2019-07-18

Bayesian Analysis of Immigration in Europe with Generalized Logistic Regression

Dalla Valle, Luciana

10.1080/02664763.2019.1642310
Journal of Applied Statistics
Taylor & Francis (Routledge)

Bayesian Analysis of Immigration in Europe with Generalized Logistic Regression

Luciana Dalla Valle [c], Fabrizio Leisen[a], Luca Rossini[d] and Weixuan Zhu[b]

[a] University of Kent, U.K.; [b] Xiamen University, China; [c] University of Plymouth, U.K. [d] Vrije Universiteit Amsterdam, The Netherlands

ARTICLE HISTORY
Compiled July 6, 2019

ABSTRACT
The number of immigrants moving to and settling in Europe has increased over the past decade, making migration one of the most topical and pressing issues in European politics. It is without a doubt that immigration has multiple impacts, in terms of economy, society and culture, on the European Union. It is fundamental to policymakers to correctly evaluate people's attitudes towards immigration when designing integration policies. Of critical interest is to properly discriminate between subjects who are favourable towards immigration from those who are against it. Public opinions on migration are typically coded as binary responses in surveys. However, traditional methods, such as the standard logistic regression, may suffer from computational issues and are often not able to accurately model survey information. In this paper we propose an efficient Bayesian approach for modelling binary response data based on the generalized logistic regression. We show how the proposed approach provides an increased flexibility compared to traditional methods, due to its ability to capture heavy and light tails. The power of our methodology is tested through simulation studies and is illustrated using European Social Survey data on immigration collected in different European countries in 2016–17.

KEYWORDS
Bayesian Inference; Generalized Logistic Regression; Empirical Likelihood; Immigration

1. Introduction

Human migration is a well-known phenomenon that dates back to the earliest periods of human history and that continues to provide opportunities as to societies as well as migrants. However, in recent times migration has proved to be a key political and policy challenge in matters such as integration, displacement, safe migration and border management [25]. People migrate for many different reasons which can be classified as economic, social, political or environmental. Some people voluntarily choose to migrate, motivated by the attractiveness of higher wages and job opportunities. On the other hand, some are forced to migrate, for reasons such as famine, natural disasters and war. One of the biggest drivers of migration in recent years has been the Syrian civil war, also known as Syrian crisis, which caused about 5.65 million people to leave the country since the start of the conflict in 2011 [7, 39]. In addition, conflicts and per-

CONTACT: Weixuan Zhu. Email: zhuweixuan@gmail.com

secutions in other parts of the world, such as the ongoing violence in Afghanistan and Iraq, abuses in Eritrea, political instability in Sudan as well as poverty in Kosovo, are forcing people to migrate. Understanding migration is fundamental for policy makers in order to effectively address evolving migration dynamics, while at the same time adequately accounting for the diverse and varied needs of migrants.

Migration to Europe has recently emerged as a critical policy issue within the European Union. Although Europe has always been characterized by large population movements, in the past three decades many countries have experienced large inflows of immigrants. An estimated 362,000 refugees and migrants risked their lives crossing the Mediterranean Sea in 2016, with 181,400 people arriving in Italy and 173,450 in Greece. In the first half of 2017, over 105,000 refugees and migrants entered Europe. Germany is the main country of destination which accounted for 31% applicants registered in 2017, followed by Italy (20%), France (14%), Greece (9%), the UK (5%) and Spain (5%) [15]. Immigration in Europe is currently a major topic of academic, policy and public concern. Major debates have arisen over refugee inflows and the recognition that many of these migrants will settle permanently in their host countries. The successful integration of immigrants, who are identifiable as ethnically different from their host countries' inhabitants and who may hold different cultural and religious values, is one of the major challenges for Europe. Opinions about migration and its effects are becoming more divergent among Europeans, according on age, education, social class and migrant heritage, and favourability towards immigration varies considerably by country [23]. In particular, some of the countries that have seen the largest migrant inflows have become more sensitive to threats from migration and this has caused an increase in anti-immigration sentiments. In the UK, the widespread concern over the numbers of people moving to the country under the EU's freedom of movement rules, was one of the main topics of discussion of the 2016 Brexit referendum.

In this climate, the political community is struggling to balance the needs of refugees, the concerns of the native population and the demands of employers. The public's views towards migration have important implications for debates surrounding the constitutional future of European countries, as states are trying to address public concern about immigration. In democratic political systems, such as European countries, in which immigration is a salient issue, public opinion has an important role in shaping immigration policy [5]. For example, in Britain, public preferences for less immigration have been among the drivers of the British immigration policy, including restrictions aimed at reaching a numerical target for estimated annual net migration. The government has explicitly claimed that its motivation to reduce the number of immigrants coming to Britain is a response to public opinion, tying its drive to reduce net migration to public concern about immigration [37, 38]. In order to identify and promote effective immigration laws and integration policies in European countries, it is fundamental to capture the society's attitude towards immigration and integration. The lack of understanding of the public's attitude towards immigration may be one of the cause of the inability of some governments to fulfill public demands for specific multicultural policy. The correct appraisal of individual sentiments towards refugees and asylum-seekers in different European countries is essential for policy makers and governmental bodies, in order to acquire a deeper understanding of social issues regarding immigration to improve laws and policies related to immigration. New research aiming at developing statistical models accurately estimating and predicting public's views towards immigration is crucial.

