Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1–24

Laplace Power-Expected-Posterior Priors for Logistic Regression*

Anupreet Porwal† and Abel Rodríguez‡

Abstract. Power-expected-posterior (PEP) methodology, which borrows ideas from the literature on power priors, expected-posterior priors and unit information priors, provides a systematic way to construct objective priors. The basic idea is to use imaginary training samples to update a (possibly improper) prior into a proper but minimally-informative one. In this work, we develop a novel definition of PEP priors for logistic regression models that relies on a Laplace expansion of the likelihood of the imaginary training sample. This approach has various advantages over previous proposals for non-informative priors in logistic regression, and can be easily extended to other generalized linear models. We study theoretical properties of the prior and provide a number of empirical studies that demonstrate superior performance both in terms of model selection and of parameter estimation, especially for heavy-tailed versions.

Keywords: generalized linear model, logistic regression, Bayesian model selection, expected-posterior priors, default priors.

1 Introduction

Generalized linear models (GLMs, e.g., see McCullagh and Nelder, 2019) are one of the main workhorses of statistical analysis. They are widely used both to model data directly and as building blocks for more complex hierarchical models. However, in spite of their broad adoption, prior elicitation for GLMs in the absence of subjective information remains an open problem, particularly in settings where the main goal is variable selection. Indeed, because standard non-informative priors for GLMs that work well for parameter estimation are often improper, they cannot be used in model selection problems, as they typically lead to ill-defined Bayes factors (e.g., see Berger et al., 2001).

Within the class of Gaussian linear models, the literature on so-called "objective" or "default" priors for model selection is extensive. Examples include point-mass spike-&-slab priors (Mitchell and Beauchamp, 1988; Geweke, 1996), g-priors (Zellner, 1986), mixtures of g-priors (Zellner and Siow, 1980; Liang et al., 2008), unit information priors (Kass and Wasserman, 1995), intrinsic Bayes factors (Berger and Pericchi, 1996a), fractional Bayes factors (O'Hagan, 1995; De Santis and Spezzaferri, 2001), non-local priors (Johnson and Rossell, 2010, 2012), power-expected-posterior priors (Fouskakis et al., 2015) and prior-based Bayesian information criterion (Bayarri et al., 2019), among other approaches. See Consonni et al. (2018) for a comprehensive review of recent approaches to objective Bayesian analysis, and Bayarri et al. (2012) for a review and discussion of

*A. Rodríguez's research was supported by NSF grant 2023495 and NSF grant 2114727.
†Department of Statistics, University of Washington, Seattle, porwaa@uw.edu
‡Department of Statistics, University of Washington, Seattle, abelrod@uw.edu

 https://doi.org/10.1214/23-BA1389

desirable properties. The literature on default priors for GLMs is more limited, with three main approaches dominating. These include those introduced by Bové and Held (2011) and Li and Clyde (2018), both of which consider modifications of mixtures of g-priors that are suitable for GLMs, and Fouskakis et al. (2018), who considers extensions of power-expected-posterior priors that rely on unnormalized power likelihoods. One feature shared by all three approaches is that they can be thought of as being based on the idea of calibrating (possibly improper) priors using either real or imaginary training samples (e.g., see Berger and Pericchi, 1996b and Pérez and Berger, 2002).

In this paper we introduce a variant of the power-expected-posterior (PEP) prior for GLMs that we call the Laplace PEP, or LPEP. While the formulation is general, this manuscript emphasizes the development of the LPEP for logistic regression models. This is because this subclass of models provides the best illustration of the theoretical and practical advantages of our approach. For example, we note that the prior described in Li and Clyde (2018) is improper when the maximum likelihood estimate (MLE) of the regression coefficients under the observed data does not exist, leading to ill-defined Bayes factors. In the case of logistic regression, this happens when there is separation among the groups (e.g., see Albert and Anderson, 1984, Heinze and Schemper, 2002 and Ghosh, 2019). Separation is reasonably common in practical applications, especially in problems with relatively small samples and several unbalanced and highly predictive risk factors. A similar issue arises with the PEP priors introduced in Fouskakis et al. (2018) since the imaginary training samples are not restricted to yield finite MLEs. Furthermore, both versions of the PEP prior proposed by Fouskakis et al. (2018) are computationally intractable, requiring the use of reversible Jump Markov Chain Monte Carlo algorithms (Green, 1995; Dellaportas et al., 2002). The LPEP is well defined under separation as long as at least one training sample exists that yields finite MLEs under the full model. Furthermore, because the LPEP can be written as a location-and-scale mixture of Gaussian distribution, it is easy to incorporate into Markov chain Monte Carlo algorithms that rely on data augmentation (e.g., Polson et al., 2013). This feature also facilitates the development of prior-based Bayesian Information Criterion (e.g., see Li and Clyde, 2018 and Bayarri et al., 2019) that avoids data augmentation (at the potential cost of accuracy). Like Bové and Held (2011), Li and Clyde (2018) and Fouskakis et al. (2018), LPEPs implicitly assume that q, the largest model size under consideration is smaller than n. However, unlike previous works, we consider the model selection consistency when both q and the number of variables p grow with n.

It is important to stress that the focus of this manuscript is on priors for variable selection that place positive probability on specific coefficients being exactly zero. An alternative approach to sparsity that is popular in the literature is to use continuous shrinkage priors (see, e.g., Bhadra et al., 2019 for a comprehensive review). Continuous shrinkage priors tend to have computational advantages and are very effective in predictive settings. However, variable selection under these priors can be performed only by thresholding. While ad-hoc techniques have been devised for this purpose (e.g., see Li and Pati, 2017), thresholding tends to work best in settings where enough prior information is available to establish practical significance.

2 Power-expected-posterior priors: A brief review

Power-expected-posterior (PEP) priors (Fouskakis et al., 2015) extend the expected-posterior (EP) priors introduced by Pérez and Berger (2002) by controlling the amount of information contained in the prior using the power approach originally developed by Ibrahim and Chen (2000) and Chen et al. (2000) in the context of subjective priors.

Briefly, let $\boldsymbol{y}$ denote the n-dimensional vector containing the observed data, $\boldsymbol{\gamma}$ index the model space, and $\boldsymbol{\beta_\gamma}$ represent vector of parameters under model $\boldsymbol{\gamma}$. We start with a (potentially improper) prior $\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta_\gamma})$ under model $\boldsymbol{\gamma}$ and introduce an n^*-dimensional vector of *imaginary* training samples arising from a distribution $m^*(\mathbf{y}^*)$. The EP prior is then constructed as

$$\pi_{\boldsymbol{\gamma}}^{EP}(\boldsymbol{\beta_\gamma}) = \int \frac{f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma}\right) \pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta_\gamma}\right)}{\int f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma}\right) \pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta_\gamma}\right) d\boldsymbol{\beta_\gamma}} m^*(\boldsymbol{y}^*) d\boldsymbol{y}^*$$

In words, the EP priors use the imaginary training sample $\mathbf{y}^*$ to update the original prior $\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta_\gamma})$, and addresses the possible effect of using any particular training sample by averaging over the distribution $m^*(\mathbf{y}^*)$. The use of a common $m^*(\boldsymbol{y}^*)$ properly calibrates the priors across the different models, even in situations where $m^*(\boldsymbol{y}^*)$ is improper. Pérez and Berger (2002) discuss various possible choices of $m^*(\boldsymbol{y}^*)$ in both informative and non-informative settings.

Note that an implicit assumption in the formulation of the PEP is that $\mathbf{y}^*$ and n^* are such that the posterior based on it is proper, i.e.,

$$\int f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma}\right) \pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta_\gamma}\right) d\boldsymbol{\beta_\gamma} < \infty \tag{1}$$

for any $\boldsymbol{y}^*$ in the support of m^* (see Pérez and Berger, 2002 for details). However, large values of n^* will produce priors that are relatively concentrated. To balance these two goals, it is common to choose n^* as the size of the minimum training sample required to satisfy (1) across all models.

