University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2018 *Paper* 123

Default Priors for the Intercept Parameter in Logistic Regressions

Philip S. Boonstra* Ryan P. Barbaro†
Ananda Sen‡

*The University Of Michigan, philb@umich.edu
†The University of Michigan
‡The University of Michigan

Default Priors for the Intercept Parameter in Logistic Regressions

Philip S. Boonstra, Ryan P. Barbaro, and Ananda Sen

Abstract

In logistic regression, separation refers to the situation in which a linear combination of predictors perfectly discriminates the binary outcome. Because finite-valued maximum likelihood parameter estimates do not exist under separation, Bayesian regressions with informative shrinkage of the regression coefficients offer a suitable alternative. Little focus has been given on whether and how to shrink the intercept parameter. Based upon classical studies of separation, we argue that efficiency in estimating regression coefficients may vary with the intercept prior. We adapt alternative prior distributions for the intercept that downweight implausibly extreme regions of the parameter space rendering less sensitivity to separation. Through simulation and the analysis of exemplar datasets, we quantify differences across priors stratified by established statistics measuring the degree of separation. Relative to diffuse priors, our recommendations generally result in more efficient estimation of the regression coefficients themselves when the data are nearly separated. They are equally efficient in non-separated datasets, making them suitable for default use. Modest differences were observed with respect to out-of-sample discrimination. Our work also highlights the interplay between priors for the intercept and the regression coefficients: numerical results are more sensitive to the choice of intercept prior when using a weakly informative prior on the regression coefficients than an informative shrinkage prior.

Default Priors for the Intercept Parameter in Logistic Regressions

Philip S. Boonstra[1*] Ryan P. Barbaro[2,3] Ananda Sen[1,4]

[1]Department of Biostatistics
[2]Division of Pediatric Critical Care
[3]Child Health Evaluation and Research Unit
[4]Department of Family Medicine
University of Michigan, Ann Arbor, MI

Abstract

In logistic regression, separation refers to the situation in which a linear combination of predictors perfectly discriminates the binary outcome. Because finite-valued maximum likelihood parameter estimates do not exist under separation, Bayesian regressions with informative shrinkage of the regression coefficients offer a suitable alternative. Little focus has been given on whether and how to shrink the intercept parameter. Based upon classical studies of separation, we argue that efficiency in estimating regression coefficients may vary with the intercept prior. We adapt alternative prior distributions for the intercept that downweight implausibly extreme regions of the parameter space rendering less sensitivity to separation. Through simulation and the analysis of exemplar datasets, we quantify differences across priors stratified by established statistics measuring the degree of separation. Relative to diffuse priors, our recommendations generally result in more efficient estimation of the regression coefficients themselves when the data are nearly separated. They are equally efficient in non-separated datasets, making them suitable for default use. Modest differences were observed with respect to out-of-sample discrimination. Our work also highlights the interplay between priors for the intercept and the regression coefficients: numerical results are more sensitive to the choice of intercept prior when using a weakly informative prior on the regression coefficients than an informative shrinkage prior.
Keywords: Bayesian Methods; Exponential-Power Distribution; Pivotal Separation; Quasi-Complete Separation; Rare Events

*Correspondence to: Philip S. Boonstra, PhD, SPH II, 1415 Washington Hts, Ann Arbor, MI, 48109; philb@umich.edu

1 Introduction

A default prior in principle falls between nearly flat/improper priors, e.g. Jeffrey's prior (Jeffreys, 1946), and informative shrinkage/variable selection priors, e.g. the Bayesian Lasso (Park and Casella, 2008). Such a prior mildly shrinks toward some null value non- or weakly identified parameters and leaves alone those well supported by the likelihood (Gelman et al., 2008; Greenland and Mansournia, 2015; Rainey, 2016). However, it does not borrow strength via shared hyperpriors as do informative shrinkage priors.

Default priors have been developed for binary data regression models, e.g. logistic regression, because of the possibility of so-called 'separation', or the existence of a linear combination of predictors that can perfectly discriminate the outcomes in the data (Albert and Anderson, 1984; Santner and Duffy, 1986). Separation can be 'complete' or 'quasi-complete', with both leading to non-finite maximum likelihood estimates (MLEs, Albert and Anderson, 1984; Santner and Duffy, 1986). A truly large association between a predictor and the outcome can cause separation, illustrating that this phenomenon is not always undesirable. Other causes include sparsity, high correlation between the predictors, or the inclusion of many binary predictors (Heinze and Schemper, 2002; Greenland and Mansournia, 2015). Regardless of cause, the lack of identification may warrant mild regularization via priors. For the intercept parameter, a very diffuse or improper prior is the usual choice; in this paper, we contend that its unique role and meaning warrant a more informative choice.

A number of authors have proposed default prior specifications for regression coefficients (Clogg et al., 1991; Bedrick et al., 1996; Zorn, 2005; Gelman et al., 2008; Hanson et al., 2014; Greenland and Mansournia, 2015). The scale-family of g-priors, or reference informative priors, is one early example of a default prior for regression coefficients (Zellner, 1983). The degree of shrinkage – and therefore the extent to which that family of priors may be viewed as 'default' – depends on the choice of scale parameter g, which is shared by all regression coefficients. If fixed at some diffuse or prespecified value, this would satisfy our definition of a default prior (Hanson et al.,

2014). If instead g is adaptively tuned via a hyperprior, as in Marin and Robert (2007), this would not, as information is shared across parameters.

Christmann and Rousseeuw (2001) propose quantitative measures of separation: $n_{\text{comp}} \geq 0$ is the size of the smallest subset of observations that, if removed from the data, would completely separate the complementary subset. Separation is equivalent to $n_{\text{comp}} = 0$; the resulting lack of numerical convergence allows for relatively easy detection. In contrast, near-separation, i.e. a small but positive n_{comp}, yields finite MLEs but manifests symptoms of separation including instability and efficiency loss. For this reason, mild regularization from default priors can be just as useful when there are a few dozen predictors as when there are hundreds or more, particularly when the number of observations is of a similar order, i.e. $p \approx n$.

The point of departure from previous work is our specific focus on the intercept parameter. In contrast to regression coefficients, there is generally not an intuitive null direction toward which shrinkage should be, and borrowing strength for the intercept is not possible (e.g. Section 3.4, Hastie et al., 2009). The usual recommendation is that a prior should be flat or effectively so (Greenland and Mansournia, 2015; Zorn, 2005; Gelman et al., 2008). We demonstrate that straightforward efficiency gains are possible by assuming that exceptionally large values of the intercept are either implausible or unverifiable in the data, such that down-weighting these regions nearly always results in efficiency gains.

This paper makes several contributions. First, we use complete separation to establish a rationale that mild shrinkage of the intercept can improve estimation of the regression coefficients themselves. Following Ghosh et al. (2017), we consider a stronger type of separation that we call 'pivotal separation'. Second, we propose to adapt the exponential-power scale-family of distributions (Box, 1953; West, 1987; Box and Tiao, 1992) for default use as a prior on the intercept in binary data regression models and develop an algorithm to determine a suitable scale for this prior. Finally, our work highlights the interplay between choice of prior on the intercept and that of the regression coefficients; our numerical studies directly compare the performance of a hierarchical

shrinkage prior on the regression coefficients against a weakly informative prior.

In Section 2, we motivate the interplay between intercept and regression coefficients under the premise of separation. Sections 3 and 4 describe our choices of the different class priors for the intercept and the regression coefficients, respectively. Findings from a comprehensive simulation study are documented in Section 5, which provides a comparative appraisal of the Bayesian estimators under varied scenarios including sparsity and $p \approx n$. The demonstration of our methodology on ten datasets in Section 6 illustrates the heterogeneity in the degree of separation in real data and, more importantly, highlights the stabilizing properties of our proposed priors in estimation of the regression coefficients. Section 7 interprets our results and discusses some counterarguments against using priors on the intercept parameter.

