = Intercept + X_1 + X_2, \tag{37}$$

$$\boldsymbol{y} = Intercept + X_1 + exp(X_2), \tag{38}$$

$$\boldsymbol{y} = Intercept + exp(X_1) + sin(X_2). \tag{39}$$

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 $\beta$'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[7].

In addition, two separate simulation runs were performed, one in an unpenalized and the other in a penalized formulation. Where for the penalized application a nuisance variable was added to compare model fit and model selection performances concurrently. The results for the unpenalized simulation can be found in Table 1, Table 2, and Table 3 below, and the results of the penalized simulation may be found in Table 4, Table 5, and Table 6 below. The Proposed Unpenalized (not the Bayesian Adaptive Lasso) Nonparametric application uniformly contained the true parameters more often, and thus had better coverage than the other methods compared. It was able to attain this by having confidence interval ranges which were smaller than the Penalized Logistic, which had the worst coverage performance

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

among the methods compared here. Furthermore, this was achieved without considering functional specifications which lack scientific interpretability as in Neural Networks[8].

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 (2021c) show that they are outperformed by the other methodologies compared. The difference in ARS on average were statistically significant for the Proposed Nonparametric method in comparison to all other models compared for Nonlinear, Mixed and Linear models over all DGPs and datasets considered.

Seaparately, the penalized simulation contained an extra nuisance parameter drawn randomly from the standard normal. Here again, the results were extremely encouraging, and consistent with the results from the previous application. The Proposed Penalized (using Bayesian Lasso) Nonparametric methodology can not only outperform existing models (including the parametric methodology) in inference, but also in classification in regards to ARS. Indeed, the classification results even outperforms the unpenalized (not using Bayesian Lasso) application and Neural Net, on average over all of the many different datasets considered. The inference results again show near perfect coverage results with smaller confidence intervals than the Penalized Logistic[9]. 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 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.

# 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.

[8]This is because deeper networks with more complex basis expansions can lead to scientifically uninterpretable models, at the cost of better classification outcomes.

[9]In the current application, since an extra nuisance variable was considered, the Penalized Logistic model was also considered for comparison.

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

| | Prop. NP. | Prop. P. | Bys. Prbt. | Pen. Logit |
|---|---|---|---|---|
| L. (NL) | 98.33% | 95.00% | 83.33% | 51.67% |
| P. (NL) | 100.00% | 95.00% | 75.00% | 46.67% |
| C. (NL) | 100.00% | 93.33% | 80.00% | 56.67% |
| L. (Mx.) | 98.33% | 91.67% | 71.67% | 61.11% |
| P. (Mx.) | 100.00% | 95.00% | 75.00% | 25.00% |
| C. (Mx.) | 100.00% | 96.67% | 81.67% | 20.00% |
| L. (L) | 100.00% | 96.67% | 85.00% | 63.33% |
| P. (L) | 98.33% | 96.67% | 80.00% | 66.67% |
| C. (L) | 98.33% | 96.67% | 81.67% | 66.67% |

Table 2: Simulation Confidence Interval Range Summary for All DGPs (at 1% Significance Level).

| | Pen. Logistic | Prop. NP. | Prop. Logit | Bys. Prbt. |
|---|---|---|---|---|
| L. (NL) | 6.47 | 5.66 | 5.37 | 2.00 |
| P. (NL) | 7.44 | 5.42 | 5.20 | 1.77 |
| C. (NL) | 7.77 | 5.65 | 4.89 | 1.88 |
| L. (Mx.) | 7.66 | 5.75 | 5.40 | 2.07 |
| P. (Mx.) | 3.94 | 5.64 | 5.12 | 1.87 |
| C. (Mx.) | 2.27 | 6.12 | 4.93 | 1.84 |
| L. (L) | 7.12 | 5.90 | 4.73 | 1.75 |
| P. (L) | 7.45 | 5.77 | 5.15 | 1.92 |
| C. (L) | 6.96 | 5.73 | 4.81 | 1.66 |

Table 3: Unpenalized Simulation Summary of ARS for All DGPs Considered.

| | 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, sample sizes and unbalancedness for all models fitted. 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. L., P., and C. indicate the Logistic, Probit and Complementary Log Log DGPs respectively. NL, Mx, and L indicate Nonlinear, Mixed and Linear DGPs respectively.