Even though the EP prior attempts to ameliorate the effect of the $\boldsymbol{y}^*$ by averaging over m^* and by using training samples that are as small as possible, in some applications the prior might still be quite concentrated. Power-expected-posterior priors (Fouskakis et al., 2015) address this by scaling the likelihood of the imaginary sample,

$$\pi_{\boldsymbol{\gamma}}^{PEP}(\boldsymbol{\beta_\gamma}) = \int \frac{\tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma}, \delta\right) \pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta_\gamma}\right)}{\int \tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma}, \delta\right) \pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta_\gamma}\right) d\boldsymbol{\beta_\gamma}} m^*(\boldsymbol{y}^* \mid \delta) f(\delta \mid \boldsymbol{\gamma}) d\delta d\boldsymbol{y}^*,$$

where $\tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma}, \delta\right) = \frac{f_{\boldsymbol{\gamma}}(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma})^{\frac{1}{\delta}}}{\int f_{\boldsymbol{\gamma}}(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma})^{\frac{1}{\delta}} d\boldsymbol{\beta_\gamma}}$ is the normalized power likelihood for the training sample $\boldsymbol{y}^*$ based on model $\boldsymbol{\gamma}$, δ is the power parameter, $m^*(\boldsymbol{y}^* \mid \delta)$ is the predictive distribution generating the imaginary samples $\boldsymbol{y}^*$, and $f(\delta \mid \boldsymbol{\gamma})$ is a hyper-prior on δ. Fouskakis et al. (2015) recommended $m^*(\boldsymbol{y}^* \mid \delta) = m_0^N(\boldsymbol{y}^* \mid \delta)$, i.e., the marginal likelihood evaluated using the power likelihood of $\boldsymbol{y}^*$ under the null model and baseline prior $\pi_{\mathbf{0}}^N\left(\boldsymbol{\beta_0}\right)$. If $\delta = 1$, then PEP prior reduces to the EP prior, while values of $\delta > 1$

yield priors with a larger variance (and therefore, less information) than the EP prior. A particularly appealing choice is $\delta = n^*$ (or, alternatively, a prior on δ that is centered around n^*), which leads to a prior that can be considered as being unit information (Kass and Wasserman, 1995). Note that δ plays a similar role to the g parameter involved in the definition of (mixtures of) g priors. Hence, treating δ as random will typically lead to priors that have heavier tails. In this paper, in addition to the unit information setting where $\delta = n^*$, we consider the hyper-g/n prior from Liang et al. (2008) and the robust prior from Bayarri et al. (2012) as two choices for hyper priors for δ. See Section 3 for more details. Furthermore, being able to use the parameter δ to control the amount of information contained in the prior means that the choice of the size of the training sample is less critical in the case of PEP priors. In the sequel, we work with $n^* = n$, a choice that is particularly convenient when dealing with regression models. Indeed, taking $n^* = n$ allows us to select $\boldsymbol{X}^*$, the design matrix associated with the training sample $\boldsymbol{y}^*$, as $\boldsymbol{X}^* = \boldsymbol{X}$ (see Section 3).

The PEP prior was originally derived for model selection in Gaussian linear models (Fouskakis et al., 2015). In that case, normalizing constant $\int f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^{\frac{1}{\delta}} d\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ associated with $\tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right)$ is usually straightforward to compute. Indeed, for most standard choices of $\pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)$, the induced PEP can be written as a location-and-scale mixture of Gaussian distributions, dramatically simplifying computation within a Markov chain Monte Carlo framework. This property, however, does not extend to other GLMs. To address this issue, Fouskakis et al. (2018) introduced two different modifications of the PEP framework that rely on the unnormalized power likelihood $f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^{\frac{1}{\delta}}$ rather than $\tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right)$: the concentrated reference PEP (CRPEP) and the diffuse reference PEP (DRPEP). However, while the use of the unnormalized power likelihood avoids some of the computational difficulties associated with the original PEP prior, many of them remain. In particular, neither $\pi_{\boldsymbol{\gamma}}^{CRPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ nor $\pi_{\boldsymbol{\gamma}}^{DRPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ belong to standard families of distributions. This prevents closed-form integration of the regression coefficients and therefore requires the use Reversible Jump Markov chain Monte Carlo algorithms (Green, 1995; Dellaportas et al., 2002). Furthermore, the definition of the CRPEP and the DRPEP and the computational approach introduced by the authors (which relies on Laplace approximations to compute certain normalizing constants needed for the acceptance probabilities of various Metropolis-Hastings steps) implicitly assume that the MLE of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ exists for any training sample $\boldsymbol{y}^*$ and model $\boldsymbol{\gamma}$.

3 The Laplace power-expected-posterior prior for logistic regression

Instead of working with the unnormalized power likelihood as in Fouskakis et al. (2018), in this paper we propose replacing the likelihood of the imaginary samples with its Laplace approximation *before* raising it to the power $1/\delta$. Hence, the name Laplace PEP, or LPEP. More concretely, let the observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be generated

from a logistic regression likelihood of the form

$$f_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \exp\left\{\sum_{i=1}^{n} y_i \boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}} - \log(1 + \exp\left\{\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right\})\right\}, \tag{2}$$

where $\boldsymbol{x}_i = (1, x_{i,1}, \ldots, x_{i,p})^T$ is the $p+1$ dimensional vector of regressors associated with observation y_i, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ is the $p+1$ dimensional vector of regression coefficients (including the intercept), $\boldsymbol{\gamma}^T = (\gamma_0, \gamma_1, \ldots, \gamma_p)$ is a binary vector of length $p+1$ such that for all $j \in \{1, \ldots, p\}, \gamma_j = 1$ if the j-th variable is included in the model (i.e., if β_j is different from zero) and $\gamma_j = 0$ otherwise, and $\boldsymbol{x}_{\boldsymbol{\gamma},i}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denote the sub-vectors of $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ with length $p_{\boldsymbol{\gamma}}+1$ where $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$ that include only those components for which the corresponding γ_j is equal to 1. In the sequel we assume that $\gamma_0 = 1$ (i.e., intercept is always included in the model) and that $n > p$. The Laplace approximation to (2) is given by

$$f_{\boldsymbol{\gamma}}\left(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) \approx f_{\boldsymbol{\gamma}}^{L}\left(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_{\gamma} - \hat{\boldsymbol{\beta}}_{\gamma}(\boldsymbol{y})\right)^T \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y})\left(\boldsymbol{\beta}_{\gamma} - \hat{\boldsymbol{\beta}}_{\gamma}(\boldsymbol{y})\right)\right\}, \tag{3}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y})$ denotes the MLE of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ based on sample $\boldsymbol{y}$, and $\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y})$ is the $(p_{\boldsymbol{\gamma}}+1) \times (p_{\boldsymbol{\gamma}}+1)$ observed information matrix

$$\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}) = \sum_{i=1}^{n} \frac{\left(1 + \exp\left\{\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y})\right\}\right)^2}{\exp\left\{\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y})\right\}} \boldsymbol{x}_{\boldsymbol{\gamma},i} \boldsymbol{x}_{\boldsymbol{\gamma},i}^T.$$

In the case of logistic regression models (and of regular exponential families more broadly), it is well known that this approximation is accurate up to an $\mathcal{O}(\frac{1}{n})$ order term (e.g., see Schwarz, 1978). With this in mind, we define the LPEP as

$$\pi_{\boldsymbol{\gamma}}^{LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \int \frac{\tilde{f}_{\boldsymbol{\gamma}}^{L}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right) \pi_{\boldsymbol{\gamma}}^{N}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)}{\int \tilde{f}_{\boldsymbol{\gamma}}^{L}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right) \pi_{\boldsymbol{\gamma}}^{N}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}} m^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*) f(\delta \mid \boldsymbol{\gamma}) d\delta d\boldsymbol{y}^*, \tag{4}$$

where $\boldsymbol{X}^*$ is the $n^* \times (p+1)$ matrix whose rows correspond to the $\boldsymbol{x}_i^T$ vectors and

$$\tilde{f}_{\boldsymbol{\gamma}}^{L}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right) \propto \delta^{-\frac{p_{\boldsymbol{\gamma}}+1}{2}} \exp\left\{-\frac{1}{2\delta}\left(\boldsymbol{\beta}_{\gamma} - \hat{\boldsymbol{\beta}}_{\gamma}(\boldsymbol{y}^*)\right)^T \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\left(\boldsymbol{\beta}_{\gamma} - \hat{\boldsymbol{\beta}}_{\gamma}(\boldsymbol{y}^*)\right)\right\}.$$

As discussed in Pérez and Berger (2002), standard choices for the baseline prior include the (improper) flat prior $\pi_{\boldsymbol{\gamma}}^{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1$ and the Jeffreys prior for GLMs (Ibrahim and Laud, 1991) where $\pi_{\boldsymbol{\gamma}}^{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto \left|\mathsf{E}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\left\{\boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y})\right\}\right|^{1/2}$. In this paper, we focus our attention on the flat prior $\pi_{\boldsymbol{\gamma}}^{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1$, which was also used in Fouskakis et al. (2018). There are two reasons for this. The first one is greater mathematical tractability, as the Jeffreys prior for GLMs does not lead to tractable expressions for the LPEP. Secondly, as we show in Section 4.3, the intrinsic prior associated with the LPEP derived under $\pi_{\boldsymbol{\gamma}}^{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1$ is the same as intrinsic prior associated with Bové and Held (2011).

