Journal of Informetrics 15 (2021) 101112

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Functional analysis of generalized linear models under non-linear constraints with applications to identifying highly-cited papers

K.P. Chowdhury

University of California, Irvine, United States

ARTICLE INFO

Article history:
Received 18 April 2020
Received in revised form
16 November 2020
Accepted 19 November 2020
Available online 16 January 2021

Keywords:
Unbalanced data
MCMC
Neural Networks
Artificial Intelligence
Machine Learning
Logistic regression
Categorical data analysis
Bayesian estimation
Model fit
Classification
Inference

ABSTRACT

This article introduces a versatile functional form for Generalized Linear Models (GLMs) through a simple, yet effective, transformation of the current framework. The models are applied through a new hierarchical bayesian estimation procedure for logistic regression to highly-cited papers in the Management Information Systems (MIS) field. The results are uniformly better, in regards to model fit and inference for in-sample and out-of-sample data, for simulation studies and real-world data applications, requiring very little time to convergence to true population parameters. In simulation studies, I show that the method contains the true parameters nearly three times as often as widely used existing GLMs, and does so while having confidence intervals that are 54.50% smaller, while requiring around two-thirds the number of MCMC iterations as existing bayesian methods. In Scientometric applications the methodology is shown to be highly robust with predictive/classification accuracy, either equaling or exceeding existing methods for identifying highly-cited articles including Artificial Neural Networks (ANN). Thus, the method is shown to be robust to the amount of asymmetry (or symmetry) of the probability of success (or failure) and robust to unbalanced samples and varying Data Generating Processes. Further, the methodology is equivalent to current methods if the data support them and is therefore complementary to existing methods, without loss of interpretability of model parameters. For the MIS field it finds that Popularity Parameter (PP) of an article Keywords can predict whether a paper will be highly-cited (top 25% of highly-cited articles) between two to three years after publication and beyond. Furthermore, given the small number of iterations needed for convergence, the methodology can also be used as a baseline method in Big Data (BD) settings for both Artificial Intelligence (AI) and Machine Learning (ML) contexts as well.

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 (Abrishami & Aliakbary, 2019; Wang, Wang, & Chen, 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

E-mail address: dukechowdhury@icloud.com

https://doi.org/10.1016/j.joi.2020.101112
1751-1577/

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 Outcome" (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 Agresti & 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, Andrews, Currim, & Jones, 2000; Maity, Pradhan, & Das, 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

$$E[\mathbf{Y}|\mathbf{X}] = \mathbf{c}(\mathbf{X})\beta + E(\epsilon|\boldsymbol{X}) = \mathbf{c(X)}\boldsymbol{\beta}. \tag{1}$$

Here the $n \times 1$ outcome variable Y is related to a $n \times (k+1)$ set of explanatory variables, $\mathbf{X} = (1, X_1, \ldots, X_k)$, through a continuous, bounded, real valued function $\mathbf{c}(\mathbf{X})$ of the same dimensions. The $(k+1) \times 1$ parameters of interest are $\beta = \{\beta_1, \ldots, \beta_{k+1}\}$. If $\mathbf{c}(\mathbf{X}) = I(\mathbf{X})$, where $I()$ is the identity 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. 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, Mai, Shen, & Yan, 2018). This is because they serve as the building blocks for various Multinomial extensions (e.g., Allenby & Rossi, 1998; Murad, Fleischman, Sadetzki, Geyer, & Freedman, 2003). However, their usage appears to be field specific. For example, Hu, Tai, Liu, and Cai (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, Zhang, & Lee, 2019; Bornmann, Leydesdorff, & Wang, 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 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 & Rubin, 2015 for an excellent summary), as well as to understand the degree of heterogeneity through finite and infinite mixture distributions (Andrews, Ansari, & Currim, 2002). In Psychology (e.g., Talukdar, 2008; Hofmans, 2017); Experimental Economics (e.g., Edelman, Luca, & Svirsky, 2017; Hallsworth, List, Metcalfe, & Vlaev, 2017); Biomedical Sciences, (e.g., Zhang et al., 2017, Davison et al., 2017; Mandal, 2017) and in the Physical Sciences (e.g., Hattab et al., 2018; Beitia-Antero, Yáñez, & de Castro, 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] 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

[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.

when the assumptions of more traditional models are violated. Finally, I reintroduce an ROC based predictive statistic ROC-Statistic (RS) (Chowdhury, 2017) 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. 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 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 4 (the frequentist algorithm is deferred to the supplemental materials [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 5. This is followed by extensive numerical simulations in Section 6 and an application to classification of highly-cited papers in Section 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 8 and finally end with some concluding thoughts in Section 9.