2 Motivation from Separation

We have n data points denoted by $\{Y_i, \boldsymbol{X}_i\}_1^n$, where $Y_i \in \{0, 1\}$, and $\boldsymbol{X}_i$ is a p-dimensional vector of covariates. A generalized linear model (GLM) takes the form $g(\Pr(Y = 1|\boldsymbol{X})) = \alpha + \boldsymbol{X}^\top\boldsymbol{\beta}$, where g is a link function, e.g. logistic, probit, or complementary log-log, mapping the unit interval to the real line. The likelihood is

$$L(\alpha, \boldsymbol{\beta}) = \prod_i g^{-1}(\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_i)^{Y_i} \left[1 - g^{-1}(\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_i)\right]^{1-Y_i}. \tag{1}$$

Partition the outcomes into sets $\mathcal{A} = \{i : Y_i = 1\}$ and $\mathcal{A}^C = \{i : Y_i = 0\}$. Complete separation holds when there exists $\mathcal{D} \in \mathcal{R}^{p+1}$ such that, for any $\{\alpha, \boldsymbol{\beta}\} \in \mathcal{D}$,

$$\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_i \geq 0, \quad \forall i \in \mathcal{A} \tag{2}$$

$$\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_i \leq 0, \quad \forall i \in \mathcal{A}^C \tag{3}$$

with the further condition that at least one of the inequalities in Equations (2) and (3) be strict. The direction of the inequalities is assumed without loss of generality. Unless noted otherwise, we use 'separation' as shorthand for 'complete separation' which differs

from the convention of Gelman et al. (2008), who use it to mean 'quasi-complete separation'. This is a weaker condition that exists when Equations (2) and (3) hold for at least one $\{\alpha, \boldsymbol{\beta}\} \neq \{0, \mathbf{0}\}$, without requiring a strict inequality. Analogous to n_{comp}, the overlap statistic n_{over} measures the minimum subset needed to remove in order to induce quasi-complete separation in the complementary subset, with $0 \leq n_{\text{over}} \leq n_{\text{comp}}$ (Christmann and Rousseeuw, 2001).

Albert and Anderson (1984) demonstrate that, in a logistic regression model, finite-valued MLEs do not exist for all parameters when the data are separated (either completely or quasi-completely). This result extends to any link such that g^{-1} is increasing in its argument. Moreover, near-separation, i.e. separation in a large subset of the data, may result in weakly identified parameter estimates. Nonetheless, in some contexts, e.g. diagnostic test evaluation, separation *is* the objective. Thus, separation – and its impact on efficiency – is neither an artificial nor pathological concern. We re-characterize results from Albert and Anderson to motivate default priors that ameliorate its symptoms.

When separation holds, then, for a given $\{\alpha, \boldsymbol{\beta}\} \in \mathcal{D}$, there exists a vector $\boldsymbol{m} = \{m_1, \ldots, m_p\}$ and a scalar $\tilde{\beta}$ such that $\tilde{\beta}\boldsymbol{m} = \boldsymbol{\beta}$ and

$$\alpha + \tilde{\beta} \times (\boldsymbol{m}^\top \boldsymbol{X}_i) \geq 0, \quad \forall i \in \mathcal{A} \tag{4}$$

$$\alpha + \tilde{\beta} \times (\boldsymbol{m}^\top \boldsymbol{X}_i) \leq 0, \quad \forall i \in \mathcal{A}^C, \tag{5}$$

Although multiple $\{\alpha, \boldsymbol{\beta}\}$ pairs may lie in $\mathcal{D}$ for a given dataset, the existence of one such pair is sufficient to separate outcomes, thus we focus on a single realization of $\{\alpha, \boldsymbol{\beta}\} \in \mathcal{D}$. Let $\mathcal{D}(\boldsymbol{m})$ be the two-dimensional region of separation so that $\{\alpha, \tilde{\beta}\} \in \mathcal{D}(\boldsymbol{m})$ separates the data as above. Without loss of generality, we may assume that $\max_j\{|m_j|\} = 1$ and $\tilde{\beta} > 0$. $\tilde{\beta}$ corresponds to the magnitude of the largest element of $\boldsymbol{\beta}$, while $\boldsymbol{m}$ is understood to be the relative size with respect to $\tilde{\beta}$ of all remaining elements of $\boldsymbol{\beta}$. If $p = 1$, then $\boldsymbol{m} = 1$ and $\mathcal{D} \equiv \mathcal{D}(\boldsymbol{m})$. When $p > 1$, this re-parameterization reduces the problem into two dimensions. The absence of finite MLEs is a consequence

of Theorem 1 in Albert and Anderson (1984), which we re-characterize as follows:

Result 1 Suppose a given dataset is separated by $\{\alpha, \tilde{\beta}\} \in \mathcal{D}(\boldsymbol{m})$, then

a. for any $k > 0$, $\{k\alpha, k\tilde{\beta}\} \in \mathcal{D}(\boldsymbol{m})$, and

b. for $0 < k_1 < k_2$, $L(k_1\alpha, k_1\tilde{\beta}\boldsymbol{m}) < L(k_2\alpha, k_2\tilde{\beta}\boldsymbol{m})$, where L is given in Equation (1).

In words, starting from an initial set of parameter values that separates the data, we can calculate a new, distinct set that also separates the data, increases the magnitude of the intercept and regression coefficients (those that are non-zero), and increases the likelihood relative to the initial set. This establishes the lack of a finite-valued maximum in the interior of the parameter space. The specific value of $\boldsymbol{m}$ in **Result 1** is not critical; the existence of just one such $\boldsymbol{m}$ gives rise to non-finite MLEs.

2.1 Pivotal Separation

A result from Ghosh et al. (2017) further characterizes the intercept-regression coefficient relationship in separation. The authors establish conditions under which the posterior mean of one or more elements of $\boldsymbol{\beta}$ does not exist. Translated to our notation, their primary result is that, when β_j is a priori Cauchy, its posterior mean does not exist if and only if there exists a particular $\boldsymbol{m} = \boldsymbol{m}^*$ satisfying Equations (4) and (5) such that (i) $m^*_{j'} = 0$ for $j' \neq j$ and (ii) $\{0, \tilde{\beta}\} \in \mathcal{D}(\boldsymbol{m})$, for some $\tilde{\beta} \neq 0$. In this case, Ghosh, et al. call the jth predictor a 'solitary separator'. Incidentally, ignoring the first condition and focusing only on the second, i.e. $\{0, \tilde{\beta}\} \in \mathcal{D}(\boldsymbol{m})$, Ghosh, et al. incidently define a stronger condition than complete separation, namely that in the absence of an intercept. Formally, we say that 'pivotal separation' holds when there exists $\mathcal{D}^* \in \mathcal{R}^p$ such that, for any $\boldsymbol{\beta} \in \mathcal{D}^*$,

$$\boldsymbol{\beta}^\top \boldsymbol{X}_i \geq 0, \quad \forall i \in \mathcal{A} \tag{6}$$

$$\boldsymbol{\beta}^\top \boldsymbol{X}_i \leq 0, \quad \forall i \in \mathcal{A}^C \tag{7}$$

with the further condition that at least one of the inequalities in Equations (6) and (7) be strict. In other words, the set of separating planes contains the point $\alpha = 0$, and

the set pivots in the neighborhood of $\alpha = 0$. We define the statistic n_{piv} as the size of the smallest subset of observations that, if removed from the data, would pivotally separate the data. Because pivotal separation is the strongest type of separation, we can extend the inequality to $0 \leq n_{\text{over}} \leq n_{\text{comp}} \leq n_{\text{piv}}$.

These relationships between the magnitude of the intercept and the regression coefficients in the likelihood are our key motivation for proposing mild shrinkage of the intercept. In Bayesian analyses, one relies on the priors' contribution – that of α as well as $\boldsymbol{\beta}$ – to control variability. Thus, in cases of complete separation, a priori supporting extreme regions of α may, a posteriori, support correspondingly extreme values of $\boldsymbol{\beta}$. Put differently, shrinking α may indirectly constrain the posterior support for $\boldsymbol{\beta}$, and vice versa. Although this relationship does not strictly hold when near-separation holds, i.e. when n_{comp} is non-zero but small, numerical instability is present in these cases as well (which we will see in the proceeding sections), and it becomes difficult to estimate well both the intercept and the regression coefficients. Based upon this, we propose an alternative default prior for α intended to limit the support of α and therefore confer more precision in estimating $\boldsymbol{\beta}$.

3 Possible Choices of Default Prior

For the remainder of the paper, we focus on the logistic link function, $g(x) = \text{logit}(x) \equiv \log_e(x/[1-x])$. We center the covariate vector to the empirical mean in the data, so that $\boldsymbol{X} = \mathbf{0}$ represents the mean, and implicitly assume that all prior formulations have a location parameter equal to zero. Probabilistically, the elements of $\boldsymbol{\beta}$ (and therefore prior interpretations) can only be framed in relative terms. For example, a log odds-ratio of $\log_e(1.3)$ increases a baseline probability of 0.004 to about 0.005 but increases a baseline probability of 0.500 to 0.565. In contrast, we can interpret α as a 1-1 function of a probability. Because the data are centered, $\text{logit}^{-1}(\alpha) = \Pr(Y = 1|\boldsymbol{X} = \mathbf{0}) \in [0, 1]$ is the probability of the outcome occurring, the 'risk', in a mathematically average observation. It will be convenient to consider $\text{logit}^{-1}(\alpha)$, because the risk scale allows

for a more intuitive interpretation of the consequences of a particular prior formulation.

3.1 Priors based upon the t-distribution

Gelman et al. (2008) suggest equipping α with a very diffuse Cauchy distribution. Greenland and Mansournia (2015) recommend an improper/flat prior for multiple reasons, including lack of an appropriate null location parameter and a sensitivity to choice of coding of the predictors. The authors caution against an informative prior on α; we discuss this important point further in the Discussion. A diffuse Cauchy prior is practically flat, and these two recommendations are essentially in agreement.