For the predictive distribution of the imaginary samples we consider

$$m^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*) \propto \tilde{m}^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*)\mathbf{1}\left(\mathbf{y}^* \in A(\boldsymbol{X}^*)\right), \tag{5}$$

where $\tilde{m}^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*)$ is an unrestricted predictive distribution of $\boldsymbol{y}^*$ given by

$$\tilde{m}^*(\boldsymbol{y}^*) = \frac{\Gamma\left(\sum_{i=1}^{n^*} y_i^* + \frac{1}{2}\right)\Gamma\left(n^* - \sum_{i=1}^{n^*} y_i^* + \frac{1}{2}\right)}{\Gamma(n^*+1)\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)}, \tag{6}$$

(a Beta-Binomial distribution with both parameters equal to $\frac{1}{2}$, which is the predictive distribution under the null model and its reference/Jeffreys prior), and $\mathbf{1}\left(\mathbf{y}^* \in A(\boldsymbol{X}^*)\right)$ is the indicator function on $A(\boldsymbol{X}^*) = \left\{\tilde{\boldsymbol{y}} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\left(\tilde{\boldsymbol{y}}, \boldsymbol{X}^*\right) \text{ exists and is finite for all } \boldsymbol{\gamma}\right\}$. Note that our choice for $m^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*)$ differs substantially from that in Fouskakis et al. (2018). For example, (5) *does not* depend on the scaling factor δ. This makes intuitive sense (there is no obvious reason why the power factor used to re-scale the information in the training sample should also affect how the training sample is generated), and simplifies both posterior computation and theoretical analysis. Furthermore, (5) explicitly depends on $\boldsymbol{X}^*$ in a way that ensures that the LPEP is proper (see Section 4.1).

At first sight, the computational implementation of (5) might seem daunting. However, the following theorem shows that, for a broad class of GLMs that includes logistic regression, it is enough to check the existence of the MLE for the full model.

Theorem 1. *Let $\ell_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \boldsymbol{y}) : S_{\boldsymbol{\gamma}} \to \mathbb{R}$ denote a log-likelihood in the regular exponential family of the form $\ell(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \mathbf{y}) = \sum_{i=1}^{n} T(y_i)\eta\left(\boldsymbol{x}_{\boldsymbol{\gamma},i}^T\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) + A\left(\boldsymbol{x}_{\boldsymbol{\gamma},i}^T\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)$. Assume $S_{\boldsymbol{\gamma}}$ is an open connected subset of $\mathbb{R}^{p+1}$ and let $\boldsymbol{\gamma}_F = (1, \ldots, 1)$ denote the full model. If*

(i) $\ell_{\boldsymbol{\gamma}_F}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_F}; \boldsymbol{y})$ is continuous and stricly concave on $S_{\boldsymbol{\gamma}_F}$

(ii) $\lim_{\boldsymbol{\beta}_{\boldsymbol{\gamma}_F} \to \boldsymbol{\beta}^*} \ell_{\boldsymbol{\gamma}_F}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_F}; \boldsymbol{y}) = -\infty$ for any $\boldsymbol{\beta}^* \in \partial S_{\boldsymbol{\gamma}_F}$, the closure of $S_{\boldsymbol{\gamma}_F}$.

Then, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y})$, the MLE, exists under any other model $\boldsymbol{\gamma}$.

The proof of Theorem 1 can be seen in supplementary Section A (Porwal and Rodríguez, 2023). Note that, in the case of logistic regression, (i) is satisfied as long as the design matrix $\mathbf{X}$ is full rank (which, in particular, requires $p < n$), while (ii) is satisfied as long as the data does not suffers from separation under the full model (Albert and Anderson, 1984). Separation checks can be carried out using the algorithms in the `R` package `detectseparation` (Kosmidis and Schumacher, 2020). Section 4.6 of Konis (2007) shows that the version of the test based on the dual program has the best empirical worst-case time and that it scales linearly in both sample size n and number of covariates p. Furthermore, the authors empirically showed that the dual program takes approximately the same time as fitting a GLM using iteratively re-weighted least squares (IRLS) algorithm.

Finally, we discuss the specification of the distribution on the power parameter δ. As mentioned in Section 2, we consider three alternatives. First, we investigate the unit

information LPEP (UI-LPEP) obtained by fixing $\delta = n^*$. We also consider a version of the hyper-g/n prior (Liang et al., 2008), $f^{HGN}(\delta) = \left(1 + \frac{\delta}{n^*}\right)^{-2}$. We call this the HGN-LPEP. The median of this hyper-g/n prior is equal to n^*, and the prior places much of its mass around this value. Our third alternative is a version of the robust prior recommended by Bayarri et al. (2012), $f^R(\delta \mid \boldsymbol{\gamma}) = \frac{1}{2(p_{\boldsymbol{\gamma}}+1)^{1/2}} \frac{(n^*+1)^{1/2}}{(\delta+1)^{3/2}} \mathbf{1}\left(\delta > \frac{n^*-p_{\boldsymbol{\gamma}}}{p_{\boldsymbol{\gamma}}+1}\right)$, which we call this the R-LPEP. Note that, under this prior, $\mathsf{E}(\delta) = 3\frac{n^*+1}{p^*_{\boldsymbol{\gamma}}+1} = \mathcal{O}(n^*)$.

4 Properties of the LPEP prior

4.1 Proper prior

The LPEP for logistic regression can be written as a location-and-scale mixture,

$$\pi^{LPEP}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \sum_{\boldsymbol{y}^* \in \{0,1\}^n} \left[\int \phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*), \delta \boldsymbol{H}^{-1}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\right) f(\delta \mid \boldsymbol{\gamma}) d\delta \right] m^*(\boldsymbol{y}^* \mid \boldsymbol{X}),$$

where $\phi_p(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the p-variate normal with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Note that, because the training samples are restricted to yield finite MLEs, $\phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*), \delta \boldsymbol{H}^{-1}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\right)$ is proper for any model γ. Furthermore, $m^*(\boldsymbol{y}^* \mid \boldsymbol{X})$, by construction, is also proper. Hence, the LPEP prior is also proper for every $\boldsymbol{\gamma}$.

4.2 Tail behavior

It is straightforward to see that unit information LPEP (where $\delta = n^*$), $\pi^{UI-LPEP}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ has Gaussian tails. On the other hand, as the following theorem shows, the hyper-g/n and the robust versions of the LPEP have heavier (polynomial) tails in every direction.