2. Preliminaries

Any Bernoulli outcome regression model with accompanying covariates $(\mathbf{Y}, \mathbf{X})$ (defined as in the introduction in Section 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 1. $y_i \in \mathbf{Y}$ are independent and Bernoulli distributed such that

$$y_i = \begin{cases} 1 & with \quad probability \quad p \\ 0 & with \quad probability \quad 1-p. \end{cases} \tag{2}$$

Assumption 2. The systematic component of the explanatory variables are considered fixed.

Assumption 3. There is a link function $g(\boldsymbol{\mu}) = \mathbf{Y}$ that relates the mean of an observation to the systematic component.

In contrast the LV models have Assumptions 4 and 5.

Assumption 4. $y_i \in \mathbf{Y}$ 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,

$$y_i = \begin{cases} 1 \ if \ y_i^* > m \\ 0 \ if \ y_i^* \leq m. \end{cases} \tag{3}$$

Assumption 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 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 ith observation. Then,

$$\mathbf{P}(\mathbf{x_i}) = \frac{\exp[\lambda(\mathbf{x_i}, \beta)]}{(1 + \exp[\lambda(\mathbf{x_i}, \beta)])}, \tag{4}$$

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

$$\ln\{\mathbf{P}(\mathbf{x_i})/(1 - \mathbf{P}(\mathbf{x_i}))\} = \lambda(\mathbf{x_i}, \beta). \tag{5}$$

The quantity on the left in (5) is the familiar log-odds ratio for the logit and is the link function assumed to hold in expectation (average) in Assumptions 1–3 for i.i.d. $y_i \in Y$ (for the probit model as in 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,

$$y_i^* = \beta' \mathbf{x_i} + \epsilon_i^* \Leftrightarrow \mathbf{P}(\mathbf{x_i}) = Pr(y_i^* > m) = F(-\epsilon_i^* < \beta' \mathbf{x_i}) = F(\beta' \mathbf{x_i}), \quad m = 0, \tag{6}$$

[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 & Trivedi, 2010).

Evidently, though no assumption is made on the link function specification in LV models, by construction of GLM the link condition Assumption 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.

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]

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,

$$g(\mathbf{P}(\mathbf{X})) = \lambda(\mathbf{X}, \beta) = \mathbf{c}(\mathbf{X})\beta \Leftrightarrow \mathbf{P}(\mathbf{X}) - g^{-1}(\mathbf{c}(\mathbf{X})\beta) = \mathbf{0}_{n\times 1}. \tag{7}$$

Since we know that (7) is not always satisfied for the ith 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,

$$\mathbf{P}(\mathbf{X}) = (g^{-1}(\mathbf{c}(\mathbf{X}))\beta)^{\boldsymbol{\alpha}^*} \Leftrightarrow \boldsymbol{\alpha}^* = \log(\mathbf{P}(\mathbf{X}))(\log((g^{-1}(\mathbf{c}(\mathbf{X})\beta))))^{-1}, \quad \boldsymbol{\alpha}^* \in \mathbf{R}^{\mathbf{n}}. \tag{8}$$

Since for certain link functions (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 observation 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 ith observation the following condition holds,

$$y_i = g(\mathbf{x}_i, \beta, \alpha_i, \delta_i) = \lambda(\mathbf{c}(x_i), \beta). \tag{9}$$

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,

$$\textit{Hypothesized Link} : g_0(\mathbf{x}_i, \beta; \alpha_i, \delta_i) = g(\mathbf{x}_i, \beta; \alpha_i = \alpha_0, \delta_i = \delta_0), \tag{10}$$