The logistic regression model for binary and multinomial responses has been routinely used in applied works for estimating and predicting immigration-related data

[2, 6, 20, 35, 41]. However, Bayesian inference for models with binomial likelihood has been out of the radar for a long time, due to estimation difficulties for the analytically inconvenient form of the likelihood function. In their seminal paper, [1] developed an exact Bayesian method for modelling categorical responses using a data augmentation algorithm applied to probit regression. This auxiliary variable approach for Bayesian probit models has been widely employed both in political science and in market research [26, 33]. Following this direction, many authors in the literature applied the same strategy to logit models, using approximations or complex extensions of the method proposed by [1] [17, 18, 21, 24]. However, Bayesian logit models have been less popular among non-statisticians than their probit counterparts, due their computational inefficiency and complexity, since they often rely on analytic approximations or numerical integrations, and are based on multiple layers of latent variables. [32] overcame these issues by introducing a new approach based on the class of Polya-Gamma distributions and by proposing a data-augmentation algorithm for the Bayesian logistic regression. [9] proved that the Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic and it guarantees the existence of a central limit theorem for Monte Carlo averages of posterior draws. [32] showed that their method outperformed previous approaches, both in ease of use and in computational efficiency. The authors opened the door of Bayesian logit models in various areas, such as network analysis, among others, leading to the publications by [11] and [12], who applied Bayesian logit models to financial and brain networks.

In the literature of logistic regression, a prolific stream of research is devoted to the introduction of flexibility into the model through the link function. [10] pointed out the consequences of link misspecification, which may lead to increased mean squared error and bias as of parameter estimates as well as predicted probabilities. In order to address these issues, [36] proposed a two-parameter class of generalized logistic models, that can approximate several symmetric and asymmetric link functions. However, in the Bayesian framework, this approach may lead to improper posteriors when noninformative improper priors, such as the improper uniform prior, are used for the regression coefficients. A different class of skewed links, yielding proper posterior distributions for the regression parameters with standard improper priors was proposed by [8]. Other contributions following this line of reaserch include the works of [29], who proposed generalized skewed-t link models using a latent variable approach; [40], who introduced a generalized extreme value link function, producing an adaptable family of models for binary data; [27], who proposed a class of symmetric power link functions, by introducing an additional power parameter on the cumulative distribution function corresponding to standard link functions; and [34], who illustrated an efficient estimation approach for the link function parameters in a Bayesian robit model.

This paper proposes a novel Bayesian approach, based on the generalized logistic regression, which introduces flexibility into the model through a new parameter of interest. The proposed approach extends the standard logistic regression model including an additional parameter, the tail parameter, which allows us to treat heavy and light tails. According to our knowledge, this is the first paper analysing and discussing the generalized regression model constructed as in [28]. Unfortunately, the likelihood function of the proposed generalized logistic regression model inherits the analytic inconvenient form of logistic models and has an additional source of complexity due to the tail parameter. To overcome these issues, we propose a novel approximate Bayesian approach that produces consistent and fast results, based on the empirical likelihood [30]. This approach is particularly suitable to address problems of intractable and complex likelihoods, such as the proposed generalized logistic model, and shows excellent

performance.

The rest of the paper is organized as follows. In Section 2, we describe the generalized logistic distribution and the regression model related to it. Section 3 illustrates the empirical likelihood strategy adopted in the paper. To validate the approach, simulated experiments on different datasets are studied in Section 4. Then, we introduce the EU immigration data and we motivate the choice of the model in Section 5. In Section 6 we fit the proposed model to the EU immigration data. Section 7 is left for final remarks.

2. Generalized Logistic Regression

In this paper, we aim at accurately estimate people's views towards immigration proposing a flexible Bayesian generalised logistic regression model. We consider a binary regression setup, in which we have n independent binary random variables $y_1, \ldots, y_n$ distributed as Bernoulli with probability of success

$$\Pr(y_i = 1|\boldsymbol{\beta}) = H(x_i^T \boldsymbol{\beta}) \tag{1}$$

where $x_i^T = (x_{i1}, \ldots, x_{ik})$ is a vector of known covariates associated to y_i, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown regression coefficients and $H : \mathbb{R} \to (0, 1)$ is a known cumulative distribution function. Probit regression assumes $H(x) = \Phi(x)$ where Φ is the cumulative distribution function of the Normal distribution. Logistic regression assumes $H(x) = S(x)$ where:

$$S(x) = \frac{e^x}{1 + e^x}, \qquad x \in \mathbb{R}. \tag{2}$$

The standard logistic distribution is described by the above cumulative distribution function and the below density function:

$$s(x) = \frac{e^x}{(1 + e^x)^2}, \qquad x \in \mathbb{R}. \tag{3}$$

Following [28], it is possible to generalize the distribution described in equation (3). Let $g(\cdot)$ be the density function of a Beta distribution with parameters (p, p) where $p > 0$. [28] considers the following transformation:

$$f(x) = g[S(x)]s(x) \tag{4}$$

where $S(x)$ and $s(x)$ are defined as equation (2) and equation (3), respectively. It is easy to see that

$$f(x) = \frac{1}{B(p, p)} \frac{e^{px}}{(1 + e^x)^{2p}} \qquad x \in \mathbb{R}. \tag{5}$$

The above distribution is known in the literature as the *Type III Generalized Logistic Distribution* (GLD). The cumulative distribution function is known up to an hypergeometric function. It could also be represented through an incomplete Beta function

as:

$$F(x) = \frac{1}{B(p,p)} B\left(\frac{e^x}{1+e^x}; p, p\right) \qquad x \in \mathbb{R}. \tag{6}$$

where $B(t; p, p) = \int_0^t x^{p-1}(1-x)^{p-1}dx$, $0 < t < 1$. Figure 1 shows the generalized logistic distribution as in (5) for different values of p. In particular, when $p = 1$, equation (5) is the logistic probability distribution of equation (3). For $p \in (0, 1)$, the GLD has heavy tails whilst for $p > 1$ it has light tails. The bottom panel of Figure 1 shows the generalised logistic cumulative distribution function. Note that for $p \in (0, 1)$ (the heavy tails case), the cumulative distribution function looks smoother, while for $p > 1$ (the light tail case) the cumulative distribution function has a sharper shape.