Theorem 2. *For any model $\boldsymbol{\gamma}$ and vector $\boldsymbol{v}$ such that $\|\boldsymbol{v}\| = 1$, let $\zeta^{HGN}(s \mid \boldsymbol{v}, \boldsymbol{\gamma}) = \pi^{HGN-LPEP}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\big|_{\boldsymbol{\beta}_{\boldsymbol{\gamma}} = s\boldsymbol{v}}$ and $\zeta^{R}(s \mid \boldsymbol{v}, \boldsymbol{\gamma}) = \pi^{R-LPEP}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\big|_{\boldsymbol{\beta}_{\boldsymbol{\gamma}} = s\boldsymbol{v}}$. Then there exist bounded functions $c^{HGN}_{\boldsymbol{\gamma}}(\boldsymbol{v})$ and $c^{R}_{\boldsymbol{\gamma}}(\boldsymbol{v})$ such that*

$$\lim_{s\to\infty} \frac{\zeta^{HGN}(s \mid \boldsymbol{v}, \boldsymbol{\gamma})}{(1 + s^2/(p_{\boldsymbol{\gamma}}+1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} = c^{HGN}_{\boldsymbol{\gamma}}(\boldsymbol{v}), \quad \lim_{s\to\infty} \frac{\zeta^{R}(s \mid \boldsymbol{v}, \boldsymbol{\gamma})}{(1 + s^2/(p_{\boldsymbol{\gamma}}+1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} = c^{R}_{\boldsymbol{\gamma}}(\boldsymbol{v}).$$

The proof is presented in supplementary Section B. One important implication of this result is that, from an estimation (rather than model selection) perspective, $\pi^{HGN-LPEP}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ and $\pi^{R-LPEP}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ are robust, in the sense of having bounded influence in the case of likelihood-prior conflict (e.g., see Andrade and O'Hagan, 2006 and Andrade and O'Hagan, 2011). A second implication relates to the existence of point estimators such as the posterior mean and the posterior variance. Ghosh et al. (2018) showed that, in the presence of separation, priors with Cauchy-like tails might lead to proper posterior distribution that might have infinite means. The results in Theorem 2 guarantee that, even in the presence of separation, the model-averaged posterior means are finite.

4.3 Intrinsic consistency

The following theorem shows that the LPEP priors for logistic regression are intrinsically consistent, i.e., that they converge to a proper prior as the size of the training sample increases (see criteria 4 of Bayarri et al., 2012).

Theorem 3. *Assume that, as n^* grows, the covariate vectors $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots$ satisfy either of the following two conditions:*

(i) If $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots$ forms a deterministic sequence, then $\frac{1}{n^*}\left[\boldsymbol{X}^*\right]^T \boldsymbol{X}^* \underset{n^*\to\infty}{\longrightarrow} \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is positive definite.

(ii) If $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots$ are random, then they are independent and identically distributed with mean $\mathbf{0}$ and finite, positive definite covariance matrix $\boldsymbol{\Sigma}$.

Then, the unit information ($\delta = n^*$), hyper-g/n and robust versions of the LPEP have proper, non-degenerate intrinsic priors of the form

$$\lim_{n^*\to\infty} \pi_{\boldsymbol{\gamma}}^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, 4\left[\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right]^{-1}\right),$$

$$\lim_{n^*\to\infty} \pi_{\boldsymbol{\gamma}}^{HGN-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \int \phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, 4\delta^*\left[\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right]^{-1}\right)(1+\delta^*)^{-2}\, d\delta^*,$$

$$\lim_{n^*\to\infty} \pi_{\boldsymbol{\gamma}}^{R-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \int \phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, 4\delta^*\left[\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right]^{-1}\right)\frac{(\delta^*)^{-3/2}}{2(p_{\boldsymbol{\gamma}}+1)^{\frac{1}{2}}}\mathbf{1}\left(\delta^* > \tfrac{1}{p_{\boldsymbol{\gamma}}+1}\right) d\delta^*,$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ is the submatrix of $\boldsymbol{\Sigma}$ that includes only the rows and columns where $\gamma_j = 1$.

A proof of this result can be seen in supplementary Section C. Interestingly, we note that these are the same intrinsic priors associated with the prior in Bové and Held (2011) under the same asymptotic regime for $\boldsymbol{X}^*$.

4.4 Model selection consistency

Model selection consistency refers to the ability of the procedure to choose the correct model as $n \to \infty$. Intuitively, when p is fixed, because the amount of information in $\pi_{\boldsymbol{\gamma}}^{LPEP}$ is kept approximately constant as n^* increases, the associated Bayes factors would behave asymptotically like those computed from the Bayesian Information Criterion, which are known to be consistent. Our results, which rely on a slight extension of those presented in Barber et al. (2016) for sparse high-dimensional logistic regression, extend this intuition to situations in which p grows with n as long as n remains larger than p and at most a moderate number of covariates $q < p$ remain active.

We consider a sequence of variable selection problems indexed by the sample size n, where $y(n)$ represents the sample for the n-th problem, p_n is the total number of covariates, $\boldsymbol{\beta}_T(n)$ is the true parameter, which is associated with the true model $\boldsymbol{\gamma}_T(n)$, and $p_{\boldsymbol{\gamma}_T(n)} = \sum_{j=1}^{p_n} \gamma_{T,j}(n)$. Our interest lies in the recovery of $\boldsymbol{\gamma}_T(n)$ over the set $\Gamma = \left\{\boldsymbol{\gamma}(n) : \boldsymbol{\gamma}(n) \in \{0,1\}^{p_n},\, p_{\boldsymbol{\gamma}(n)} \leq q_n\right\}$. An implicit assumption is that the true model

is contained in the set of models under consideration. The following theorem, a proof of which can be seen in supplementary Section D, formalizes the result for UI-LPEP prior.

Theorem 4. *Assume that:*

(i) $q_n = n^{\psi}$ *for* $0 \leq \psi < 1/3$.

(ii) $p_n = n^{\kappa}$ *for* $\psi < \kappa < 1$.

(iii) $\boldsymbol{\beta}^{min}_{\boldsymbol{\gamma}_T(n)}(n) = \min_{j:\boldsymbol{\gamma}_{T,j}(n)=1} \left|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{T,j}(n)}(n)\right| \geq n^{-\phi/2}$ *for some* $0 \leq \phi < 1 - \psi$.

(iv) $\|\boldsymbol{\beta}_T(n)\|_2 \leq a_0$ *for a fixed constant* $a_0 \in (0, \infty)$.

(v) For every $i = 1, 2, \ldots$, *the vector* $\boldsymbol{x}_i$ *is such that* $\|\boldsymbol{x}_i\|_2$ *is bounded by a constant.*

(vi) For all n, *the smallest eigenvalue of* $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ *is bounded from below by a positive constant.*

(vii) $P(\boldsymbol{\gamma}(n)) \propto \binom{p_n}{p_{\boldsymbol{\gamma}(n)}}^{-1} \mathbb{I}\{p_{\boldsymbol{\gamma}(n)} \leq q_n\}$.

Define $\tilde{\boldsymbol{\gamma}}(n) = \arg\max_{\boldsymbol{\gamma}(n)}\{P(\boldsymbol{\gamma}(n)) \times m^{UI-LPEP}_{\boldsymbol{\gamma}(n)}(\boldsymbol{y}(n))\}$, *where* $m^{UI-LPEP}_{\boldsymbol{\gamma}(n)}(\boldsymbol{y}(n))$ *is the marginal likelihood of* $\boldsymbol{y}(n)$. *Then* $\mathsf{Pr}(\tilde{\boldsymbol{\gamma}}(n) = \boldsymbol{\gamma}_T(n) | \boldsymbol{y}(n)) \to 1$ *as* $n \to \infty$.

Conditions (i) and (ii) are statements about the rate of growth of the number of parameters and the maximum size of the model, while condition (iii) relates them to the rate of *decrease* in the minimum signal size. Note that when $\phi = 0$, (iii) holds true for all $0 \leq \psi < 1/3$. However, in the worst case, when q_n grows at the rate that is $\approx n^{1/3}$, we require $\boldsymbol{\beta}^{min}_{\boldsymbol{\gamma}_T(n)}(n) \geq n^{-1/3}$. Condition (iv) is necessary to avoid separation. Conditions (v) and (vi) are standard conditions for the existence of maximum likelihood estimators (e.g., see Wedderburn, 1976). Assumptions (iv), (v) and (vi), or their implications, have appeared previously in the literature (e.g., see Chen and Chen, 2012 and Luo and Chen, 2013). Condition (vii) limits the size of the models under consideration using a truncated Beta-Binomial prior. Assuming priors on models that heavily penalize large models is also common in high-dimensional regression (e.g., see Rossell et al., 2021).