Let $t_\nu(\sigma)$ denote a t-distribution centered at zero with ν degrees of freedom and scale parameter σ. To ensure the existence of the first two posterior moments, we consider a lighter-tailed prior than that recommend by Gelman, et al., namely $\alpha \sim t_3(10)$. To two significant digits, the $t_3(10)$ distribution implies $\Pr(0.001 < \text{logit}^{-1}(\alpha) < 0.999) = 0.46$; in other words, about 54% of the prior mass falls below 0.001 or above 0.999. This is the most diffuse prior we consider in this paper. We also consider $\alpha \sim t_\infty(10) \equiv N(\sigma = 10)$, that is, a normal distribution with mean zero and standard deviation 10. The prior mass of a normal prior is more concentrated towards the center: $\Pr(0.001 < \text{logit}^{-1}(\alpha) < 0.999) = 0.51$. The density functions of these scale-families are given in the first two rows of Table 1, and Figure 1 plots the densities (truncated to the interval $[-8, 8]$) for the particular choices of scales that we evaluated.

3.2 Alternative Prior Formulations

Generally, there are two decisions made (either implicitly or explicitly) in the formulation of a prior. First, a family of distributions must be selected. Then, a particular parametrization must be selected from within that family, which controls where the prior mass falls. For example, the scaled t-distribution implies a bell-shaped prior, with the choice of degrees of freedom and scale parameter determining the spread or concentration of the mass. We discuss these two choices in reverse order.

3.2.1 Default parameter values

Explicitly limiting the prior mass on extreme values of α is preferred both intuitively and theoretically. First, because the covariates are centered, α (properly transformed) represents the risk for an *average* patient. An a priori assumption that it is very small or very large may be counter to intuition or – at the least – unverifiable in all but very large datasets. Second, very large or small values of α may indirectly admit extreme values of $\boldsymbol{\beta}$, as argued in Section 2. In this section, we describe **Algorithm 1** which, when paired with a scaled-family of priors, identifies the value of the scale parameter such that the prior probability of α falling in an extreme region is small, given the sample size n. The algorithm requires a user-provided definition of *extreme* and *small*.

We motivate the algorithm by considering the $\text{logit}^{-1}(\alpha)$, or risk, scale. Consider the hypothetical most extreme data configurations: $\sum Y_i = 0$ or $\sum Y_i = n$. Although these configurations suggest that $\text{logit}^{-1}(\alpha)$ is likely to be very close to 0 or 1, they may provide little information on exactly how close it should be. Our algorithm approximates the boundaries of the interval where precision degrades. As the sample size n increases, we can precisely estimate more extreme values of $\text{logit}^{-1}(\alpha)$. Thus, the algorithm depends on n. Specifically, it ensures that a proportion, $1 - q$, of the prior mass on $\text{logit}^{-1}(\alpha)$ is in the interval $[1 - s_n, s_n]$, where $s_n \in (0.5, 1)$. Ignoring Jensen's bias and thus equating the average event rate with the event rate at the mean, i.e. $E_{\boldsymbol{X}}\Pr(Y = 1|\boldsymbol{X}) \approx \Pr(Y = 1|\boldsymbol{X} = E[\boldsymbol{X}]) = \text{logit}^{-1}(\alpha)$, then the likelihood of the most extreme configuration is $\Pr(\sum Y_i = n|\{\boldsymbol{X}_i\}) \approx \left(\text{logit}^{-1}(\alpha)\right)^n$. In this case, the maximized likelihood, equal to 1, is achieved at $\hat{\alpha}_{\text{MLE}} = \infty$, and so the likelihood ratio evaluated at s_n is $(s_n)^n / \sup_\alpha L(\alpha; \sum Y_i = n) = (s_n)^n$. The likelihood ratio test statistic, $-2n\log_e(s_n)$, approaches zero as s_n is made closer to 1, and a tail region of the $\text{logit}^{-1}(\alpha)$ space exists to the right of s_n within which values of $\text{logit}^{-1}(\alpha)$ become increasingly indistinguishable from $\text{logit}^{-1}(\hat{\alpha}_{\text{MLE}}) \equiv 1$ with respect to the likelihood ratio. We identify the value of s_n such that the likelihood ratio test statistic $-2n\log_e(s_n)$ equals δ, i.e. $s_n = \exp\{-\delta/(2n)\}$. By construction, then, when $\sum Y_i = n$, the likelihood ratio statistic comparing s_n to $\text{logit}^{-1}(\hat{\alpha}_{\text{MLE}})$ is δ. The case is symmetric when

$\sum Y_i = 0$. Next, we use this derived value of s_n to identify the scale parameter in a family of priors such that a proportion $1 - q$ of the prior mass falls in the interval $[1 - s_n, s_n]$; equivalently, the proportion of mass outside this interval is q. The steps are as follows:

Algorithm 1

1. Choose constants δ and q such that $0 < \delta < -2n\log_e(0.5)$ and $0 < q < 1$. The bounds on δ ensure that $0.5 < s_n < 1$.
2. Calculate $s_n = \exp\{-\delta/(2n)\}$, where n is the sample size.
3. Given a scale-family of priors on α parametrized by σ, select $\sigma = \sigma_n$ so that $\Pr(1 - s_n < \text{logit}^{-1}(\alpha) < s_n | \sigma = \sigma_n) = 1 - q$, equivalently $\Pr(\text{logit}(1 - s_n) < \alpha < \text{logit}(s_n) | \sigma = \sigma_n) = 1 - q$

The algorithm is semi-automatic in that it requires user-selected constants δ and q. Larger values of δ and smaller values of q will result in smaller values of σ_n. The values that we use here, $\delta = 1$ and $q = 0.01$ work well in our numerical study.

3.2.2 Choice of prior family

Although the student-t scale family of priors can be applied to **Algorithm 1**, it is not necessarily ideally suited for doing so: when n is small, the algorithm-based scale parameter σ_n will be correspondingly small, and the prior mass will be concentrated around $\alpha = 0$. This is potentially as unsatisfactory as having mass concentrated at the tails, because $\alpha = 0$ generally not an appropriate null value for the intercept. It is not only the scale of the prior on α that is important but also the shape. Modern developments in computational statistics, such as the STAN probabilistic modeling language (Stan Development Team, 2018, 2017; Carpenter, 2017), have opened new avenues for priors driven by statistical sensibility and not computational convenience, e.g. prior conjugacy. We use a family of distributions called the *exponential-power* distribution (Box, 1953; West, 1987; Box and Tiao, 1992) with parameters $\sigma, k > 0$. Denoted by $\text{EP}_k(\sigma)$, its density function is $\pi(\alpha|\sigma, k) \propto \exp\{-|\alpha/\sqrt{2}\sigma|^k\}$. The supplement derives further properties of the EP family. With $k = 1$ or 2, respectively,

Table 1: Densities of families of priors on the intercept parameter α, up to a multiplicative constant.

Name	$\pi(\alpha) \propto$
$t_\nu(\sigma)$	$(\sigma^2 + \alpha^2)^{-(\nu+1)/2}$
$t_\infty(\sigma)$	$\exp\{-\alpha^2/(2\sigma^2)\}$
$\text{EP}_k(\sigma)$	$\exp\{-\lvert\alpha/\sqrt{2}\sigma\rvert^k\}$
$\text{Logis}(\sigma)$	$\frac{e^{-\alpha/\sigma}}{(1+e^{-\alpha/\sigma})^2}$

this reduces to the Laplace and normal families of distributions. Using $k > 2$ results in platykurtic distributions which, used in conjunction with **Algorithm 1** herein, yield a more locally uniform distribution of mass within the extreme boundaries while smoothly vanishing to zero outside of s_n. This is illustrated in Figure 1, which plots EP distributions with $k = 2$, 4, and 10. Given the numerical results in Section 2, platykurtosis may be a desirable feature for a prior on α.

For completeness, we also consider a Logistic prior on α, with scale estimated using **Algorithm 1**. In total, we consider six specific priors for the intercept parameter in a multivariable logistic regression: $t_3(10)$, $t_\infty(10)$, $\text{EP}_2(\sigma_n)$, $\text{EP}_4(\sigma_n)$, $\text{EP}_{10}(\sigma_n)$, and $\text{Logis}(\sigma_n)$, where σ_n denotes estimation of σ using **Algorithm 1**, with $\delta = 1$ and $q = 0.01$. When $n = 250$, these six distributions are plotted in Figure 1. Table S1 (supplement) gives calculations of σ_n from **Algorithm 1** as n changes.