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 $\{\alpha_i^*, \delta_i^*\}$,

$$\textit{Correct Link} : g^*(\mathbf{x}_i, \beta; \alpha_i, \delta_i) = g(\mathbf{x}_i, \beta; \alpha_i = \alpha_i^*, \delta_i = \delta_i^*). \tag{11}$$

Crucially, (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,

$$\textit{argmin}_\beta (Y - \mathbf{c}(\mathbf{X})\beta)^d \ \ s.t. \ \ g^*(\mathbf{x}_i, \beta; \alpha_i, \delta_i) = \lambda(\mathbf{c}(\mathbf{x_i}), \beta), \quad \forall i \in \{1, \ldots, n\}; \quad 1 \le d < \infty, \tag{12}$$

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

where p represents the appropriate p-norm in $L^p(E)$[5] . The proof of this statement can be found in Theorem 5. Since by construction, observations are independent, one can further impose

$$\mathbf{E}(\alpha^*) = \mathbf{E}(\alpha_i); \quad \mathbf{E}(\delta^*) = \mathbf{E}(\delta_i), \tag{13}$$

which follows easily from our identically distributed assumption. Thus,

$$\mathbf{E}[g^*(\mathbf{x}_i, \beta; \alpha_i, \delta_i)] = \mathbf{E}[\lambda(\mathbf{c}(\mathbf{x_i}), \beta)] = \mathbf{E}[\lambda(\mathbf{c}(\mathbf{X}), \beta)], \quad \forall i \in \{1, \ldots, n\}. \tag{14}$$

Therefore, we need only ensure that this assumption holds for each of the ith observations. To show the importance of this formulation, below I first prove the existence and uniqueness of (11) for the logistic LM, 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.

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.

$$g_0(\mu, \alpha, \delta) = \frac{\mathbf{P}^{\alpha}}{(\mathbf{1} - \mathbf{P})}^{\delta}. \tag{15}$$

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

$$g_0(\mu, \alpha^*, \delta^* = 1) = \mathbf{P}^{\boldsymbol{\alpha}^*}(\mathbf{1} - \mathbf{P})^{-1}. \tag{16}$$

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 ith observation consider,

$$g_0(\mu, \alpha_i, \delta_i)^{(1/\delta_i)} = \frac{P_i^{\alpha_i/\delta_i}}{(1 - P_i)}, \tag{17}$$

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

$$\lim_{\alpha_i \to \infty} \frac{P_i^{\alpha_i/\delta_i}}{(1 - P_i)} = 0 \quad \text{and} \quad \lim_{\alpha_i \to -\infty} \frac{P_i^{\alpha_i/\delta_i}}{(1 - P_i)} = \infty. \tag{18}$$

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 $\infty, \quad -\infty$ *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

$$\mathbf{x_i} \textit{ such that } \{P_i : \{\alpha_i, \delta_i\} \notin \{-\infty\, \infty\} \quad \text{and} \quad \delta_i \neq 0\}. \tag{19}$$

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

$$\alpha_i \in R\backslash\{-\infty, \infty\}, \tag{20}$$

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,

$$g_0(\mu, \alpha_i, \delta_i)^{(1/\delta_i)} = g_0(\mu, \alpha_i^*, \delta_i^* = 1), \tag{21}$$

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 1. *There exists a family of link functions given by a monotonic transformation of the Generalized Odds function,* $\mathbf{P}^{\boldsymbol{\alpha}^*}(\mathbf{1} - \mathbf{P})^{-1}$ *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 Online Appendix [1.1], and as a result the model can interpolate

[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.

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

Theorem 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 establishes these results.

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* $\alpha^* \in R^n \backslash \{-\infty, \infty\}$*, given* $\mathbf{P_i} \notin \{0, 1\}$*,* $\mathbf{x_i} \notin \{0, \infty, -\infty\}$ *for each* $i \in \{1, \ldots, n\}$ *and* $\beta_j \notin \{\infty, -\infty\}$ *with* $j \in \{1, \ldots, (k+1)\}$*.*

A somewhat technical proof of this result is given in the Online Appendix [1.2]. Using this result, I provide the foundations for the extension of the specific result to all GLMs below.