5 Computation

5.1 Markov chain Monte Carlo sampling

The LPEP prior can be easily combined with the Polya-Gamma augmentation of Polson et al. (2013) to generate an efficient Markov chain Monte Carlo algorithm for variable selection in logistic regression. For this purpose, it is convenient to re-express (4) as

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*, \delta, \boldsymbol{\gamma} \sim \mathsf{N}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*), \delta\{\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\}^{-1}\right), \tag{7}$$

with $y^* \sim m^*(y^* \mid \boldsymbol{X}^*)$ and $\delta \mid \boldsymbol{\gamma} \sim f(\delta \mid \boldsymbol{\gamma})$. We can then use this hierarchical framework to first sample from the conditional posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta \mid \boldsymbol{y}^*, \boldsymbol{y})$ followed by sampling from the full conditional posterior distribution of $(\boldsymbol{y}^* \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta, \boldsymbol{y})$.

Consider first sampling $(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta \mid \boldsymbol{y}^*, \boldsymbol{y})$. From Theorem 1 of Polson et al. (2013),

$$f_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto \prod_{i=1}^{n} \left(\exp\left\{(y_i - 1/2)\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right\} \int_0^{\infty} \exp\left\{-\frac{\omega_i}{2}\left(\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^2\right\} f(\omega_i \mid 1, 0) d\omega_i \right),$$

where $f(\omega \mid a, b)$ denotes the density of a Pòlya-Gamma random variate with parameters a and b. Therefore, after introducing a vector of auxiliary variables $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$,

$$f(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y}) \propto f(\boldsymbol{\gamma}) f(\delta \mid \boldsymbol{\gamma})$$
$$\phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*), \delta \boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{y}^*)\right) \phi_n\left(\boldsymbol{z} \mid \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\Omega}^{-1}\right), \tag{8}$$

where $\boldsymbol{z} = ((y_1 - 1/2)/\omega_1, \ldots, (y_n - 1/2)/\omega_n)^T$, $\boldsymbol{\Omega} = \text{diag}\{\omega_1, \ldots, \omega_n\}$, $f(\boldsymbol{\gamma})$ is a prior on 2^p dimensional model space and $f(\delta \mid \boldsymbol{\gamma})$ is the prior on scale parameter δ. It is straightforward to see that $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ can be integrated out of (8), yielding

$$f(\boldsymbol{\gamma}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y}) = \int f(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}} \propto f(\boldsymbol{\gamma}) f(\delta \mid \boldsymbol{\gamma}) \phi_n(\boldsymbol{z} \mid \boldsymbol{m}_{\boldsymbol{z}}^{\boldsymbol{\gamma}}, \boldsymbol{V}_{\boldsymbol{z}}^{\boldsymbol{\gamma}}), \tag{9}$$

where $\boldsymbol{m}_{\boldsymbol{z}}^{\boldsymbol{\gamma}} = \boldsymbol{X}_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)$, $\boldsymbol{V}_{\boldsymbol{z}}^{\boldsymbol{\gamma}} = \boldsymbol{\Omega}^{-1} + \delta \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{y}^*) \boldsymbol{X}_{\boldsymbol{\gamma}}^T$. Then, various versions of Metropolis-Hastings algorithms can be implemented to explore the space of models (e.g., see section 4.5 of George and McCulloch, 1997).

Once the model $\boldsymbol{\gamma}$ and the exponent δ have been updated, the regression coefficients can be sampled using the fact that $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \delta, \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y} \sim \mathsf{N}(\boldsymbol{m}_{\boldsymbol{\gamma},\boldsymbol{\omega}}, \boldsymbol{V}_{\boldsymbol{\gamma},\boldsymbol{\omega}})$, where

$$\boldsymbol{m}_{\boldsymbol{\gamma},\boldsymbol{\omega}} = \boldsymbol{V}_{\boldsymbol{\omega}} \left(\boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\Omega} \boldsymbol{z} + \frac{1}{\delta} \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \right), \quad \boldsymbol{V}_{\boldsymbol{\gamma},\boldsymbol{\omega}} = \left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{\Omega} \boldsymbol{X}_{\boldsymbol{\gamma}} + \frac{1}{\delta} \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \right)^{-1},$$

and each ω_i can be updated from $f(\omega_i \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta, \mathbf{y}^*, \mathbf{y})$, which corresponds to an updated Pólya-Gamma distribution. Finally, $(\boldsymbol{y}^* \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta, \boldsymbol{y})$ can be easily updated using either Gibbs sampling or random-walk Metropolis-Hastings steps. Further details of the computational algorithm can be seen in supplementary Section E.

5.2 Model search using a prior-based Bayesian information criteria

In order to accelerate computation, Li and Clyde (2018) propose to use their default prior, $\beta_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \delta \sim \mathsf{N}\left(\mathbf{0}, \delta\{\boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y})\}^{-1}\right)$, to construct a prior-based Bayesian Information Criterion (pBIC). In situations where p is at least moderately large, this pBIC is then embedded into a random walk Metropolis-Hastings on the model space that has many similarities with the algorithm described in the previous section. This approach, can be applied across a wide variety of GLMs. In logistic regression, it sidesteps the need to perform the kind of data augmentation with Pólya Gamma random variables, potentially leading to computational gains (at the potential expense of accuracy).

A similar approach can be developed for the LPEP priors. In particular, we can use the Laplace approximation in (3) (applied this time to the likelihood of the observed data) in combination with (7) to get

$$f(\mathbf{y} \mid \gamma, \delta, \mathbf{y}^*) \approx f^L(\mathbf{y} \mid \gamma, \delta, \mathbf{y}^*) = f_{\boldsymbol{\gamma}}\left(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right) \delta^{-\frac{p_{\boldsymbol{\gamma}}+1}{2}} \left(\frac{|\boldsymbol{H}_{\boldsymbol{\gamma}}|}{|\tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}}|}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^T \boldsymbol{H}_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} + \frac{1}{\delta}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{*T} \boldsymbol{H}_{\boldsymbol{\gamma}}^* \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^* - \tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}}^T \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}} \tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}}\right]\right\}, \tag{10}$$

where

$$\tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}} = \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}}^{-1}\left[\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y})\hat{\boldsymbol{\beta}}_{\gamma}(\boldsymbol{y}) + \frac{1}{\delta}\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\hat{\boldsymbol{\beta}}_{\gamma}(\boldsymbol{y}^*)\right], \quad \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}} = \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}) + \frac{1}{\delta}\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*). \tag{11}$$

Equation (10) can be used to approximate the acceptance probabilities of a Metropolis-Hastings algorithm that explores the posterior distribution $f(\boldsymbol{\gamma}, \delta, \mathbf{y}^* \mid \mathbf{y})$ by alternating between sampling from $f(\boldsymbol{\gamma}, \delta \mid \mathbf{y}^*, \mathbf{y})$ and $f(\mathbf{y}^* \mid \boldsymbol{\gamma}, \delta, \mathbf{y})$. Then, samples for the coefficients can be obtained from the approximate posterior $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \delta, \mathbf{y}^*, \mathbf{y} \sim \mathsf{N}\left(\tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}}, \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}}^{-1}\right)$. Further details are provided in supplementary Section F.

6 Simulation studies

We conducted two simulation studies to compare the performance of Laplace PEP priors with other existing model selection techniques. This section discusses the results from the first study. Results for the second one can be seen in supplementary Section I.

The simulation study uses $n = 500$ and $p = p_{\boldsymbol{\gamma}_F} = 100$, with the predictors being drawn independently from a zero-mean, unit-scale multivariate normal distribution with correlations given by $cor(x_{i,j}, x_{i,j'}) = r^{|j-j'|}$ for $1 \leq j < j' \leq p$. It involves eight scenarios, which differ in terms of the sparsity level in the coefficients and the correlation structure. More specifically, we consider all combinations of four levels of sparsity ($p_{\boldsymbol{\gamma}_T} \in \{0, 5, 10, 20\}$, please see Table 1) and two correlation coefficients ($r \in \{0, 0.9\}$).