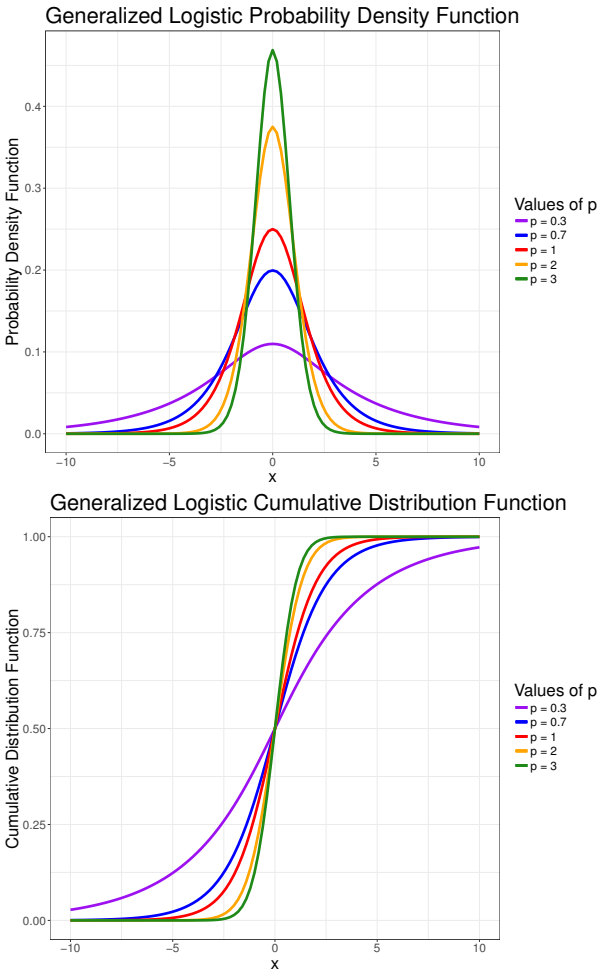

Figure 1.: Probability density function (top panel) and cumulative distribution function (bottom panel) of the generalized logistic for different values of p: $p = 0.3$ (purple line), $p = 0.7$ (blue line), $p = 1$ (red line), $p = 2$ (orange line) and $p = 3$ (green line).

The distribution displayed in equation (5) implicitly assumes that the location is $\mu = 0$ and the scale is $\sigma = 1$. The density function of the GLD with location $\mu \in \mathbb{R}$

and scale $\sigma > 0$ is

$$f(x) = \frac{1}{\sigma B(p,p)} \frac{e^{p\left(\frac{x-\mu}{\sigma}\right)}}{\left(1+e^{\left(\frac{x-\mu}{\sigma}\right)}\right)^{2p}} \qquad x \in \mathbb{R}. \tag{7}$$

To simulate from the above distribution is straightforward. The procedure is as follows.

(1) Generate $Y \sim \text{Beta}(p,p)$.
(2) Compute $X = S^{-1}(Y)$ where S is the cumulative distribution function of the standard logistic distribution in equation (2).
(3) Step 1 and 2 generate an observation from a GLD with $\mu = 0$ and $\sigma = 1$. We can obtain a GLD with different μ and σ by simply multiplying X by the scale parameter σ and adding the location parameter μ.

In this paper, a generalization of the usual logistic regression is proposed by setting $H(x) = F(x)$ in equation (1), where $F(x)$ is the cumulative distribution function introduced in (6). This model is more flexible than the standard logistic regression since it has an extra parameter p which controls the tails of the distributions. The usual logistic regression can be recovered when $p = 1$. Summing up, we consider the following Bayesian model:

$$\begin{aligned} y_i|\boldsymbol{\beta}, p &\sim \mathcal{B}e(F(x_i^T\boldsymbol{\beta})) \\ \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{v}, B) \\ p &\sim \mathcal{G}a(a,b) \end{aligned} \tag{8}$$

where $\mathcal{B}e$ denotes the Bernoulli distribution, $\mathcal{N}$ denotes the Normal distribution, $\mathcal{G}a$ denotes the Gamma distribution, and $\mathbf{v}, B, a$ and b are suitable hyperparameters. Posterior inference for this model is nontrivial due to the form of the cumulative distribution function displayed in equation (6). The next Section will illustrate a computational strategy for efficiently estimating the parameters of the newly proposed model.

3. Inference Sampling Strategy: the Empirical Likelihood Approach

In this section, we describe a parameter estimation approach which is a recent proposal in the literature and makes use of the empirical likelihood (EL). We found that this method is reasonably fast and accurate.