Theorem 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* $\alpha^* \in R^n \backslash \{-\infty, \infty\}$*, given* $\mathbf{P_i} \notin \{0, 1\}$*,* $\mathbf{x_i} \notin \{0, \infty, -\infty\}$ *for each* $i \in \{1, \ldots, n\}$ *and* $\beta_j \notin \{\infty, -\infty\}$ *with* $j \in \{1, \ldots, (k+1)\}$*.*

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,

$$\alpha^* = \log(P(\mathbf{X}))(\log((g^{-1}(\mathbf{c}(\mathbf{X})\beta))))^{-1}, \quad \alpha^* \in \mathbf{R^n} \backslash \{-\infty, \infty\}. \tag{22}$$

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

$$\mathbf{P}(\mathbf{X}) \neq \mathbf{0}_{n\times 1} \quad \text{and} \quad g^{-1}(\mathbf{c}(\mathbf{X})\beta)) \neq \mathbf{0}_{n\times 1}, \tag{23}$$

(22) has a specification and a solution, by the same argument as I had proceeded with the Generalized Logistics 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,

$$\mathbf{E}[g(\mathbf{P}(\mathbf{X}))] = \mathbf{E}[(\beta' \mathbf{c}(\mathbf{x_i}))]. \tag{24}$$

Therefore, a sufficient condition for (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 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.* $g^*(\mathbf{x}_i, \beta; \alpha_i, \delta_i) = \lambda(\mathbf{c}(\mathbf{x_i}), \beta), \ \forall i \in \{1, \ldots, n\}; 1 \leq d < \infty; \{|\mathbf{c}(\mathbf{X})|, |\beta|\} < \infty$.

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 $\mathbf{R}\backslash\{\infty, -\infty\}$. Let $E \subseteq \mathbf{R}\backslash\{\infty, -\infty\}$. Then for each continuous and bounded function f in $L^p(E)$ there exists closed and bounded intervals $[a, b]_j$ such that

$$\cup_{j=1}^{n}[a, b]_j = E, \ \ n \in \{1, 2, \ldots\} \tag{25}$$

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 $\mathbf{R}\backslash\{-\infty, \infty\}$ 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 (online 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 5.

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 online appendix [1.2]). The extension of these algorithms to any GLM, follows similarly from the existence and uniqueness results discussed previously.

4.1. Estimation of the Generalized Logistic Link

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

$$\log\left\{\frac{\mathbf{P}^{\alpha^*}}{(1-\mathbf{P})}\right\} = \lambda(\mathbf{c}(\mathbf{X}),\beta) \Leftrightarrow \log\left\{\frac{F(\lambda(\mathbf{c}(\mathbf{X}),\beta))^{\alpha^*}}{(1-F(\lambda(\mathbf{c}(\mathbf{X}),\beta)))}\right\} - \lambda(\mathbf{c}(\mathbf{X}),\beta) = 0, \tag{26}$$

$$\Rightarrow \alpha^* = \frac{\lambda(\mathbf{c}(\mathbf{X}),\beta) + \log(1-F(\lambda(\mathbf{c}(\mathbf{X}),\beta)))}{\log(F(\lambda(\mathbf{c}(\mathbf{X}),\beta)))}. \tag{27}$$

Clearly, there is no analytical solution here for $\beta|\alpha^*$. However, for any particular value of β, we can solve for $\left\{\alpha^*|\beta, \delta^* = 1\right\}$ 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 (27) 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 [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 supplemental materials for this paper [1.2] while the bayesian formulation is given in [4.2].

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]

$$y_i^* = \lambda(\mathbf{c}(x_i),\beta) + \epsilon_i^*, \tag{28}$$

$$y_i^* \overset{i.i.d.}{\sim} Logistic(\lambda(\mathbf{c}(x_i),\beta), \pi^2/3), \tag{29}$$

the full probability model can be written as,

$$p(\beta,\alpha^*|y) = \int_{y^*} p(\alpha^*,\beta,y^*|y) \Leftarrow p(\beta|\alpha^*,y) \propto L(X,\beta)p(\alpha^*|\beta,y)p(\beta). \tag{31}$$

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

$$F(\epsilon_i^*) \overset{i.i.d.}{\sim} Logistic(0, \pi^2/3), \tag{32}$$

$$f(\alpha_i^*|\mathbf{x_i},\beta) \overset{i.i.d.}{\sim} \theta \exp(-\theta g(\mathbf{x_i},\beta)), \tag{33}$$

$$g(\mathbf{x_i},\beta) = \frac{\lambda(c(\mathbf{x_i}),\beta) + \log(1-F(\lambda(c(\mathbf{x_i}),\beta)))}{\log(F(\lambda(c(\mathbf{x_i}),\beta)))}, \tag{34}$$

$$f(\beta) \sim N(\mu_0, \sigma_0^2). \tag{35}$$

Thus, for suitable values of the hyper-parameters (Sections 6 and 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.