A total of 100 datasets were generated for each of our 8 scenarios. We apply both Bayesian procedures and penalized likelihood approaches to each dataset. In terms of Bayesian procedures, we implement the LPEP prior using (i) the "exact" MCMC

$p_{\boldsymbol{\gamma}_T}$	$\beta_{\boldsymbol{\gamma}_T,0}$	$\beta_{\boldsymbol{\gamma}_T,1:5}$	$\beta_{\boldsymbol{\gamma}_T,6:10}$	$\beta_{\boldsymbol{\gamma}_T,11:15}$	$\beta_{\boldsymbol{\gamma}_T,16:20}$
0	-0.5	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$
5	-0.5	$\boldsymbol{b}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$
10	-0.5	$\boldsymbol{b}$	$\mathbf{0}$	$\boldsymbol{b}$	$\mathbf{0}$
20	-0.5	$\boldsymbol{b}$	$0.5\boldsymbol{b}$	$\boldsymbol{b}$	$0.5\boldsymbol{b}$

Table 1: Value of intercept and coefficients in the true logistic regression model where $\boldsymbol{b} = (2, -1, -1, 0.5, -0.5)^T$.

procedure discussed in 5.1 (denoted as LPEPE in the sequel) and (ii) the "approximate" MCMC discussed in 5.2 (denoted as LPEPL). We also consider the methodology of Li and Clyde (2018), which relies on a mixture of g-priors using (i) a Laplace approximation to compute the associated marginal likelihood (denoted LCL in the sequel), as well as (ii) an "exact" version of their procedure based on a latent-variable augmentation similar to the one described in Section 5.1 (denoted as LCE). Comparing LPEPE, LPEPL, LCE and LCL allows us to disentangle the effect of the Laplace approximation from that of the prior choice on the performance of these techniques. We use the `R` package `BAS-V1.5.5` (Clyde, 2020) to implement LCL, and a slight modification of our own code to implement LCE. In all cases, we assume a Beta-Binomial(1,1) prior over the model space, and run the MCMC chain for $2^{17} \approx 131,000$ iterations after a burn-in of $10,000$ iterations. We do not include the CRPEP and DRPEP priors from Fouskakis et al. (2018) in this simulation study for two reasons. First, the computational complexity of the methods makes a simulation study like this prohibitive. Not only is each iteration of the algorithm more expensive than those of the other approaches, but the algorithm mixes more slowly. Secondly, the algorithm provided by the authors broke down for a number of our simulated datasets. In terms of penalized likelihood methods, we compare against LASSO (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010). We use the default implementations of `R` package `glmnet` (Friedman et al., 2010) to implement LASSO, and the package `ncvreg` (Breheny and Huang, 2011) for SCAD and MCP.

We evaluate the performance of these various methods in terms of model selection performance using three metrics. First, we report the frequency with which the MAP model matches the true model $\boldsymbol{\gamma}_T$ (see Table 2). For the penalized likelihood approaches (for which a single model is reported for each dataset) the equivalent metric is simply the number of datasets for which the technique reported the correct model.

Bayesian methods clearly outperform penalized likelihood approaches. Furthermore, most Bayesian approaches tend to perform very well when the data is generated from the null model, both in the uncorrelated and highly correlated cases. On the other hand, as the number of non-zero coefficients increases, we observe that all approaches struggle to identify the true model, particularly when the covariates are highly correlated. When $p_{\boldsymbol{\gamma}_T} = 20$, none of the procedures is able to identify the true model. Nonetheless, it appears that, overall, LPEPE (and, specially, the robust and the hyper-g/n versions of LPEPE) perform the best, and that the exact versions of the procedures (LPEPE and LCE) perform better than their approximate counterparts (LPEPL and LCL).

While the MAP metric we discussed above provides some insights into model performance, it tends to be less informative when there is substantial uncertainty on the posterior distribution over the model space. Therefore, we also compute for each dataset the F_1 score for the MAP (Bayesian procedures) or selected (penalized likelihood procedures) model (see Figure 1). In this setting, the F_1 score is defined as the harmonic mean of the proportion of true positives among selected covariates (the precision) and the proportion of selected covariates among true positive covariates (the recall). We focus on precision and recall rather than false positive and false negatives because of the class imbalance implied by the sparse nature of the true models used in our simulation. Results are not presented for the null model since the F_1 score is not well defined

p		100							
$p(\boldsymbol{\gamma})$		Beta-Binomial(1,1)							
$p_{\boldsymbol{\gamma}_T}$		0		5		10		20	
r		0	0.9	0	0.9	0	0.9	0	0.9
$\delta = n$	LPEPE	99	**100***	47	0	**13**	0	0	0
	LPEPL	**100***	**100***	47	0	12	0	0	0
	LCE	**100***	**100***	**48**	0	9	0	0	0
	LCL	**100***	**100***	44	0	9	0	0	0
$\delta \sim$ robust	LPEPE	99	**100***	**50**	0	**14***	0	0	0
	LPEPL	**100***	**100***	**50**	0	8	0	0	0
	LCE	99	**100***	39	0	0	0	0	0
	LCL	**100***	**100***	45	0	2	0	0	0
$\delta \sim$ hyper g/n	LPEPE	**99**	**98**	**51***	0	**11**	0	0	0
	LPEPL	98	**98**	50	0	6	0	0	0
	LCE	97	**98**	21	0	0	0	0	0
	LCL	66	80	4	0	0	0	0	0
	LASSO	59	66	0	0	0	0	0	0
	SCAD	57	62	0	0	0	0	0	0
	MCP	73	70	**8**	0	**3**	0	0	0

Table 2: Number of times (**over 100 replications**) that the **MAP** model coincides with the true model in the logistic regression ; **BOLD** represent group maximum; * represent overall maximum.

in that scenario. In all cases, the methods based on LPEP priors tend to have higher F_1 scores, with the robust and hyper-g/n versions performing slightly better than the unit information prior. We also see that, while all Bayesian procedures have very similar performance under the unit information prior, exact versions of LPEP and the Li and Clyde (2018) prior tend to outperform approximate versions under the robust and hyper-g/n priors (in some cases, quite dramatically). We see a similar pattern among methods in average model size selected for each data set (see supplementary Section G).

Next, we compare the procedures using the average mean squared error (AMSE) of the estimated coefficients, $AMSE(\boldsymbol{\beta}) = \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \beta_{j,\boldsymbol{\gamma}_T})^2$, where $\hat{\beta}_j$ and $\beta_{j,\boldsymbol{\gamma}_T}$ are the estimated and true values of j^{th} covariate, respectively. For the Bayesian procedures, model-averaged posterior mean estimates are used for this calculation. For penalized likelihood methods, the sparse point estimates of the coefficients are used. The results in Table 3 indicate that, as the true model size $p_{\boldsymbol{\gamma}_T}$ and the true correlation between covariates increases, the AMSE increases for all techniques. However, as for other metrics, heavy tailed versions of LPEP significantly outperform all other techniques in terms of estimation performance under non-null true model scenarios.

Further discussions of the tradeoff between computational complexity and accuracy for LPEPE, LPEPL, LCE and LCL can be found in supplementary Section H.