In particular, we follow a novel approximate Bayesian approach for addressing posterior inference proposed by [30] and based on the EL. The authors' idea is to replace the likelihood function with an approximation, called EL [31]. This EL-based approximate Bayesian approach takes advantage of the approximation device provided by the well-established EL to perform posterior inference. This approach is particularly appealing in applications where the likelihood function is complicated or impossible to evaluate. Along the same lines, many other similar algorithms were proposed, for example by [42] and [22].

The empirical likelihood is a robust non-parametric alternative to classical likelihood approach. Assume that we have i.i.d. data $\boldsymbol{y} = (y_1, ..., y_n)$ from a distribution F. Starting by defining the parameters of interest θ as functionals of F, the empirical

likelihood profiles a non-parametric likelihood through a set of constraints of the form

$$\mathbb{E}_F[h(\boldsymbol{y}, \theta)] = 0,$$

where the dimension of h sets the constraints unequivocally defining θ. The EL is defined as

$$L_{EL}(\theta|\boldsymbol{y}) = \max_{p_1,\ldots,p_n} \prod_{i=1}^{n} p_i$$

for $p_i \in [0,1]$ with constraints $\sum_{i=1}^{n} p_i = 1$; $\sum_{i=1}^{n} p_i h(y_i, \theta) = 0$.

Here we describe the Bayesian EL algorithm briefly. Let $L_{el}(\theta_j|\boldsymbol{y})$, with $j = 1, ..., M$, denote the estimate of the empirical likelihood at the point θ_j given the observed data $\boldsymbol{y}$. The so-called Bayesian Computation with the empirical likelihood algorithm (BC_{el}) generates values $\theta_j, j = 1, \ldots, M$, from the prior distribution of θ, and uses the values $L_{el}(\theta_j|\boldsymbol{y})$ as weights in an importance sampling framework. The sampler works as follows

---

***BC*$_{el}$.** Bayesian Computation with the empirical likelihood

---

for $j = 1$ to M do

(1) Generate θ_i from the prior distribution $\pi(\cdot)$
(2) Set the weight $w_j = L_{el}(\theta_j|\boldsymbol{y})$

end for

The output is a sample of size M of parameters with associated weights, which operates as an importance sampling output. This means that a posterior sample of simulated parameters of size N is sampled with replacement from the M parameters with corresponding weights w_j's [30].

For the generalized logistic regression, we set as constraint in the EL approach that the sum of the score functions, namely $\sum_{i=1}^{n} \frac{\partial \mathcal{L}(\theta, y_i)}{\partial \theta}$, is restricted to 0. Since the explicit expressions of the score functions are extremely difficult to obtain due to the incomplete beta function involved, we resort to the R numerical approximation in practice.

4. Simulation Studies

In this section, we analyse the performance of the proposed methodology for different values of p. In particular, we choose three different values of the tail parameter p: $p = 0.1$ (heavy tails), $p = 1$ (logistic regression) and $p = 3$ (light tails), and we test the performance of the BC_{el} algorithm on simulated data. For each value of p, we simulated 20 different datasets of sample sizes $n = 500$ and $n = 1,000$ respectively. We focus on the five dimensional case, where $k = 5$ and the vector of unknown coefficients $\boldsymbol{\beta}$ is a (5×1) vector with both positive and negative values. In particular, we choose the following values for $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$: $\beta_1 = 1$, $\beta_2 = -1$, $\beta_3 = -3$, $\beta_4 = 1$ and $\beta_5 = 3$.

For the simulations, we consider vague priors for $\boldsymbol{\beta}$ and p. More precisely, for $\boldsymbol{\beta}$ we choose a multivariate normal prior, $\mathcal{N}_5(\mathbf{v}, B)$, with mean vector $\mathbf{v} = \mathbf{0}$ and covariance matrix $B = 5 \cdot \mathbb{I}_5$, and for p we choose a gamma prior, $\mathcal{G}a(a, b)$, with hyperparameters $a = b = 1$.

Considering that the dimensionality of the parameters is not low, we propose to use the latin hypercube sampling strategy to sample from the priors, with the goal of achieving the maximum inference by varying multiple parameters at the same time. Essentially, the d-dimensional parameter probability space is divided into $M = 20,000$ equally sized subdivisions in each dimension, and then $M = 20,000$ random samples, one from each sub-division, are sampled.

	p	β_1	β_2	β_3	β_4	β_5
Real value	0.1	1	-1	-3	1	3
$BC_{el},\ n = 500$	0.0899(0.01)	0.9547(0.29)	-1.0052(0.25)	-3.1382(0.28)	1.0888(0.25)	3.2259(0.25)
$BC_{el},\ n = 1000$	0.0895(0.01)	0.9493(0.33)	-1.1088(0.25)	-3.0382(0.28)	1.1270(0.30)	3.0471(0.28)
Real value	1	1	-1	-3	1	3
$BC_{el},\ n = 500$	0.9004(0.13)	1.1542(0.17)	-1.1546(0.14)	-3.4051(0.25)	1.1834(0.17)	3.4388(0.22)
$BC_{el},\ n = 1000$	0.9886(0.34)	1.0681(0.25)	-1.0456(0.35)	-3.2263(0.49)	1.0426(0.26)	3.2398(0.51)
Real value	3	1	-1	-3	1	3
$BC_{el},\ n = 500$	2.1979(0.64)	1.2306(0.15)	-1.2498(0.21)	-3.6796(0.32)	1.1944(0.18)	3.6118(0.32)
$BC_{el},\ n = 1000$	2.2329(0.56)	1.2442(0.12)	-1.2161(0.16)	-3.7258(0.38)	1.2103(0.20)	3.7416(0.45)

Table 1.: Posterior means over the 20 different simulated datasets estimated by using the BC_{el} method compared with the true values of $p = 0.1$ (top), $p = 1$ (middle) and $p = 3$ (bottom) and $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (1, -1, -3, 1, 3)$. The values in brackets are the standard deviations over the 20 different simulations.