Table 4: Summary of Penalized Simulation Coverage Percentage for All DGPs (at 1% Significance Level).

| | Unp. NP. | Pen. NP. | Prop. P. | Bys. Prbt. | Pen. Logit |
|---|---|---|---|---|---|
| L.(NL) | 100.00 | 100.00 | 98.75 | 92.50 | 70.00 |
| P. (NL) | 100.00 | 98.68 | 97.37 | 84.21 | 65.79 |
| C. (NL) | 100.00 | 98.75 | 98.75 | 81.25 | 65.00 |
| L. (Mx.) | 100.00 | 100.00 | 100.00 | 81.94 | 70.83 |
| P. (Mx.) | 97.22 | 97.22 | 94.44 | 81.94 | 70.83 |
| C. (Mx.) | 100.00 | 100.00 | 96.67 | 81.67 | 72.06 |
| L. (L) | 98.75 | 100.00 | 100.00 | 88.75 | 75.00 |
| P. (L) | 96.25 | 95.00 | 97.50 | 80.00 | 73.75 |
| C. (L) | 100.00 | 100.00 | 100.00 | 87.50 | 75.00 |

Table 5: Penalized Simulation Confidence Interval Summary for All DGPs (at 1% Significance Level).

| | Unp. NP. | Pen. NP. | Prop. Logit | Bys. Prbt. | Pen. Logit |
|---|---|---|---|---|---|
| L. (NL) | 5.96 | 5.77 | 5.52 | 1.87 | 5.64 |
| P. (NL) | 6.04 | 5.60 | 4.93 | 1.99 | 5.81 |
| C. (NL) | 5.88 | 5.43 | 5.63 | 2.03 | 6.18 |
| L. (Mx.) | 5.64 | 5.87 | 5.67 | 1.80 | 6.22 |
| P. (Mx.) | 5.67 | 5.78 | 5.22 | 1.97 | 6.03 |
| C. (Mx.) | 5.97 | 5.44 | 5.44 | 1.97 | 6.07 |
| L. (L) | 5.74 | 5.66 | 5.67 | 1.78 | 6.25 |
| P. (L) | 5.53 | 5.15 | 4.86 | 1.96 | 6.32 |
| C. (L) | 5.74 | 5.97 | 5.60 | 1.94 | 6.46 |

Table 6: Penalized Application Summary of ARS for All DGPs Compared.

| | Unp. NP. | Pen. NP. | P. L. | Bay. P. | Neu. 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, sample sizes and unbalancedness for all models fitted. 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. L., P., and C. indicate the Logistic, Probit and Complementary Log Log DGPs respectively. NL, Mx, and L indicate Nonlinear, Mixed and Linear DGPs respectively. NP. is nonparametric, Unp. implies nonparametric, Pen. implies penalized, Bay. implies Bayesian and Neu. Net. implies Neural Net.

## 4.1 Detecting Heavy Drinking Events Using Smartphone Data

To illustrate the efficacy of the model, I 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,

$$\begin{aligned} TAC = Intercept + x - axis\ reading + \\ y - axis\ reading + z - axis\ reading. \end{aligned} \tag{40}$$

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 Test Data (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 (2021c) (the relevant plots can be found in Figure 1 and Figure 2). The strength of the