4 Prior on $\boldsymbol{\beta}$

Our objective is not to directly compare priors on $\boldsymbol{\beta}$ but rather to isolate the impact of the choice of prior on α, given a typical choice of prior on $\boldsymbol{\beta}$. Unavoidable, however, is that α and $\boldsymbol{\beta}$ will be a posteriori correlated, and thus the choice of prior on $\boldsymbol{\beta}$ matters. As discussed in the Introduction, default priors on $\boldsymbol{\beta}$ are meant to provide automatic weak shrinkage of unstable components of $\boldsymbol{\beta}$, whereas informative shrinkage priors proactively distinguish between signal and noise in the components of $\boldsymbol{\beta}$ by borrowing strength across parameters. We consider one of each type.

Hierarchical Shrinkage (HS)(Carvalho et al., 2009, 2010; Piironen and Vehtari,

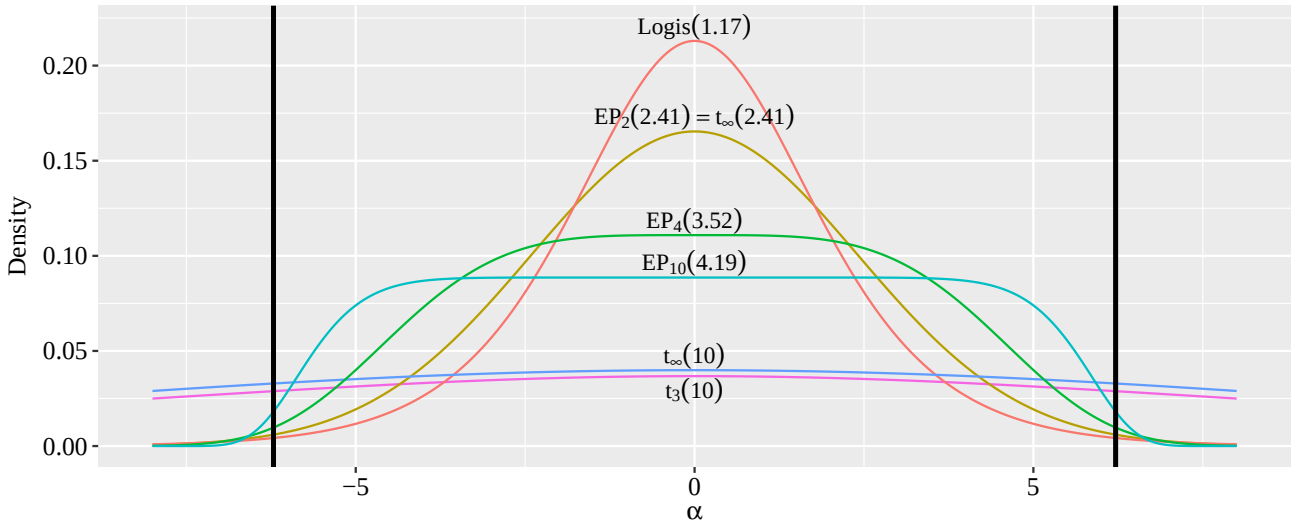

Figure 1: Densities of six prior distributions on α, arranged by the height of the density at $\alpha = 0$. The scales of the logistic ('Logis') and exponential-power ('EP') priors are selected according to **Algorithm 1** with $n = 250$, using $\delta = 1$ and $q = 0.01$, which ensures that only 1% of the prior mass falls outside of the interval $[-6.21, 6.21]$, denoted by the vertical lines.

2015, 2017). Independently for each of the j components of $\boldsymbol{\beta}$,

$$\tau \sim t_1^+(1),$$

$$\lambda_j \sim t_1^+(1),$$

$$\theta_{jn} \equiv \left(1/c^2 + 1/\left[\tau^2\lambda_j^2\psi_n^2\right]\right)^{-1/2}, \tag{8}$$

$$\beta_j|\{\lambda_j, \tau\} \overset{ind}{\sim} N(\theta_{jn}) \tag{9}$$

When $c = \infty$, each β_j is conditionally a mean-zero, Normal-scale-mixture of a local shrinkage parameter λ_j, itself distributed as a standard half-Cauchy, a global shrinkage parameter τ, which is also half-Cauchy, and a fixed, user-defined scale ψ_n, possibly dependent on n. Choosing instead a value for c that is large but finite has the effect of soft-thresholding weakly or under-identified coefficients that are *under*-shrunk by the heavy-tailed Cauchy prior, e.g. when the data are separated (Piironen and Vehtari, 2017); the prior variance for any coefficient will never exceed c^2. We fix $c = 15$ here.

Piironen and Vehtari (2016) demonstrate that ψ_n should scale with $n^{-1/2}$. In logistic regressions, they derive a formula to approximate the effective number of non-zero

parameters implied by any choice of prior that can be formulated as a Normal-scale-mixture, as in Equation (9). Specifically, the parameter $\kappa_j \equiv (1+[\sqrt{n}\theta_{jn}/2]^2)^{-1} \in (0,1)$ quantifies the amount of shrinkage towards zero relative to the MLE of β_j, with larger values corresponding to greater shrinkage. The effective total number of non-zero parameters assumed by the prior, p_{eff}, is thus approximated by $\sum_j (1-\kappa_j)$. Based upon this, for a desired value of $\tilde{p}_{\text{eff}}$, we use the value of ψ_n solving $\tilde{p}_{\text{eff}} = \sum_j (1-E[\kappa_j])$, where the expectation is taken over the prior distributions of τ and each λ_j, $j = 1, \ldots, p$. The choice of $\tilde{p}_{\text{eff}}$ depends on the context of the study; usually, it will be much smaller than p, particularly when p is large, to reflect that most associations are expected to be close to zero. We use HS($\tilde{p}_{\text{eff}}$) to denote the hierarchical shrinkage prior in Equation (9) with ψ_n based upon the choice of $\tilde{p}_{\text{eff}}$; our choices for $\tilde{p}_{\text{eff}}$ are given in Section 5.1.

Logistic (Logis) Greenland and Mansournia (2015) compare several weakly informative default priors on $\boldsymbol{\beta}$, including a standard Logistic distribution. That is, each of the j components of $\boldsymbol{\beta}$ is identically and independently distributed as a standard logistic random variable, $\beta_j \overset{iid}{\sim} \text{Logis}(\sigma = 1)$. The authors recommend *not* to scale the covariates when using this prior, advice that we adhere to here.

5 Simulation Study

We conduct a simulation study to compare the performance of six priors on α in terms of their impact on estimation (of the regression coefficients $\boldsymbol{\beta}$) and discrimination of observations. We designed the scenarios to represent the challenging regressions that may lead to separation, namely rare events (small values of α), large associations ($\boldsymbol{\beta}$ large in magnitude), correlated $\boldsymbol{X}$, binary $\boldsymbol{X}$, and/or $p \approx n$.

5.1 Generating Models

We considered a total of nine scenarios of varying p. These are summarized in Table 2 and detailed in the Supplement (S2). In each, the sample size n is varied such that p/n ranges between a small to a large fraction. The covariate vector $\boldsymbol{X}$ is either multivariate

normal or Bernoulli; in the case of the latter, we generated them jointly as indicators of orthant sets of an underlying multivariate normal. This is akin to a framework containing correlated binary exposure variables that represent threshold crossing of a continuous marker. For simulating the datasets, the parameters were fixed at the given values of $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ and $\alpha = \alpha^*$; the covariates $\boldsymbol{X}$ and outcome Y are random. Because $\boldsymbol{X}$ is centered to its empiric mean prior to model fitting, the third column of Table 2, $\alpha^* + E(\boldsymbol{X})^\top \boldsymbol{\beta}^*$ is the value to which the intercept priors correspond.

A single simulation consisted of sampling the random vector $\{Y, \boldsymbol{X}\}$ for n observations and fitting the data to a Bayesian logistic regression using either an HS or Logistic prior on $\boldsymbol{\beta}$ and one of the considered priors on α. For the HS prior on $\boldsymbol{\beta}$, the covariates were standardized to have mean zero and unit standard deviation. We selected the scale parameter ψ_n as described in Section 4, setting the approximate expected number of non-zero parameters to be $\tilde{p}_{\text{eff}} = 2$ for Scenario 1 (for which $p = 4$) and $\tilde{p}_{\text{eff}} = 10$ for Scenarios 2–9 ($p = 25$, 75, or 150). For the Logistic prior on $\boldsymbol{\beta}$, the covariates were only centered, as in Greenland and Mansournia (2015) .