Table 1 shows the posterior means over the 20 different simulated datasets, for the three values of p and the 5-dimensional vector of unknown coefficients. The values in brackets are the standard deviations over the 20 different simulations. As one can see in Table 1, the posterior means of p estimated with the EL approach are very close to the true values, particularly in the case of heavy tails. In addition, the posterior means of the vector of unknown $\boldsymbol{\beta}$ converge quickly to the real values.

We then fitted the simulated data assuming both the standard logistic regression and the generalized logistic regression, with the aim of showing the cost of ignoring the tail behaviours of logistic distributions. The two candidate models are compared using the Deviance Information Criterion (DIC) and the Log Pseudo Marginal Likelihood (LPML) [19], where lower DIC or higher LPML values indicate a better-performing model. The comparison in terms of DIC ad LPML is repeated for each one of the 20 simulated datasets and the best performing model is recorded each time. Table 2 shows the percentage of best performance of the generalized logistic over the standard logistic model out of the 20 simulations. From Table 2, we clearly see the advantage of the proposed generalized logistic model over the standard logistic regression. The generalized logistic model performs best when the dataset is simulated with heavy and light tails (corresponding to the scenarios $p = 0.1$ and $p = 3$), 95% and 75% of the times in terms of DIC, and 90% and 80% of the times in terms of LPML, respectively, with $n = 500$. When the sample size increases to $n = 1,000$, the generalized logistic model wins in both scenarios 100% of the times in terms of DIC, and 100% and 95%

of the times in terms of LPML. The standard logistic regression model outperforms the generalized logistic in slightly more simulations when $p = 1$. However, this is not surprising due to the fact that the standard logistic distribution is a special case of the generalized logistic distribution.

	True p	% Lowest DIC	%Highest LPML
$n = 500$	$p = 0.1$	95%	90%
	$p = 1$	40%	45%
	$p = 3$	75%	80%
$n = 1000$	$p = 0.1$	100%	100%
	$p = 1$	45%	45%
	$p = 3$	100%	95%

Table 2.: Percentage of best performance of the generalized logistic model over the standard logistic model out of 20 simulations. The best performance is determined each time by the lowest DIC value and highest LPML value.

5. The Immigration Dataset and Model Motivation

The aim of this paper is to correctly estimate public opinions towards immigration, via the proposed Bayesian generalised logistic regression model. We consider data selected from the European Social Survey (ESS), an academically driven cross-national survey, which has been administered in over 30 countries since 2001. The data have been collected following hour-long face-to-face interviews covering questions on a variety of core topics. In this paper we consider data regarding attitudes towards immigration from the ESS8 edition 1.0 published in October 2017, and collected in Great Britain (GB), Germany (DE) and France (FR) between August 2016 and March 2017 [13, 14]. During that period, the above listed European countries' citizens were heavily exposed by media to news related to events concerning immigration, igniting discussion among members of the government and the public. During the first half of 2017, migrants made more than 30,000 illegal attempts to get into UK from Calais, by crossing the Channel tunnel or by surreptitiously boarding the cargo area of lorries heading for ferries crossing the English Channel [16]. In December 2016, a truck was deliberately driven into the Christmas market in Berlin, leaving 12 people dead and 56 others injured. The perpetrator was a Tunisian failed asylum seeker [3]. In July 2016, a truck was driven into crowds in Nice, resulting in the death of 86 people and injuring 434. The driver was a Tunisian resident in France [4]. Most likely these facts affected public opinions towards refugees and asylum seekers and urged policy makers to develop suitable immigration strategies. The correct appraisal of people's attitude towards migration in different European countries is essential and requires a flexible methodological approach on carefully selected data.

From the ESS data, we considered 12 variables comprising subject-specific information as well as individual opinions. Subject-specific variables include the highest level of education (`edulvlb`), with 26 levels from not completion of primary education to doctoral degree; the household's total net income (`hinctnta`), from the first to the tenth income decile; age (`agea`), from 15 to 100 years old, and the dichotomous variables `rlgblg` and `blgetmg`, indicating whether the interviewee belongs to particular religion or to a minority ethnic group, respectively. Opinion variables include answers

to questions ranging from 0 (most negative opinion) to 10 (most positive opinion), such as: do you think that most people try to take advantage of you (`pplfair`)? Do you trust your country's parliament (`trstprl`)? Do you trust your country's legal system (`trstlgl`)? Do you trust the European parliament (`trstep`)? Do you trust the United Nations (`trstun`)? In addition, to measure personal well-being, we considered the variable (`happy`), where 0 corresponds to extremely unhappy and 10 corresponds to extremely happy. The dependent variable is `immig`, indicating whether the respondent would allow immigrants from poorer countries outside Europe, with `immig` $= 1$ if the respondent is against immigration, and `immig` $= 0$ if the respondent is in favour of immigration. The dependent variable was obtained by dichotomizing the ESS variable `impcntr`.

The total number of observations for the three considered European countries is $5,354$. For individual countries the number of observations are: $1,419$ for GB; $2,284$ for DE and $1,651$ for FR.