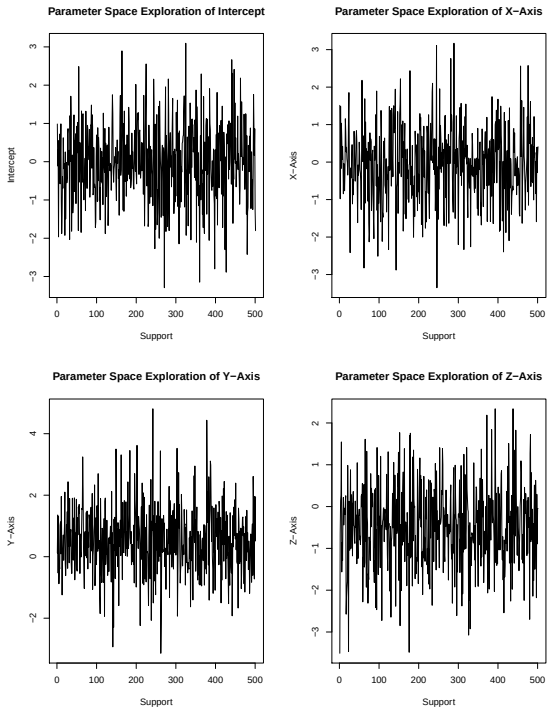


Figure 1: Unpenalized Draws.

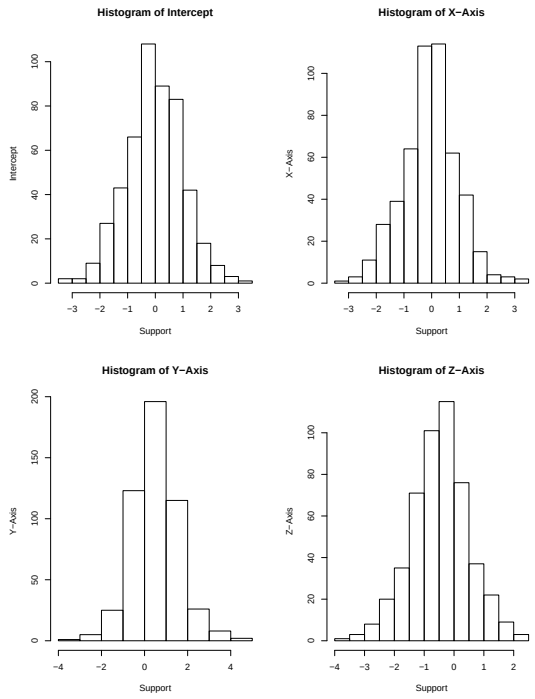


Figure 2: Unpenalized Histograms.

methodology to perform both model fit and model selection at the same time as seen in the simulation studies were also evident here. Specifically, in the penalized application the Adaptive Bayesian Lasso of (Leng et al. (2014)) was applied in a Hierarchical framework. Contrary to the accepted norm that we cannot perform model fit and model selection at

the same time, the TeD application had perfect predictive accuracy (the relevant plots for the methodology can be found in the supplementary materials for this article). 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. However, I defer further discussions on these findings till the discussion section.

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 7 below. The results bring to mind the image of a heavily intoxicated individual 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 (0.95) application with Proposed Parametric Logistic (0.97) 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.

## 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.

$$Signal/Noise = Intercept + \sum_{i=1}^{28} Feature_i. \tag{41}$$

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 for the unpenalized application may be found below in Figure 3, Figure 4. The penalized application figures can be found in the supplementary materials for this paper. 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 9 below. The unpenalized application was especially good for the Test Dataset (TeD), with the penalized version also

Table 7: Heavy Drinking Event Detection Parameter Summary for All Methodologies.

| Method | Predictor | Est. | CL | CH |
|---|---|---|---|---|
| (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 |

Table 8: Heavy Drinking Event Detection ARS Summary

| | (1)-TrD. | (1)-TeD. | (2)-TrD. | (2)-TeD. |
|---|---|---|---|---|
| ARS | 0.30 | 0.00 | 0.34 | 0.00 |

Note: (1) Nonprametric, (2) Penalized Nonparametric, (3) Parametric, (4) Existing Bayesian, (5) MLE Logistic, (6) Penalized Logistic. Est. indicates Estimate. CL and CH indicate Confidence Interval Low and Confidence Interval High respectively. TrD. and TeD. indicate Training and Test Datasets (20% of the data) respectively

Table 9: Signal/Noise Detection Summary of ARS for Nonparametric Application to Exotic Particle Detection Data.

| | Unp. NP. (TrD) | Unp. NP. (TeD) | Pen. NP. (TrD) | Pen. NP. (TeD) |
|---|---|---|---|---|
| ARS | 0.36 | 0.06 | 0.44 | 0.14 |

Note: Unp. NP. (TrD) is the unpenalized application on the training data, and Unp. NP. (TeD) is the unpenalized nonparametric application on the test data (last 500,000 observations). Pen. NP. (TrD) is the penalized application on the training data, and Pen. NP. (TeD) is the penalized application on the test data (last 500,000 observations).