We quantified performance using root mean-squared error (RMSE) (Armagan et al., 2013) and area under the receiver-operator characteristic (AUC). Let Σ_X denote the true covariance matrix of $\boldsymbol{X}$, and let F_{post} generically denote denote the posterior distribution. The metrics are respectively defined as

$$\text{RMSE} = \sqrt{(E_{F_{\text{post}}}[\boldsymbol{\beta}] - \boldsymbol{\beta}^*)^\top \Sigma_X (E_{F_{\text{post}}}[\boldsymbol{\beta}] - \boldsymbol{\beta}^*)} \tag{10}$$

$$\text{AUC} = \frac{\sum \sum_{i,j:Y_{i,\text{new}}=0,Y_{j,\text{new}}=1}^{n_{\text{new}}} 1\left[(X_{i,\text{new}} - X_{j,\text{new}})^\top E_{F_{\text{post}}}[\boldsymbol{\beta}] < 0\right]}{\sum \sum_{i,j:Y_{i,\text{new}}=0,Y_{j,\text{new}}=1}^{n_{\text{new}}}} \tag{11}$$

RMSE measures the distance between the posterior mean of $\boldsymbol{\beta}$ and its true value, standardized to the variance of $\boldsymbol{X}$. AUC measures the ability to discriminate between observations that will and will not experience the outcome. Neither RMSE nor AUC depend directly upon the posterior distribution of α. For each of 30 scenario-plus-sample-size combinations, we simulate 200 datasets. To stratify performance based upon the degree of separation, we calculated n_{comp} and n_{piv}, the minimum number

Table 2: Summary of nine scenarios considered in the simulation study. The parameters α^* and $\boldsymbol{\beta}^*$ are the true values of the intercept and regression coefficients, fixed across simulation iterations, and p is the length of the parameter $\boldsymbol{\beta}^*$. The data are centered prior to analysis, meaning that the third column, $\alpha^* + E(\boldsymbol{X})^\top \boldsymbol{\beta}^*$ is the value of the intercept parameter to which the priors correspond. The fourth column, $\Pr(Y = 1|\boldsymbol{X} = E\boldsymbol{X})$, is the risk at the average covariate pattern, equal to $\text{logit}^{-1}(\alpha^* + E(\boldsymbol{X})^\top \boldsymbol{\beta}^*)$. The fifth column is the average risk in the population, equivalently $E_X \text{logit}^{-1}(\alpha^* + \boldsymbol{X}^\top \boldsymbol{\beta}^*)$. The sixth column is the standard deviation of the joint linear predictor. The final column qualitatively describes possible causes of separation for that scenario

Scenario	p	$\alpha^* + E(\boldsymbol{X})^\top \boldsymbol{\beta}^*$	$\Pr(Y = 1\|\boldsymbol{X} = E\boldsymbol{X})$	$E_{\boldsymbol{X}} \Pr(Y = 1\|\boldsymbol{X})$	$\text{sd}(\boldsymbol{X}^\top \boldsymbol{\beta}^*)$	Causes of Separation
1	4	-1.5	0.182	0.202	0.7	{BX}
2	25	-1.6	0.165	0.186	0.7	{BX, $p \approx n$}
3	25	-1.6	0.165	0.167	0.2	{BX, $p \approx n$}
4	25	-5.0	0.007	0.075	2.7	{RE, LA, BRX}
5	25	-2.5	0.076	0.181	2.9	{LA, BRX, $p \approx n$}
6	75	-4.0	0.018	0.071	2.0	{RE, $p \approx n$}
7	75	-3.0	0.047	0.067	0.9	{BX, $p \approx n$}
8	150	-3.0	0.047	0.060	0.7	$\{p \approx n\}$
9	150	-3.0	0.047	0.050	0.3	$\{p \approx n\}$

'BX' = binary $\boldsymbol{X}$;
'BRX' = binary, rare (mean = 0.05) $\boldsymbol{X}$;
'$p \approx n$' = at least one setting considered with $n \leq 2p$;
'RE' = rare events = $\Pr(Y = 1|\boldsymbol{X} = E\boldsymbol{X}) < 0.02$;
'LA' = large association = standard deviation of linear predictor ≥ 2

of observations needed to remove to induce complete and pivotal separation (Section 2.1), respectively using **Algorithm 2**, described in the Supplement.

5.2 Results

Tables 3 and S2 (in the supplement) respectively report the median RMSE and median AUC calculated over 200 simulated datasets from each scenario+sample size configuration. The best value (i.e. lowest median RMSE, highest median AUC) is italicized (calculated separately for each of the two priors on $\boldsymbol{\beta}$). Bold-faced values correspond to those priors that out-performed the italicized prior in at least 33% of the 200 simulated datasets, which means that the prior was not too suboptimal. Figure 2 plots the $\log_2$ ratio of RMSE comparing the $t_3(10)$ prior on α to each of the other priors for each of the 6000 individual simulated datasets from all scenarios combined (200 datasets per each of 30 scenario + sample size configurations) stratified by the value of the separation statistics n_{comp} and n_{piv}. Figure S2 in the Supplement gives the analogous

plot for AUC.

Focusing on the HS prior on $\boldsymbol{\beta}$ in Table 3, the choice of prior on α has relatively little impact on estimation of $\boldsymbol{\beta}$ in Scenario 1 ($p = 4$). The median values of both n_{comp} and n_{piv} in this scenario were generally much larger than zero, ranging from 9 ($n = 50$) to about 84 ($n = 400$). In subsequent Scenarios with larger p, the EP and Logistic priors (using σ_n based upon **Algorithm 1**), generally have better estimation error among the smaller sample sizes (n) considered, with all six priors yielding about equal RMSEs as the sample size n, and consequently the statistics n_{comp}, n_{piv}, increase. In some scenarios, e.g. Scenario 6 with $n = 100$, the $t_3(10)$ and $t_\infty(10)$ priors have the highest median values (155 and 153, respectively, compared to 102 from the EP_4 prior), but they are also boldfaced. Thus, these heavy-tailed priors also exhibit greater within-scenario variation. There were greater differences when considering the Logistic prior on $\boldsymbol{\beta}$, both between, i.e. comparing HS to Logistic priors on $\boldsymbol{\beta}$, and within, i.e. comparing the six priors on α. The HS prior on $\boldsymbol{\beta}$ outperforms the Logistic prior except for Scenarios 4 and 5, and the $\text{EP}_4(\sigma_n)$ prior is nearly always smallest when using a Logistic prior on $\boldsymbol{\beta}$. Overall, fewer priors are in boldface, pointing to greater variation between priors on α.

The final row of Table 3 gives the average rank of each prior over all datasets combined, with lower numbers corresponding to smaller, i.e. better, RMSE. For HS, the EP_4 prior had the best average rank of 3.15, followed by EP_2 and EP_4 (both 3.25), a Logistic prior (3.35), and finally the heavier tailed t_∞ and t_3 priors, respectively, 3.94 and 4.07. For the Logis prior on $\boldsymbol{\beta}$, the $\text{EP}_1 0$ prior had the highest average rank (2.63), followed by EP_4 (2.80), EP_2 (3.07), Logistic (3.44), t_∞ (4.35), and t_3 (4.72).

From Figure 2, the improvement in RMSE is most prominent when the data are pivotally separated ($n_{\text{piv}} = 0 \Leftrightarrow$ the "0!" category on the x-axis) or completely but not pivotally separated ($n_{\text{comp}} = 0$ & $n_{\text{piv}} > 0 \Leftrightarrow$ the "0/{0!}" category), although better performance is also noted for many datasets in which n_{comp} was small.

From Table S2 and Figure S2 in the Supplement, differences in terms of discrimination were less pronounced. Under the HS prior on $\boldsymbol{\beta}$, there was always a < 0.04

Table 3: Median RMSEs (×100), across 200 datasets, for estimating $\boldsymbol{\beta}$ with its posterior mean for all combinations of six priors on the intercept parameter α and two priors on $\boldsymbol{\beta}$. Smaller numbers are better. In each row, values in *italics* correspond to the smallest median RMSE, and values in **bold** had a smaller RMSE than the italicized value in at least 33% of the datasets. All metrics were calculated separately for each prior on $\boldsymbol{\beta}$. The columns n_{comp} and n_{piv} give the median values of these statistics across the 200 simulated datasets.