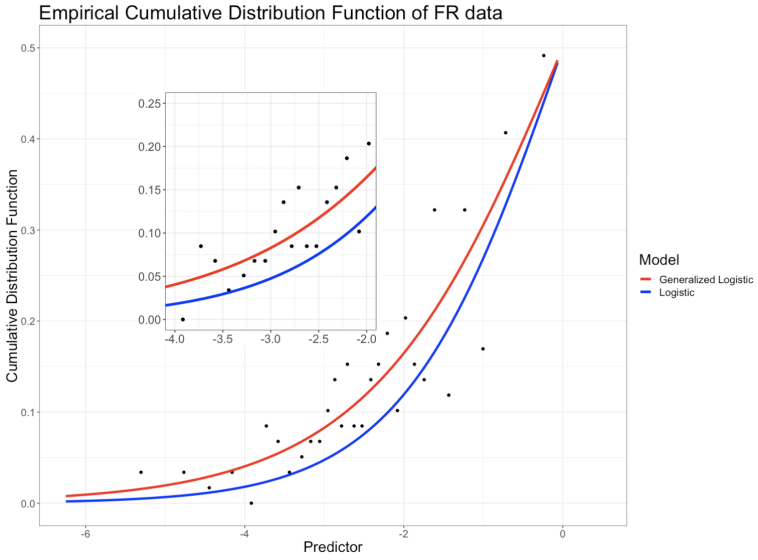

Figure 2.: Cumulative distribution function of FR data fitted with the standard logistic regression model (blue line) and the generalized logistic regression model (red line). The data is categorized into 30 categories based on the linear predictor.

We modelled the probability of `immig` $= 1$ using the remaining 11 variables described above as covariates, according to the standard and to the generalised logistic regression. In order to demonstrate the plausibility of both models, we implemented the Hosmer-Lemeshow test, categorizing the observations into 30 groups based on the linear predictor. Both models pass the test, with a p-value of 0.5107 for the standard, and 0.4398 for the generalized logistic regression. In addition, to compare the fitness of both models to the data, in Figure 2 we plotted their cumulative distribution functions against the linear predictor of the FR data. We display the outcomes obtained using the FR data, since data from the other countries gave very similar results. The points represent the observations, which are categorized in 30 subgroups, as explained before. The blue line shows the cumulative distribution function of the Fr data fitted with the standard logistic regression, while the red line shows the cumulative distribution function of the generalized logistic regression model. Unfortunately, the standard logistic regression does ot fit the data well, due to its rather sharp shape. On the contrary, the the generalized logistic regression models shows a good fit, thanks to its smooth shape,

which is able to capture heavy tails, as illustrated in Section 2, Figure 1. Therefore, the most suitable model for the immigration data is the generalized logistic regression, which is flexible enough to capture different tail shapes and it is able to provide improved estimates and predictive power compared to the standard logistic model.

6. Data Analysis

In this Section, we further demonstrate the suitability of the generalized logistic model compared to the standard logistic regression and we examine the effects of the covariates on the attitude towards immigration for people living in the three different European countries under consideration. As discussed earlier, we first modelled the probability of the respondents being against immigration, that is, `immig`=1, using all the 11 remaining variables from ESS data. In order to select the most informative covariates, we adopted a simple variable selection procedure, by fitting all the three countries' data with a standard logistic regression and excluding the variables giving posterior support to zero. The remaining 7 variables are `pplfair, trstep, trstun, happy, agea, edulvlb` and `hinctnta`. Then, we fitted the remaining covariates for each country with the generalized logistic regression. We chose vague priors for the regression coefficients, that is, normal priors centered around the maximum likelihood estimators and with a standard deviation 5 times bigger than the one estimated with the standard logistic regression. A $\mathcal{G}a(1,1)$ prior is selected for the parameter p. We adopted the BC_{el} algorithm with the same prior settings on the coefficients, and with $M = 20,000$, to estimate both the generalized logistic and the standard logistic regression.

The results are summarized in Table 3, which shows, for each European country subset, the DIC and LPML values, the parameter estimates given by the posterior means and the associated 95%-credible intervals. The analysis of the posterior means obtained with the two different models reveals some differences in the estimation of covariate effects on the immigration attitude. In general, there is much more posterior support for zero in the parameters estimated with the standard logistic regression, indicating that this approach is not flexible enough to model the immigration data. The generalized logistic model suggests that the higher the level of education the higher the probability that people are favourable towards immigration, especially in UK and France. This is confirmed by the standard logistic regression, but only for the French subset. In addition, the generalized logistic model indicates that people who trust the European parliament, particularly those living in Germany and France, tend to be in favour of immigration. Another important effect, according to the generalized logistic model, is `pplfair`, since trustful people generally show a more positive attitude towards immigration, as shown by the results of the German subset. Therefore, the determinants of public opinions towards immigration change in different European countries. The generalized logistic approach allows us to obtain insights about the determinants of immigration's attitude, that would not be possible using the standard logistic regression.

The proposed generalized logistic model outperforms the standard logistic regression for all countries, always showing the lowest DIC and highest LPML. This result is confirmed by the estimated values of the tail parameters p, that are lower than 1 for all countries, denoting heavy tails and suggesting that the standard logistic regression is not a suitable model for the immigration data. However, we note that the 95% credible intervals for the p parameters include the value one, due to the effect of

the prior variance. The results demonstrate the flexibility of the generalized logistic approach, that, thanks to the additional tail parameter, is able to capture non-standard tail behaviours.