[8] Please note that fixing the variance parameter is according to the formulation given in Albert and Chib in Albert and Chib (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.

1. Draw from the truncated logistic distribution for each observation.
2. Given the realized values of $y_i'^{*}s$, draw from $f(\alpha^*|X, \beta)$, making any transformations as necessary for every observation.
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.

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 $E(\alpha^*|\beta)$ for existing models such as logit ($\alpha^* = \mathbf{1}$), we can use large-sample results under independence through Assumptions 1 and 4 to test the hypothesis that our model results vary from traditional GLM fits. In particular, we know for GLM,

$$E[\alpha^*|\beta] = \log(P(X))(\log((g^{-1}(c(X)\beta)))\,)^{-1}. \tag{36}$$

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

$$\alpha^* \sim N(E(\alpha^*|\beta^*), E(\alpha^* - E(\alpha^*|\beta^*)|\beta^*)^2), \tag{37}$$

$$\hat{\bar{\alpha^*}} \overset{asymp.}{\sim} N(\frac{\sum_{i=1}^{n}\alpha_i}{n}, \frac{\sum_{i=1}^{n}(\alpha_i - \alpha)^2}{n-1}). \tag{38}$$

β^* above represents the optimized estimated value. Thus, we can check our hypothesis that $\bar{\alpha^*} = k$, for some $k \in \mathbf{R}\backslash\{-\infty, \infty, 0\}$ 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 $\mathbf{E}[\alpha^*] = \mathbf{E}[\delta^*] = 1$.

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,

$$\mathbf{Y} = Intercept + \mathbf{X_1} + (\mathbf{X_2})^2, \tag{39}$$

$$\mathbf{Y} = Intercept + \mathbf{X_1} + \exp(\mathbf{X_2}), \tag{40}$$

$$\mathbf{Y} = Intercept + \exp(\mathbf{X_1}) + \sin(\mathbf{X_2}). \tag{41}$$

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

[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.

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

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 observation 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 jth independent variable ($j \in \{1, 2\}$) for the ith row.[11] [12] Thus, we can consider the calculated **Y** values along with the generated **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 β's compared to actual β's, number of MCMC iterations required and Akaike Information Criteria (AIC). Where the AIC is defined as,

$$2 \times (-\log - likelihood + Number\ \ of\ \ Parameters)/Number\ \ of\ \ Observations \tag{42}$$

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 1 and Fig. 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 (Fig. 1). However, most importantly, the proposed model contained the true parameters 84.20% (341 out of 405 total) of the time, as 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

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

Table 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	1849.95	7.65	4.76
β_2 C.I. Rng.	1.96	279.44	4.54	3.95
β_3 C.I. Rng.	2.33	9174.67	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 8000 MCMC iterations with a 4000 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).

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.

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 25th 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 (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 25th 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 number of topics and let N be the number of keywords in each topic obtained from the LDA analysis. Define $k_{m,n}$ $\{m \in M,\ \ n \in N\}$ as the nth keyword for the mth topic. Thus, from ResearchGate for each $k_{m,n}$ we can obtain the number of questions ($q_{m,n}$) related to it. From

[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.