			median		$\boldsymbol{\beta} \sim$ HS($\tilde{p}_{\text{eff}}$)						$\boldsymbol{\beta} \sim$ Logis(1)					
Scenario	p	n	$n_{\text{comp.}}$	$n_{\text{piv.}}$	$t_3(10)$	$t_\infty(10)$	EP$_2(\sigma_n)$	EP$_4(\sigma_n)$	EP$_{10}(\sigma_n)$	Logis(σ_n)	$t_3(10)$	$t_\infty(10)$	EP$_2(\sigma_n)$	EP$_4(\sigma_n)$	EP$_{10}(\sigma_n)$	Logis(σ_n)
1	4	50	9	9	**51**	**51**	**49**	**49**	**50**	***49***	65	65	59	63	65	***59***
1	4	100	20	20	**41**	**41**	**41**	**41**	***41***	**42**	46	46	45	46	46	***45***
1	4	200	40	40	**36**	**36**	**36**	***36***	**37**	**37**	***33***	**33**	**33**	**33**	**33**	**33**
1	4	400	84	84	**28**	***27***	**27**	**28**	**27**	**28**	**24**	**24**	**24**	**24**	**24**	***24***
2	25	50	0	1	95	97	67	71	75	***67***	229	230	***207***	209	214	**208**
2	25	100	2	7.5	58	58	58	58	58	***57***	188	188	181	185	187	***179***
2	25	200	22	24.5	**34**	**34**	**34**	**34**	**34**	***34***	119	119	118	119	119	***117***
3	25	50	0	1	95	93	45	51	57	***42***	233	234	**214**	216	220	***213***
3	25	100	2	8	30	31	30	31	31	***29***	193	193	186	189	193	***184***
3	25	200	22	25	23	24	**23**	23	23	***23***	123	124	122	124	123	***122***
4	25	100	0	0	425	371	279	267	***264***	284	**180**	***180***	189	190	187	189
4	25	200	1	2	238	219	**159**	159	***158***	**160**	**137**	***136***	141	141	139	141
4	25	400	2	7	**114**	**115**	110	109	***108***	111	**102**	***102***	105	104	103	105
5	25	50	1	1	744	711	664	**639**	***609***	705	***191***	**191**	194	193	193	193
5	25	100	2	2	492	496	519	***465***	**465**	569	**152**	***152***	153	153	**152**	154
5	25	200	6	6	**146**	**147**	***144***	**145**	**145**	145	**114**	***114***	115	**114**	**115**	115
6	75	100	0	0	**155**	**153**	119	110	***102***	121	784	743	448	380	***354***	579
6	75	200	0	1	**66**	**66**	**49**	***46***	**48**	51	988	953	611	447	***358***	787
6	75	400	0	8	**34**	**34**	**31**	***31***	**33**	**31**	1076	1052	710	484	***358***	905
7	75	100	0	0	107	108	**82**	**82**	**83**	***82***	333	334	282	254	***236***	298
7	75	200	0	2	**74**	**75**	**64**	**65**	**67**	***64***	390	390	348	322	***301***	361
7	75	400	0.5	12	**30**	**30**	**30**	**30**	**30**	***29***	348	348	326	317	***305***	332
8	150	100	0	0	75	74	**72**	***72***	**72**	**72**	880	830	572	***561***	606	698
8	150	200	0	0	**72**	**72**	**71**	**71**	**71**	***71***	1161	1107	752	583	***514***	978
8	150	400	0	4	**68**	**68**	**68**	**68**	***68***	**68**	1541	1490	1072	717	***514***	1373
8	150	600	0	14	**62**	***61***	**62**	**62**	**62**	**63**	1670	1634	1218	789	***535***	1523
9	150	100	0	0	42	40	***31***	**31**	**31**	**31**	854	785	528	***515***	560	658
9	150	200	0	0	**30**	**30**	**29**	***29***	**30**	**29**	1139	1072	719	552	***484***	951
9	150	400	0	3	**28**	**28**	**28**	**28**	**28**	***28***	1495	1436	1025	681	***485***	1328
9	150	600	0	11	**28**	***28***	**28**	**28**	**28**	**28**	1656	1615	1197	758	***508***	1506
Avg. Rank (1-6)					4.07	3.94	3.25	3.15	3.25	3.35	4.72	4.35	3.07	2.80	2.63	3.44

difference in AUC between the best- and worst-performing prior on α; usually this difference was within a point. Nonetheless, alternative priors on α discriminated slightly better, as evidenced by the final row in Table S2, with a nearly 0.5 point improvement in average ranking over standard heavy tailed priors. Interestingly, for the Logistic prior, the heavy tailed priors were generally best; however, this is mitigated by the preference of the HS prior on $\boldsymbol{\beta}$ over the Logistic prior on $\boldsymbol{\beta}$.

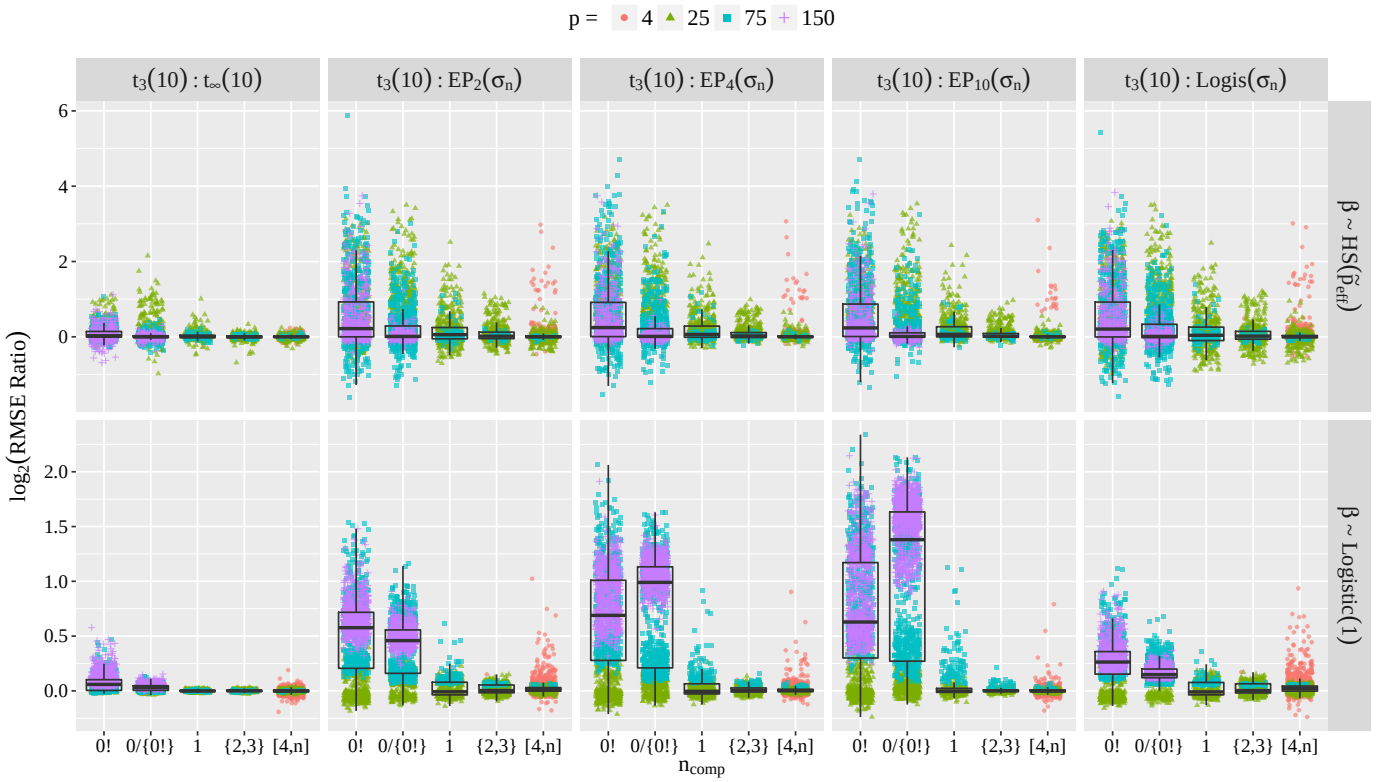

Figure 2: Comparison of $t_3(10)$ prior on α against five alternative priors on α (columns) under two different priors on $\boldsymbol{\beta}$ (rows). Each point represents an individual simulated dataset. The y-axis gives root mean-squared error (RMSE) ratios on the $\log_2$ scale when estimating $\boldsymbol{\beta}$ with its posterior mean, and the x-axis defines groups based upon separation: "0!" indicates pivotal separation ($n_{\text{piv}} = 0$); "0/{0!}" indicates complete but not pivotal separation ($n_{\text{piv}} > n_{\text{comp}} = 0$); and the remaining categories correspond to value(s) of n_{comp}. **Positive values on the y-axis indicate that the given prior on α yielded more efficient estimation of $\boldsymbol{\beta}$ than a $t_3(10)$ prior on α.** Different plot characters are used to indicate p, the number of predictors. In total, each panel contains 6000 points (30 unique scenario-sample size configuration times 200 simulated datasets per configuration).

6 Data Examples

We demonstrate our proposed priors on ten exemplar datasets with varying p/n ratios and degrees of separation and sparsity. Although these datasets have been previously reported, our re-analysis is novel in two ways. First, it quantifies heterogeneity in the degree of separation between datasets using the statistics n_{comp}, etc. Although Christmann and Rousseeuw (2001) have already studied this in five of these ten datasets, **Algorithm 2** was able to identify a tighter upper-bound on n_{comp} or n_{over} in two

of those datasets. Second, by sequentially removing increasing subsets of the n_{comp} observations from each dataset, which protect it from separation-induced variability, we characterize the stabilizing properties our proposed intercept priors have on the regression coefficients. We consider only the HS prior on $\boldsymbol{\beta}$, which was preferred to the Logistic prior on $\boldsymbol{\beta}$. The datasets are as follows:

REMISSION $n = 27$ cancer patients. The outcome is remission (33%, 9/27), with $p = 6$ patient characteristics (Lee, 1974; Christmann and Rousseeuw, 2001).