7. Conclusions

This paper introduces a novel generalized logistic regression model to correctly estimate the opinions on immigration of citizens belonging to European countries. The model can accommodate heavy and light tails in the distribution of the predictors, while including the standard logistic approach as a special case. Despite the complex form of the likelihood, due to the additional parameter, we obtain fast and accurate results adopting a Bayesian empirical likelihood strategy. Simulation results show that the proposed approach outperforms the standard logistic model under various tail scenarios. The EU immigration data show non-standard tail behaviour, demonstrating the need for the proposed model. The estimates of the covariate effects obtained with the generalized logistic approach reveal insights about the determinants of public opinions towards immigration that would not be available adopting the standard logistic model. The results show that the determinants of people's views on immigrants vary between European countries. The proposed approach allows an accurate estimation of people's attitude towards migration, that is fundamental for developing suitable social and political immigration strategies.

Acknowledgments

The second author was supported by the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no: 630677. The third author acknowledges financial support from the European Union Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 796902. The fourth author is supported by the Chinese Fundamental Research Funds for the Central Universities No 20720181062.

References

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

[2] Bakker, B. N., M. Rooduijn, and G. Schumacher (2016). The psychological roots of populist voting: Evidence from the united states, the netherlands and germany. *European Journal of Political Research 55*(2), 302–320.

[3] BBC (2016a). Germany attacks: What is going on? `https://www.bbc.com/news/world-europe-36882445`. Accessed: 2018-09-03.

[4] BBC (2016b). Nice attack: What we know about the bastille day killings. `https://www.bbc.com/news/world-europe-36801671`. Accessed: 2018-09-03.

[5] Blinder, S. (2015). Imagined immigration: the impact of different meanings of ???immigrants??? in public opinion and policy debates in britain. *Political Studies 63*(1), 80–100.

[6] Castro, L., A. Felix, and R. Ramírez (2015). The limits of latinidad? immigration attitudes across latino national origin groups. *Minority Voting in the United States [2 volumes]*, 233.

[7] Chen, B., A. Shrivastava, R. C. Steorts, et al. (2018). Unique entity estimation with application to the syrian conflict. *The Annals of Applied Statistics 12*(2), 1039–1067.

[8] Chen, M.-H., D. K. Dey, and Q.-M. Shao (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association 94*(448), 1172–1186.
[9] Choi, H. M., J. P. Hobert, et al. (2013). The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics 7*, 2054–2064.
[10] Czado, C. and T. J. Santner (1992). The effect of link misspecification on binary regression inference. *Journal of statistical planning and inference 33*(2), 213–231.
[11] Durante, D. and D. B. Dunson (2014). Nonparametric bayes dynamic modelling of relational data. *Biometrika 101*(4), 883–898.
[12] Durante, D., D. B. Dunson, et al. (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis 13*(1), 29–58.
[13] ESS8 (2016). ESS Round 8: European Social Survey Round 8 Data. *Data file edition 2.0. NSD - Norwegian Centre for Research Data, Norway ??? Data Archive and distributor of ESS data for ESS ERIC*.
[14] ESS8 (2018). ESS Round 8: European Social Survey. *ESS-8 2016 Documentation Report. Edition 2.0. Bergen, European Social Survey Data Archive, NSD - Norwegian Centre for Research Data for ESS ERIC*.
[15] Eurostat (2018). Asylum in the EU member states. `http://ec.europa.eu/eurostat/web/main/home`. Accessed: 2018-08-06.
[16] Express (2017). Migrants make more than 30,000 illegal attempts to get into UK from Calais this year. `https://www.express.co.uk/news/uk/838727/migrant-uk-border-calais-france-emmanuel-macron-channel-tunnel-dover`. Accessed: 2018-09-03.
[17] Frühwirth-Schnatter, S. and R. Frühwirth (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis 51*(7), 3509–3528.
[18] Frühwirth-Schnatter, S. and R. Frühwirth (2010). Data augmentation and MCMC for binary and multinomial logit models. In *Statistical modelling and regression structures*, pp. 111–132. Springer.
[19] Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 501–514.
[20] Goodman, S. W. and M. Wright (2015). Does mandatory integration matter? effects of civic requirements on immigrant socio-economic and political outcomes. *Journal of Ethnic and Migration Studies 41*(12), 1885–1908.
[21] Gramacy, R. B., N. G. Polson, et al. (2012). Simulation-based regularized logistic regression. *Bayesian Analysis 7*(3), 567–590.
[22] Grazian, C., B. Liseo, et al. (2017). Approximate bayesian inference in semiparametric copula models. *Bayesian Analysis 12*(4), 991–1016.
[23] Heath, A. and R. Ford (2016). How do europeans differ in their attitudes to immigration? In *3rd international ESS conference: Understanding key challenges for European societies in the 21st century.*
[24] Holmes, C. C., L. Held, et al. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis 1*(1), 145–168.
[25] International Organization for Migration (2017). Migration and migrants: A global overview. In *IOM (2017) World Migration Report 2018*, Chapter 2. Geneva: IOM.
[26] Jackman, S. (2009). *Bayesian analysis for the social sciences*, Volume 846. John Wiley & Sons.
[27] Jiang, X., D. K. Dey, R. Prunier, A. M. Wilson, and K. E. Holsinger (2013). A new class of flexible link functions with application to species co-occurrence in cape floristic region. *The Annals of Applied Statistics*, 2180–2204.
[28] Jones, M. C. (2004). Families of distributions arising from distribution of order statistics. *Test 13*(1), 1–43.
[29] Kim, S., M.-H. Chen, and D. K. Dey (2007). Flexible generalized t-link models for binary response data. *Biometrika 95*(1), 93–106.
[30] Mengersen, K. L., P. Pudlo, and C. P. Robert (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences 110*(4), 1321–1326.