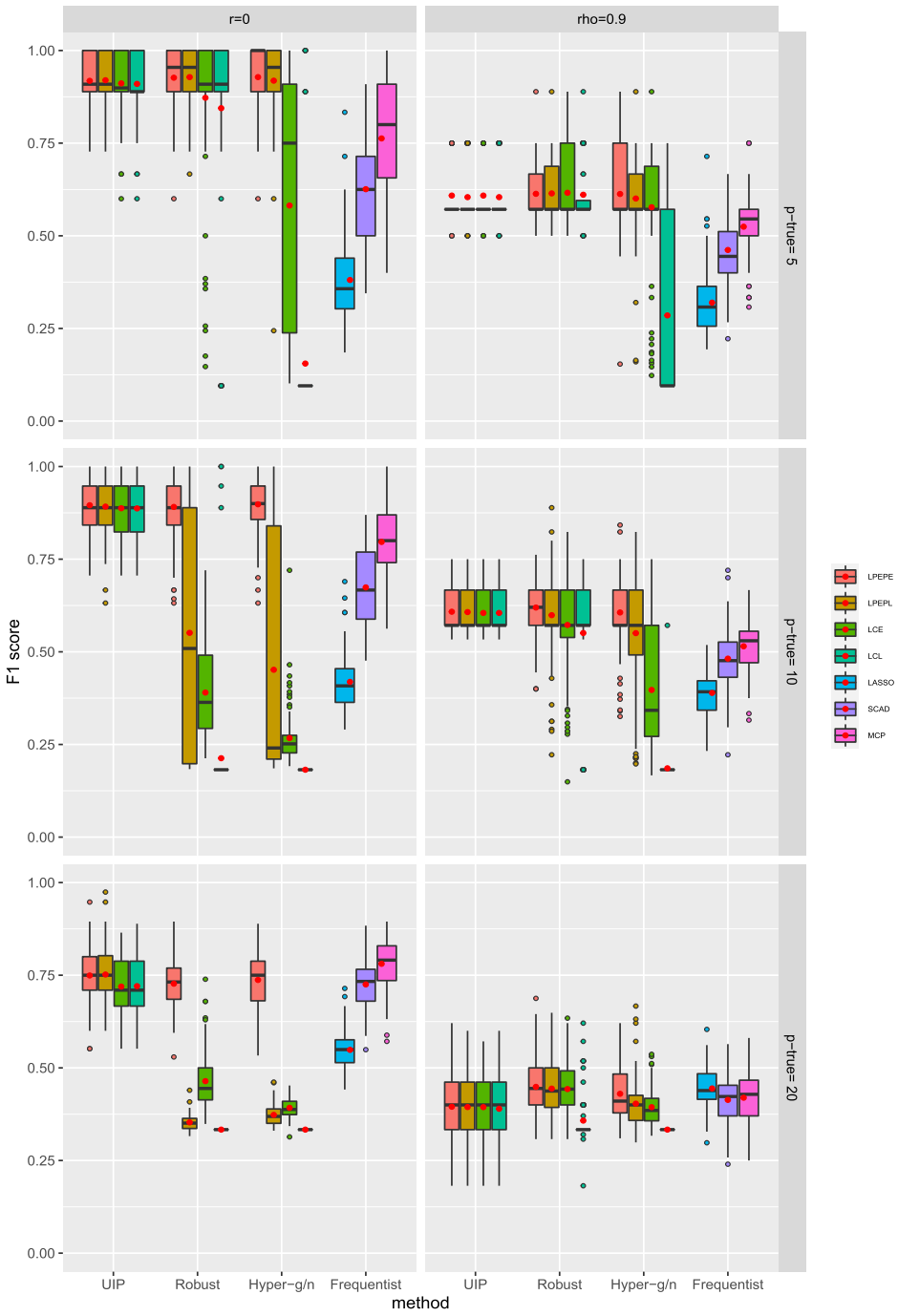

Figure 1: F1 score for the MAP model estimated by various methods and prior combinations for 100 simulated datasets ($n = 500, p = 100$) under different scenarios of correlation ($r = 0$: left; $r = 0.9$: right) and true number of non-zero coefficients specified in rows ($p - true = p_{\boldsymbol{\gamma}_T}$); Red dots represent the average F1 score.

p		100							
$p(\boldsymbol{\gamma})$		Beta-Binomial(1,1)							
$p_{\boldsymbol{\gamma}_T}$		0		5		10		20	
r		0	0.9	0	0.9	0	0.9	0	0.9
$\delta = n$	LPEPE	0.11	**0.10***	2.93	16.06	7.84	31.56	15.89	**59.19**
	LPEPL	**0.10***	**0.10***	**2.73**	**15.84**	**6.64**	**31.46**	**14.65**	59.24
	LCE	0.11	**0.10***	3.05	16.17	8.30	31.92	16.96	59.94
	LCL	**0.10***	**0.10***	2.87	16.05	7.09	31.57	16.31	60.09
$\delta \sim$ robust	LPEPE	0.12	**0.10***	2.64	14.58	**6.35***	26.61	**13.47**	48.96
	LPEPL	0.11	**0.10***	**2.51***	**14.37**	42.64	**26.34***	125.96	**47.21***
	LCE	0.12	**0.10***	5.57	14.47	68.70	29.67	147.02	58.43
	LCL	**0.10***	**0.10***	8.80	14.58	3067.22	41.31	1744.03	112.74
$\delta \sim$ hyper g/n	LPEPE	**0.15**	0.14	2.71	**14.11***	**6.59**	**26.52**	**13.29***	**48.16**
	LPEPL	0.18	0.15	**2.58**	14.14	62.95	27.61	146.06	55.29
	LCE	0.22	**0.12**	8.02	14.45	53.17	33.20	76.12	55.21
	LCL	0.30	0.43	35.33	27.52	146.80	73.60	164.25	94.32
	LASSO	0.25	**0.21**	7.08	19.62	16.96	**33.66**	28.99	**54.68**
	SCAD	**0.21**	0.79	3.07	**19.31**	6.85	41.61	14.99	70.28
	MCP	0.22	0.23	**2.82**	19.96	**6.63**	45.80	**14.78**	69.89

Table 3: 1000 times the AMSE for estimated coefficients over 100 replications; **BOLD** represent group minimum; * represent overall minimum.

7 Real data applications

This section discusses the performance of the LPEP in two real datasets. Three additional datasets are considered in supplementary Sections J, K and L.

7.1 `URINARY`: Determinants of urinary incontinence

The `URINARY` data set (Potter, 2005; Mansournia et al., 2018) describes the results of a small drug study with 21 subjects. The response corresponds to whether the subject developed urinary incontinence after receiving the drug. The explanatory variables capture drug-induced physiological changes. While small, the data set is challenging to analyze because it exhibits full separation.

Table 4 presents estimates of the regression coefficients for various Bayesian and penalized likelihood methods. The results for LPEPE, CRPEP and DRPEP are based on 10,000 iterations of the MCMC algorithm obtained after a burn-in period of 10,000 iterations. On the other hand, for LCL we use the full model enumeration procedure in the `R` package `BAS`. For Bayesian procedures, we present model-averaged posterior means and 95% credible intervals. Confidence intervals for the penalized likelihood procedures are not presented because the `R` packages used to fit the models do not provide them.

Note that LCL always produces large point estimates and very wide credible for the model coefficients. This is no surprise; the prior proposed by Li and Clyde (2018) is proper only for models for which the MLEs are finite. Hence, for a data set like `URINARY`, some of the Bayes factors associated with LCL are ill-defined. This is also why we do not

		β_0	β_1	β_2	β_3
$\delta = n$	LPEPE	0.56 (−1.66, 2.85)	−0.70 (−2.32, 0.10)	−0.39 (−0.81, −0.13)	0.15 (0.00, 0.37)
	LCL	−83.84 (−6009.26, 5897.13)	−2333.88 (−161488.87, 158312.76)	−1578.58 (−109266.01, 107118.14)	296.17 (−19896.81, 20678.41)
	CRPEP	−1.15 (−3.21, 0.52)	−0.70 (−1.88, 0.30)	−0.34 (−0.63, −0.10)	0.00 (−0.00, 0.00)
	DRPEP	0.69 (−0.55, 2.13)	−1.00 (−2.09, −0.19)	0.00 (0.00, 0.00)	0.06 (−0.03, 0.16)
$\delta \sim$ robust	LPEPE	0.71 (−1.74, 3.55)	−0.98 (−3.82, 0.05)	−0.52 (−1.89, −0.12)	0.19 (0.00, 0.54)
	LCL	−83.84 (−6250.14, 5980.74)	−2148.67 (−161065.38, 154146.29)	−1453.30 (−108979.51, 104298.98)	272.66 (−19890.08, 20102.78)
$\delta \sim$ hyper g/n	LPEPE	0.61 (−1.65, 3.07)	−0.75 (−2.74, 0.10)	−0.41 (−1.02, −0.09)	0.15 (0.00, 0.39)
	LCL	−83.84 (−6252.44, 5650.45)	−1288.82 (−124412.67, 113166.12)	−871.72 (−84179.77, 76570.77)	163.55 (−15457.94, 14685.16)
	CRPEP	−1.04 (−3.04, 0.62)	−0.66 (−1.78, 0.30)	−0.33 (−0.65, −0.08)	0.00 (−0.00, 0.00)
	DRPEP	−0.89 (−3.11, 0.69)	0.00 (0.00, −0.00)	−0.36 (−0.76, −0.11)	0.00 (−0.00, 0.00)
	LASSO	0.36	−0.70	−0.31	0.11
	SCAD	0.41	−0.23	−0.20	0.07
	MCP	0.40	−0.17	−0.20	0.07

Table 4: Estimated BMA coefficients and 95% credible intervals for Bayesian techniques for `urinary` dataset; For frequentist techniques, estimated coefficient is displayed.

show results for LCE and LPEPL; the posterior distribution for the associated MCMC algorithms is improper if the full model is included in the analysis. Furthermore, note that CRPEP yields point estimates that appear to be different from those generated by LPEPE, DRPEP, and the penalized likelihood methods. This is clearer when looking at the intercept of the model, which is negative with high probability under CRPEP but positive with high probability under LPEPE and DRPEP under all hyperpriors.