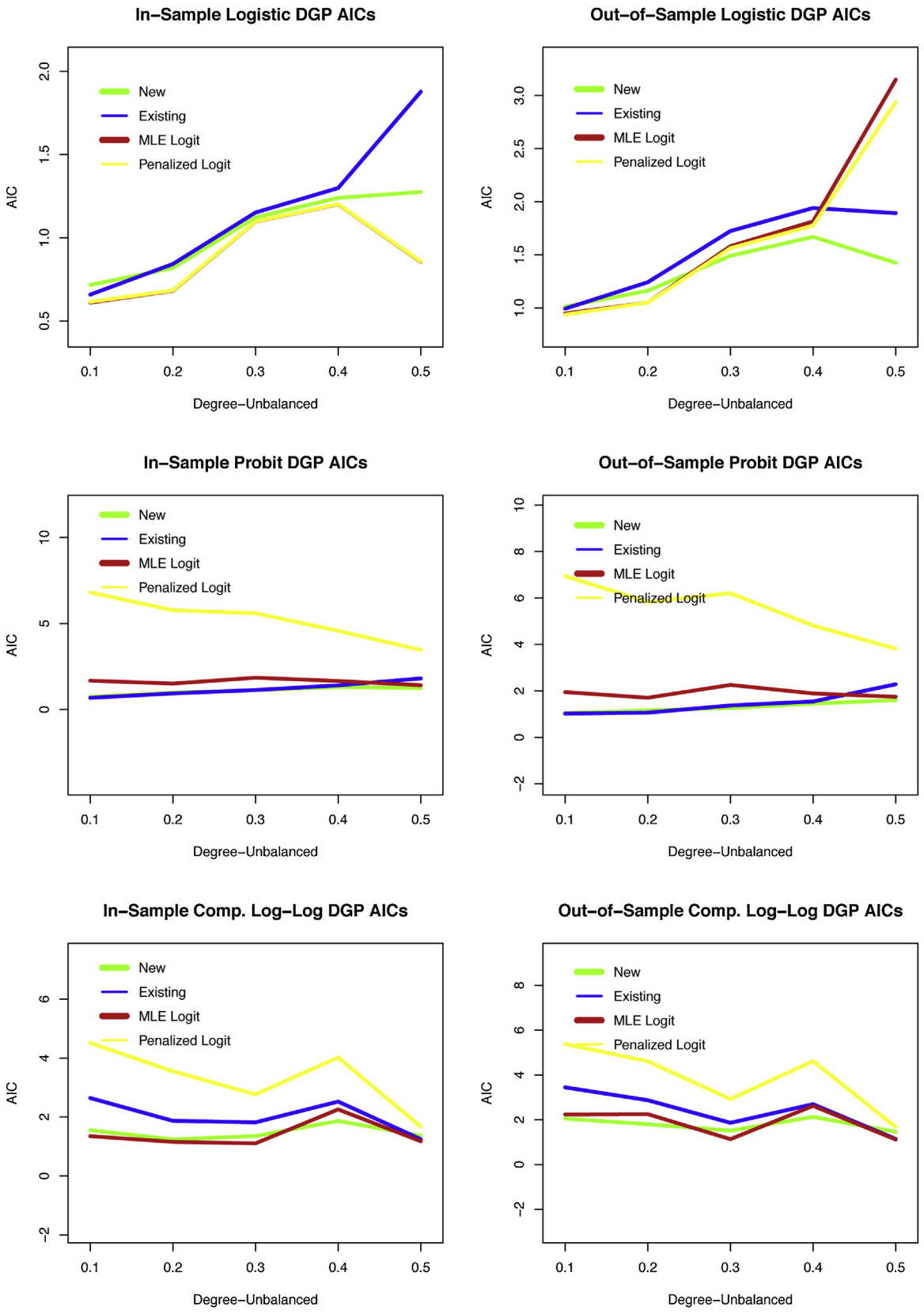

Fig. 1. Simulation results summary for all three DGPs in-sample and out-of-sample for all three linear and non-linear models considered. Note: The AICs for MLE-Logit and penalized logit were ∞ for multiple datasets and thus only finite AIC values are graphed here. Comp. Log–Log refers to Complementary

Table 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)

Table 3
ROC-statistic for Management Information System and marketing.

	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.