VASOCONSTRICTION $n = 39$ experiments studying the presence of 'transient reflex vasoconstriction' in subjects' fingers. The outcome is vasoconstriction (51%, 20/39) with $p = 2$ predictors (Finney, 1947; Christmann and Rousseeuw, 2001).

FOODSTAMP $n = 150$ surveyed individuals who may enroll in a foodstamp program. The outcome is participation (16%, 24/150) with $p = 3$ factors (Künsch et al., 1989; Schauberger and Tutz, 2014; Christmann and Rousseeuw, 2001).

BIRTHWEIGHT $n = 189$ births at a Springfield, MA medical center in 1986. The outcome is low-birth-weight status (31%, 59/189) with $p = 8$ possible risk factors (Jaeger et al., 1997; Venables and Ripley, 2002; Christmann and Rousseeuw, 2001).

ECMO $n = 178$ pediatric patients receiving extracorporeal membrane oxygenation (ECMO) at inpatient critical care units in three academic hospitals in North America. The outcome is short-term mortality (26%, 47/178), with $p = 22$ patient-level risk factors (Barbaro et al., 2016, 2018).

PIMA $n = 532$ records of Pima Indian heritage women tested for diabetes. The outcome was a positive test (33%, 177/532) with $p = 7$ risk factors (Venables and Ripley, 2002; Smith et al., 1988).

NES1964, NES1968 Two years' worth of national election surveys (NES), 1964 ($n = 1062$) and 1968 ($n = 851$). The outcome is preference for the Republican candidate (respectively 32%, 344/1062; 54%, 461/851), with $p = 3$ predictors. (Section 5.10, Gelman and Hill, 2007)

IVC $n = 3200$ in-vitro experiments on the thrombus-capturing efficacy of a filter in a model for the inferior vena cava (IVC). The outcome is efficacy (77%, 2452/3200),

with $p = 4$ predictors (Jaeger et al., 1997; Christmann and Rousseeuw, 2001).

HYDRAMINOS $n = 2992$ neonatal pregnancies at an academic hospital over the period of 1969-1975. The outcome is neonate mortality (0.56%, 17/2992), with $p = 14$ risk factors (Neutra et al., 1978; Greenland and Mansournia, 2015; Sullivan and Greenland, 2013).

The attributes of these ten datasets, including the separation statistics calculated using **Algorithm 2**, are given in the first column of Table 4. Quasi-complete separation, i.e. $n_{\text{over}} = 0$, exists in two of these datasets (ECMO and NES1964); none of the datasets are completely separated.

We analyzed each of the full datasets with a hierarchical shrinkage prior on $\boldsymbol{\beta}$ and one of the six priors on α, recording the posterior mean of $\boldsymbol{\beta}$. To select ψ_n in Equation (9), we used $\tilde{p}_{\text{eff}} = p/2$. We determined the set of n_{comp} observations, the removal of which would separate the data, and removed all but $k \in \{0, 1, 2\}$ of these. Thus, $k = 0$ yields separation, and $k = 1$ or 2 yield near-separation. Variability in estimation will increase with smaller values of k. Thus, we compared the posterior means of $\boldsymbol{\beta}$ from each of these subsetted analyses to the posterior mean from the full analysis with a pseudo-RMSE treating the full analysis as the 'truth':

$$\text{pRMSE}_k = \sqrt{(E_{F_{\text{post}}^{\text{SUB}_k}}[\boldsymbol{\beta}] - E_{F_{\text{post}}^{\text{FULL}}}[\boldsymbol{\beta}])^\top \Sigma_X (E_{F_{\text{post}}^{\text{SUB}_k}}[\boldsymbol{\beta}] - E_{F_{\text{post}}^{\text{FULL}}}[\boldsymbol{\beta}])}. \qquad (12)$$

Σ_X is estimated from the predictors in each full dataset, $F_{\text{post}}^{\text{SUB}_k}$ denotes the posterior given the removal of all but k of the separation-inducing observations, and $F_{\text{post}}^{\text{FULL}}$ denotes the posterior given the full the data. pRMSE_k quantifies the ability of a prior to stabilize the induced variation coming from removal of key observations. For $k = 0, 1, 2$, there are, respectively, 1, n_{comp}, and $n_{\text{comp}}(n_{\text{comp}} - 1)/2$ possible subsets; when n_{comp} or $n_{\text{comp}}(n_{\text{comp}}-1)/2$ exceeded 100, we randomly selected 100 such subsets.

Table 4 gives the values of pRMSE_k for the priors on α across all ten datasets. For $k = 1$ or 2, the median values of pRMSE_k over the selected subsets are given; the smallest such value is italicized. The EP_{10} prior on α most often yields estimates of

$\boldsymbol{\beta}$ that are closest to those coming from the full-data analysis. In several datasets, the diffuse priors, $t_3(10)$, $t_\infty(10)$, lead to somewhat dramatically different estimates of $\boldsymbol{\beta}$ as separation is approached. These results demonstrate the stabilizing effect from using locally uniform priors and reducing the scale parameter to down-weight implausible regions of the α space.

7 Conclusion

Understanding the role the intercept parameter plays in a GLM is conceptually difficult. When the covariates are not centered and the support of $\boldsymbol{X}$ lies far from the origin, the parameter lacks meaningful interpretation. Yet prediction based on a model without an intercept will, in general, incur significant error. The intercept plays the crucial role of balancing the prediction plane. When the covariates are centered, the intercept is a function of the regression coefficients, as in the third column, Table 2. In both centered and non-centered cases, it is thus apparent that efficient estimation of the intercept has bearing on efficient estimation of $\boldsymbol{\beta}$. This relationship between the intercept and regression coefficients is most explicit in binary data regressions subject to separation, which can arise due to strong associations, sparsity (rare events), binary covariates, or relatively high dimensionality of the predictor space, among other causes. Given this, we seek to better estimate the intercept and therefore regression coefficients.

Maximum likelihood is not equipped to handle separation. When the information in the data is lopsided in an extreme manner, sensible estimates of the parameters are obtained through shrinkage, e.g. Bayesian estimation. We have demonstrated that modest performance gains in estimation in a logistic regression are obtainable by a priori downweighting implausible regions of the intercept parameter space. Classical results on complete separation in logistic regression answer why this makes sense and modern Bayesian methodology and software demonstrate how to do so. The key idea is to use priors in which extreme values of the intercept are down-weighted, and **Algorithm 1** precisely defines 'extreme', given a particular sample size. Although our algorithm can

Table 4: Values of the pseudo-RMSE (pRMSE_k) metric for ten exemplar datasets described in the text, when all but $k = 2$, 1, or 0 of the n_{comp} observations have been removed; pRMSE_k is the standardized distance between the posterior means of $\boldsymbol{\beta}$ from the full data analysis and the reduced data analysis. The data are completely separated when $k = 0$.

		pRMSE_k					
Dataset	k	$t_3(10)$	$t_\infty(10)$	$\text{EP}_2(\sigma_n)$	$\text{EP}_4(\sigma_n)$	$\text{EP}_{10}(\sigma_n)$	$\text{Logis}(\sigma_n)$
REMISSION	2	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*
$\{n, p, \bar{Y}\} = \{27, 6, 0.333\}$	1	0.85	0.92	0.80	0.84	0.83	*0.78*
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{2, 2, 2\}$	0	3.11	3.17	2.76	2.96	3.03	*2.67*
VASOCONSTRICTION	2	1.19	1.24	1.16	1.28	1.20	*1.16*
$\{n, p, \bar{Y}\} = \{39, 2, 0.513\}$	1	1.56	1.68	*1.53*	1.62	1.71	1.58
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{3, 3, 3\}$	0	11.42	11.48	10.99	11.15	11.36	*10.64*
FOODSTAMP	2	8.52	8.66	3.87	3.32	*3.27*	4.70
$\{n, p, \bar{Y}\} = \{150, 3, 0.160\}$	1	14.34	13.86	4.83	3.71	*3.44*	6.61
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{17, 17, 6\}$	0	22.12	18.20	5.10	3.89	*3.49*	7.49
BIRTHWT	2	2.91	2.92	2.38	2.48	2.61	*2.34*
$\{n, p, \bar{Y}\} = \{189, 8, 0.312\}$	1	3.96	3.99	*2.83*	2.84	2.93	2.87
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{47, 46, 5\}$	0	15.51	10.53	3.37	3.24	*3.21*	3.57
ECMO	2	3.11	3.04	3.21	*2.90*	2.91	3.45
$\{n, p, \bar{Y}\} = \{178, 22, 0.264\}$	1	3.86	4.03	3.91	*3.56*	3.61	4.30
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{21, 20, 0\}$	0	5.51	5.40	4.84	4.36	*4.14*	5.75
PIMA	2	8.44	8.48	7.04	*6.89*	7.00	7.17
$\{n, p, \bar{Y}\} = \{532, 7, 0.333\}$	1	12.80	12.83	8.89	8.16	*7.85*	9.58
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{122, 102, 102\}$	0	29.85	24.81	10.85	9.03	*8.32*	13.40
NES1964	2	6.64	6.43	4.92	4.49	*4.24*	5.36
$\{n, p, \bar{Y}\} = \{1062, 3, 0.324\}$	1	8.71	8.41	5.54	4.77	*4.38*	6.58
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{346, 346, 0\}$	0	17.47	12.62	6.23	4.98	*4.46*	8.47
NES1968	2	6.53	6.50	7.19	6.47	*6.24*	7.72
$\{n, p, \bar{Y}\} = \{851, 3, 0.542\}$	1	7.01	7.03	8.19	7.25	*6.86*	8.84
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{316, 316, 3\}$	0	10.46	*10.06*	12.54	11.21	10.66	13.36
IVC	2	25.80	24.44	17.33	11.71	*8.29*	22.15
$\{n, p, \bar{Y}\} = \{3200, 4, 0.766\}$	1	35.81	31.42	19.34	12.12	*8.37*	27.92
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{488, 458, 213\}$	0	49.35	37.62	20.74	12.39	*8.42*	33.67
HYDRAMINOS	2	0.89	0.88	0.53	0.49	*0.45*	0.58
$\{n, p, \bar{Y}\} = \{2992, 14, 0.006\}$	1	1.76	1.57	0.66	0.55	*0.47*	0.77
$\{n_{\text{piv}}, n_{\text{comp}}, n_{\text{over}}\} = \{16, 14, 1\}$	0	3.79	2.36	0.65	0.51	*0.39*	0.92