[31] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*(2), 237–249.

[32] Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association 108*(504), 1339–1349.

[33] Rossi, P. E., G. M. Allenby, and R. McCulloch (2012). *Bayesian statistics and marketing*. John Wiley & Sons.

[34] Roy, V. (2014). Efficient estimation of the link function parameter in a robust bayesian binary regression model. *Computational Statistics & Data Analysis 73*, 87–102.

[35] Storm, I., M. Sobolewska, and R. Ford (2017). Is ethnic prejudice declining in britain? change in social distance attitudes among ethnic majority and minority britons. *The British journal of sociology 68*(3), 410–434.

[36] Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association 83*(402), 426–431.

[37] United Kingdom Border Agency (2011a). *Migration Permanent Limit (Tier 1 andTier 2 of the Points-Based System): Impact Assessment*. London: Home Office.

[38] United Kingdom Border Agency (2011b). *Reform of the Points-Based Student Immigration System: Impact Assessment*. London: Home Office.

[39] United Nations High Commissioner for Refugees (2018). Syria emergency. `http://www.unhcr.org/syria-emergency.html`. Accessed: 2018-08-06.

[40] Wang, X., D. K. Dey, et al. (2010). Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. *The Annals of Applied Statistics 4*(4), 2000–2023.

[41] Wright, M., M. Levy, and J. Citrin (2016). Public attitudes toward immigration policy across the legal/illegal divide: The role of categorical and attribute-based decision-making. *Political Behavior 38*(1), 229–253.

[42] Zhu, W., J. M. Marin, and F. Leisen (2016). A bootstrap likelihood approach to bayesian computation. *Australian & New Zealand Journal of Statistics 58*(2), 227–244.

Country	DIC	LPML	Model	const	pplfair	trstep	trstun	happy	agea	edulvlb	hinctnta	p
GB	**912.681**	**-461.3604**	generalized logistic	-0.4202	-0.1822	-0.1731	-0.03721	-0.01326	0.00944	-0.0021	-0.0440	0.8218
			2.50%	*-2.4377*	*-0.3176*	*-0.3796*	*-0.2354*	*-0.2892*	*-0.0139*	*-0.0039*	*-0.1856*	*0.4284*
			97.50%	*1.4487*	*0.0113*	*0.0447*	*0.1409*	*0.1209*	*0.0358*	*-0.0006*	*0.1228*	*1.5696*
	1085.599	-571.9422	standard logistic	-1.2053	-0.1168	-0.0793	0.0041	0.0140	0.0103	-0.0024	-0.0265	
			2.50%	*-4.3530*	*-0.5254*	*-0.3468*	*-0.3521*	*-0.3352*	*-0.0215*	*-0.0062*	*-0.3358*	
			97.50%	*1.8652*	*0.2427*	*0.1941*	*0.3533*	*0.4066*	*0.0478*	*0.0003*	*0.2765*	
DE	**1217.856**	**-621.5027**	generalized logistic	-1.1452	-0.1612	-0.1524	-0.0112	-0.0873	0.0124	-0.0007	-0.1171	0.7338
			2.50%	*-2.4617*	*-0.3291*	*-0.2839*	*-0.2115*	*-0.2438*	*-0.0061*	*-0.0026*	*-0.2352*	*0.2955*
			97.50%	*0.6941*	*-0.0093*	*-0.0107*	*0.1847*	*0.0659*	*0.0375*	*0.0016*	*0.0414*	*1.7036*
	1439.653	-750.7174	standard logistic	-0.6765	-0.1710	-0.1041	-0.0033	-0.0656	0.0110	-0.0006	-0.0908	
			2.50%	*-5.0062*	*-0.5073*	*-0.4353*	*-0.3359*	*-0.4471*	*-0.0233*	*-0.0039*	*-0.3574*	
			97.50%	*2.5959*	*0.2658*	*0.1662*	*0.3785*	*0.2669*	*0.0493*	*0.0027*	*0.2153*	
FR	**1313.0390**	**-656.4381**	generalized logistic	0.8442	-0.0921	-0.1912	-0.0514	0.0191	-0.0043	-0.0038	-0.1061	0.7845
			2.50%	*-0.3443*	*-0.1898*	*-0.3592*	*-0.2047*	*-0.0609*	*-0.0155*	*-0.0056*	*-0.2390*	*0.3991*
			97.50%	*1.4314*	*0.0193*	*-0.0114*	*0.1417*	*0.1545*	*0.0091*	*-0.0027*	*0.0286*	*1.2809*
	1515.8030	-770.1424	standard logistic	-0.3798	-0.0756	-0.0969	-0.0029	0.0081	0.0050	-0.0029	-0.0546	
			2.50%	*-2.7743*	*-0.2725*	*-0.3029*	*-0.2662*	*-0.2503*	*-0.0228*	*-0.0053*	*-0.4103*	
			97.50%	*4.1248*	*0.2373*	*0.1920*	*0.3349*	*0.3153*	*0.0286*	*-0.0002*	*0.2266*	

Table 3.: Generalized logistic and standard logistic models comparison for the EU immigration data.