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 $i \in \{1, \ldots, I\}$, with i defined as the total number of articles in the jth year. Accordingly, for the jth year the measures can be defined as follows:

$$TP^i_j = \sum_{m=1}^{M}\sum_{n=1}^{N} q_{m,n}\theta^i_m\phi^m_n \tag{43}$$

$$PP^i_j = \sum_{m=1}^{M}\sum_{n=1}^{N} p^j_{m,n}\theta^i_m\phi^m_n \tag{44}$$

$$NP^i_j = \sum_{m=1}^{M}\sum_{n=1}^{N} e^j_{m,n}\theta^i_m\phi^m_n \tag{45}$$

$$WPP^i_j = \sum_{m=1}^{M}\sum_{n=1}^{N} w^j_{m,n}\theta^i_m\phi^m_n \tag{46}$$

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.

$$VP_j^i = \sum_{m=1}^{M}\sum_{n=1}^{N} v_{m,n}^j \theta_m^i \phi_n^m. \tag{47}$$

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 (42) 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).

The statistic is defined as follows,

$$ROC - Statistic = \frac{FP}{TP}, \tag{48}$$

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 population DGP (represented by the out-of-sample hold-out data). Therefore, the underlying assumptions is that MIP performances based on true population parameters should be more robust and give better models fits out-of-sample, in addition to having reasonable MIP 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, 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.

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,

$$Highly\ \ Cited = Intercept + Journal\ \ Impact\ \ Factor + Popularity\ \ Parameter. \tag{49}$$

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 $f(\alpha^*|X, \beta)$, 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 fit 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

Table 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.

the ROC-Statistics can be found in Table 3 and the AIC based summary is given in Table 4. The estimates with the relevant standard errors can be found in Tables 5 and 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 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 0.05α-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, Yi, and Lee (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.

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

Table 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.

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.

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,

[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.

Table 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.

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 3) and model fit (Table 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 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 comparison 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 logistic and Penalized logistic models 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

[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.

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 & 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 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 **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 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.

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.

[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, Garnier, Papanicolaou, & Solna, 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]).

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

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.

Author statement

The author declares that this article is the result of my own work from inception to implementation and from conceptualization to verification of all proofs and computer codes relevant to the manuscript. All codes relevant to the implementation of the material is also the sole work of myself. In addition, the manuscript is written by me including all versions submitted for the review process.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.joi.2020.101112.

References

Abramson, C., Andrews, R. L., Currim, I. S., & 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., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, *13*(2), 485–499.

Agresti, A., & Kateri, M. (2011). *Categorical data analysis*. Springer.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, *89*(1–2), 57–78.

Andrews, R. L., Ansari, A., & 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., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics*, *13*(1), 407–418.

Beitia-Antero, L., Yáñez, J., & de Castro, A. I. G. (2018). On the use of logistic regression for stellar classification. *Experimental Astronomy*, *45*(3), 379–395.

Bornmann, L., Leydesdorff, L., & 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., & Trivedi, P. K. (2010). . *Microeconometrics using stata* (Vol. 2) TX: Stata Press College Station.

Choi, J., Yi, S., & 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. (2017). Supervised machine learning and heuristic algorithms for outlier detection in irregular spatiotemporal datasets. *Journal of Environmental Informatics*, *S.l*, 1–16. ISSN 1684-8799. Available at: <http://www.jeionline.org/index.php?journal=mys&page=article&op=view&path%5B%5D=201700375>. Date accessed: 28 Dec. 2020

Davison, N., Warren, R., Mason, K., McElhone, K., Kirby, B., Burden, A., et al. (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.

Edelman, B., Luca, M., & 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., & Solna, K. (2007). . *Wave propagation and time reversal in randomly layered media* (Vol. 56) Springer Science & Business Media.

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

Guerrero, V. M., & 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., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, *148*, 14–31.

Hattab, M., de Souza, R., Ciardi, B., Paardekooper, J., Khochfar, S., & 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., & 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., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Li, K., Mai, F., Shen, R., & Yan, X. (2018). *Measuring corporate culture using machine learning*. pp. 3256608. Available at SSRN.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*(2), 81.

Maity, A. K., Pradhan, V., & Das, U. (2018). Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. *The American Statistician*, 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., & 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., & Nguyen, C.-T. (2008). *A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference. JGibbLDA*. 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.

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., & 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.

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., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics*, *10*(4), 1166–1177.
Wang, M., Wang, Z., & 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., & 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., et al. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics*, *12*(4), 1099–1117.