be applied to any scale family of priors, one additional contribution of this paper is the novel application of an existing family of distributions, the exponential-power (EP) family, of which the Laplace and normal distributions are members. Critically, the shape of the EP family can be made platykurtotic, i.e. locally uniform, in an interval

around $\alpha = 0$ while simultaneously and smoothly vanishing for large values of $|\alpha|$.

A complementary contribution of this paper is **Algorithm 2** in the Supplement, a heuristic approach to calculate the n_{comp}, n_{over}, and n_{piv} statistics measuring the degree of separation. We are not aware of simulations studies that have stratified results based upon these statistics, as we have done here. Based upon our simulation study, using **Algorithm 1** to adapt the prior scale of whichever prior distribution is being used matters more than changing the shape of the prior distribution itself. However, our analysis of ten datasets demonstrates the stabilizing properties of the EP-family of priors, which are more locally uniform, in a variety of true data configurations.

The potential costs of our approach are small in that there is relatively little downside to excluding regions of the intercept parameter space, within which precise estimation is not possible. In cases that the likelihood suggests that the intercept is very small or very large, the parameter will still only be weakly identified. Although the efficiency gains relative to typical approaches diminish with sample size, there is generally little loss of efficiency. The scale parameter selected by **Algorithm 1** also increases with sample size, so our priors become appropriately diffuse in these data-rich(er) scenarios. Moreover, these efficiency gains held across two different priors on the regression coefficients $\boldsymbol{\beta}$, one being an adaptive shrinkage prior and the other a weakly informative prior. In this regard, our recommendation satisfies our working definition of 'default'.

Nonetheless, it is important to evaluate objections to shrinking the intercept, specifically those posed by Greenland and Mansournia (2015). One of these is that the interpretation of α is sensitive to which predictors are in the model and how they are coded. To ameliorate concerns about covariate coding, we center all covariate(s), including binary covariates, to the empiric mean. Centering covariates in this way may result in an intercept that has no physical interpretation (an observation cannot take on a proportion of a binary predictor), but it does retain an intuitive mathematical interpretation as being the log-odds (in the case of logistic regression) corresponding to an average observation. All quantities can be back-transformed to the natural scale.

Second, and more importantly, Greenland and Mansournia observe that centering

the prior at $\alpha = 0$ may not make sense in some contexts, for example, when the intercept is a function of the sampling design. When observations are retrospectively sampled based upon their outcome, the intercept does not reflect the prevalence of the outcome but rather the outcome-sampling ratio. In this case, the modeler is no longer agnostic about the intercept. This is related to the choice of location for the prior, and none of the priors considered here, all of which share a common location, would be suitable. Instead, introducing a location parameter in the EP (or any other location-scale family) could be used to reflect the outcome-sampling ratios. More to the point, assuming that the regression is focused on obtaining sensible estimates of β, we emphasize that *any* choice of prior encapsulates such assumptions and will have consequences for this objective. It is possible to be too agnostic. A very diffuse prior allows for a very small or large baseline prevalence, which will be unverifiable in most datasets. Thus, the priors that we recommend here are not unique in this limitation.

8 Acknowledgements

All numerical analyses were conducted in the R statistical environment (R Core Team, 2016; Wickham, 2009; Stan Development Team, 2018) This work was partially supported by the National Institutes of Health [P30 CA046592]

References

Albert, A. and Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71,** 1–10.

Armagan, A., Dunson, D., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica* **23,** 119–143.

Barbaro, R., Boonstra, P., Kuo, K., Selewski, D., Bailly, D., Stone, C., Chow, J., Annich, G., Moler, F., and Paden, M. (2018). Evaluating pediatric mortality risk prediction among children receiving extracorporeal respiratory support. *Submitted* .

Barbaro, R., Boonstra, P., Paden, M., Roberts, L., Annich, G., Bartlett, R., Moler, F., and Davis, M. (2016). Development and validation of the pediatric risk estimate score for children using extracorporeal respiratory support (Ped-RESCUERS). *Intensive Care Medicine* **42,** 879–888.

Bedrick, E., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* **91,** 1450–1460.

Box, G. (1953). A note on regions for tests of kurtosis. *Biometrika* **40,** 465–468.

Box, G. and Tiao, G. (1992). *Bayesian inference in statistical analysis*. John Wiley & Sons, New York, NY, Wiley Classics Library Edition edition.

Carpenter, B. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* **76,** 1–32.

Carvalho, C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97,** 465–480.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80.

Christmann, A. and Rousseeuw, P. (2001). Measuring overlap in binary regression. *Computational Statistics & Data Analysis* **37,** 65–75.

Clogg, C., Rubin, D., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* **86,** 68–78.

Finney, D. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34,** 320–334.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press New York, NY, USA.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2,** 1360–1383.

Ghosh, J., Li, Y., and Mitra, R. (2017). On the use of cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis* **doi:10.1214/17-BA1051,**.

Greenland, S. and Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine* **34,** 3133–3143.

Hanson, T., Branscum, A., Johnson, W., et al. (2014). Informative g-priors for logistic regression. *Bayesian Analysis* **9,** 597–612.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2 edition.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21,** 2409–2419.

Jaeger, H. J., Mair, T., Geller, M., Kinne, R. K., Christmann, A., and Mathias, K. D. (1997). A physiologic in vitro model of the inferior vena cava with a computer-controlled flow system for testing of inferior vena cava filters. *Investigative radiology* **32,** 511–522.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **186,** 453–461.

Künsch, H., Stefanski, L., and Carroll, R. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association* **84,** 460–466.

Lee, E. (1974). A computer program for linear logistic regression analysis. *Computer Programs in Biomedicine* **4,** 80–92.

Marin, J. and Robert, C. (2007). *Bayesian core: A practical approach to computational Bayesian statistics*. Springer Science & Business Media, New York, NY.

Neutra, R. R., Fienberg, S. E., Greenland, S., and Friedman, E. A. (1978). Effect of fetal monitoring on neonatal death rates. *New England Journal of Medicine* **299,** 324–326.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103,** 681–686.

Piironen, J. and Vehtari, A. (2015). Projection predictive variable selection using Stan+RarXiv preprint arXiv:1508.02502.

Piironen, J. and Vehtari, A. (2016). On the hyperprior choice for the global shrinkage parameter in the horseshoe priorarXiv preprint arXiv:1610.05559.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11,** 5018–5051.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rainey, C. (2016). Dealing with separation in logistic regression models. *Political Analysis* **24,** 339–355.

Santner, T. and Duffy, E. (1986). A note on a. albert and ja anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **73,** 755–758.

Schauberger, G. and Tutz, G. (2014). *catdata: Categorical Data*. R package version 1.2.1.

Smith, J. W., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In

Proceedings of the Annual Symposium on Computer Application in Medical Care, page 261. American Medical Informatics Association.

Stan Development Team (2017). *Stan Modeling Language User's Guide and Reference Manual, Version 2.17.0*. http://mc-stan.org/.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.17.3.

Sullivan, S. G. and Greenland, S. (2013). Bayesian regression in sas software. *International Journal of Epidemiology* **42,** 308–317.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74,** 646–648.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, NY.

Zellner, A. (1983). Applications of Bayesian analysis in econometrics. *Journal of the Royal Statistical Society. Series D (The Statistician)* **32,** 23–34.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis* **13,** 157–170.