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.

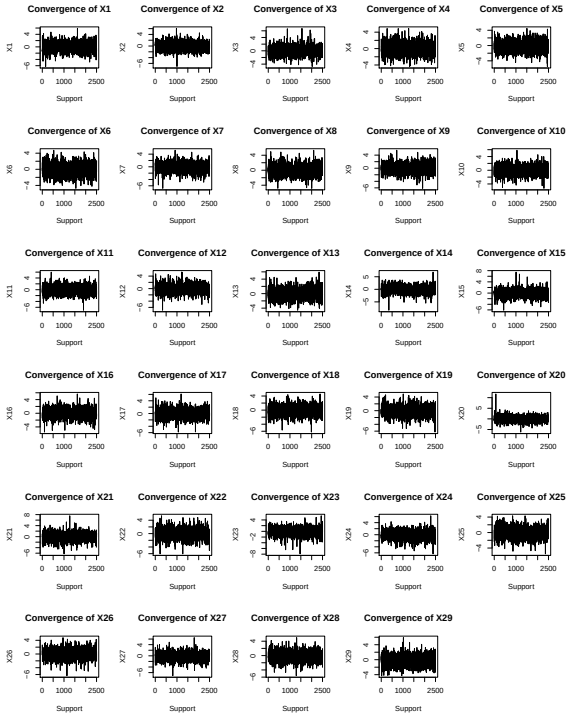


Figure 3: Unpenalized Draws.
Note: The first plot in the upper left corner represents the intercept (X1). All other plots are sequential from left to right as presented in Baldi et al. (2014).

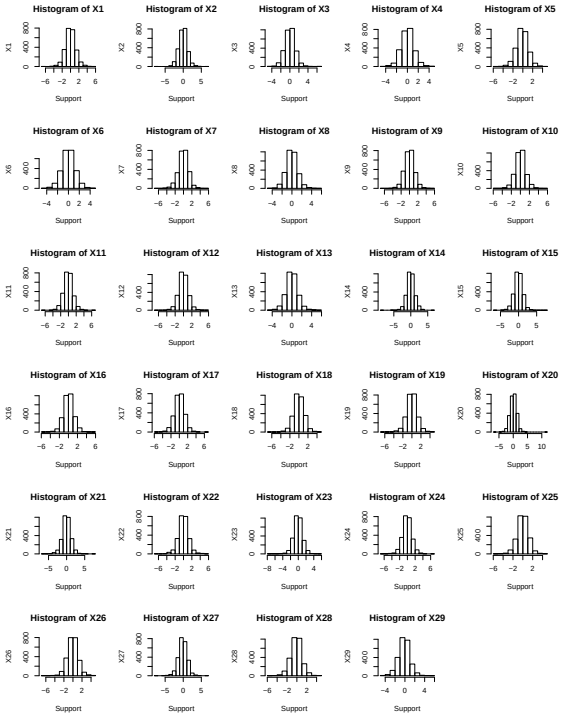


Figure 4: Unpenalized Histograms.
Note: The first plot in the upper left corner represents the intercept (X1). All other plots are sequential from left to right as presented in Baldi et al. (2014).

# 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, and had 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 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.

# 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.

# References


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.

Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

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.

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.

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.

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¿*.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

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). Nonparametric application of functional analysis of generalized linear models under nonlinear constraints. In *Symposium on Data Science and Statistics*. American Statistical Association.

Greene, W. (2003). *Econometric analysis Pearson Education India*.

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.

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.

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.

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.

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.

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.

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.

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.

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.