Next, we present in Table 5 the posterior inclusion probabilities (PIPs) associated with each of the three variables, along with the model selected by each of the penalized likelihood methods. In all cases, LCL consistently places probability one on all variables.This is consistent with the results generated by the penalized likelihood methods. On the other hand, LPEPE places very high PIP for the second and third variables, but only a moderately high PIP for the first one. These results are consistent for all specification of δ. In contrast, the results for the original PEP procedures in Fouskakis et al. (2018) are completely different and, more importantly, inconsistent across the CRPEP and DRPEP and various choices of δ. For example, while the CRPEP consistently favors excluding the third covariate, the DRPEP favors dropping either the second, or both the first and the third covariates depending on which hyperprior is used for δ.

		$P(\gamma_1 \neq 0 \mid \boldsymbol{y})$	$P(\gamma_2 \neq 0 \mid \boldsymbol{y})$	$P(\gamma_3 \neq 0 \mid \boldsymbol{y})$
$\delta = n$	LPEPE	0.725	0.996	0.908
	LCL	1.000	1.000	1.000
	CRPEP	1.000	1.000	0.000
	DRPEP	1.000	0.000	1.000
$\delta \sim$ robust	LPEPE	0.743	0.996	0.901
	LCL	1.000	1.000	1.000
$\delta \sim$ hyper g/n	LPEPE	0.732	0.992	0.891
	LCL	1.000	1.000	1.000
	CRPEP	1.000	1.000	0.000
	DRPEP	0.000	1.000	0.000
	LASSO	1.000	1.000	1.000
	SCAD	1.000	1.000	1.000
	MCP	1.000	1.000	1.000

Table 5: Marginal posterior inclusion probabilities (PIPs) for `urinary` dataset (Bayesian procedures) and variables included in the model (penalized likelihood methods).

7.2 `GUSTO-I`: Survival to treatments for occluded coronary arteries

Next, we consider data from the Global Utilization of Streptokinase and TPA for Occluded Coronary Arteries (GUSTO-I) trial (Califf et al., 1996), which was previously analyzed in Held et al. (2015) and Li and Clyde (2018). We aim to model the binary endpoint of 30-day survival for a subgroup of $n = 2188$ patients using 17 clinical covariates described in supplementary Section M.

Figure 2 displays the PIPs for Bayesian methods and the inferred model under the penalized likelihood techniques. For all other Bayesian techniques, we use 131,000 iterations with a burn-in of 10,000 iterations. As in our simulation studies, all penalized likelihood techniques select denser models than the Bayesian procedures. In line with Li and Clyde (2018) and Held et al. (2015), we observe that `AGE`, `KILLIP`, `HYP`, `HRT` and `STE` have high PIPs under all methods. However, the different versions of LCL perform quite differently. In particular, the version of LCL that relies on a hyper-g/n hyperprior tends to explore very dense models leading, to PIPs close to 0.5 for all variables. Similarly, the hyper-g/n versions of CRPEP and DRPEP seem to differ from their $\delta = n$ versions with respect to `PMI` and `SEX` variables. On the other hand, the different versions of LPEP are roughly in agreement for all variables.

We also compare the different procedures in terms of their out-of-sample predictive performance using 10-fold cross-validation. Table 6 presents the average value of four different metrics across all 10 folds: the area under the ROC curve (AUC), the Calibration Slope (CS), the Logarithmic Score (LS) and the Brier score (BRIER). AUC and CS allow us to evaluate discrimination and calibration. In both cases, scores closer to 1 indicate better performance. On the other hand, LS and BRIER measure the predictive accuracy of methods. In both cases, lower scores indicate better performance. Most

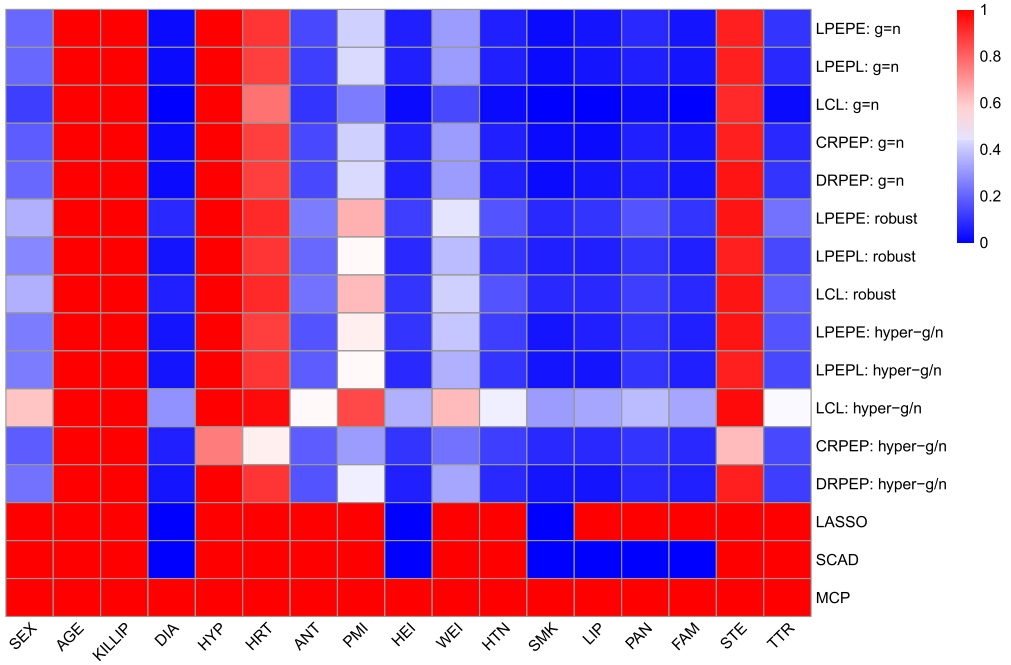

Figure 2: Marginal posterior inclusion probabilities (PIPs) for `GUSTO-I` data (Bayesian procedures) and variables selected in the model (penalized likelihood methods).

methods perform similarly under these metrics. The main exceptions are both versions of CRPEP and DRPEP, which seem to substantially underperform. LPEP procedures slightly outperform other methods in terms of AUC and CS. On the other hand, LASSO seems to slightly outperform all Bayesian procedures in terms of LS and Brier score.

8 Discussion

Our results show that LPEP priors for logistic regression are superior to existing techniques, both in terms of model selection and of parameter estimation. The differences are particularly striking when comparing the LPEP with the original CRPEP and DRPEP proposed in Fouskakis et al. (2018), and for heavy-tailed versions of mixtures of g-priors (Li and Clyde, 2018). When compared against the CRPEP and DRPEP, LPEP priors substantially reduce the computational burden associated with the use of imaginary samples. Furthermore, our empirical analyses show that the results generated by the CRPEP and DRPEP can differ substantially from each other and from the consensus of other methods, and that they can are affected by the choice of hyperpriors.

We were surprised by the poor behavior of the heavy-tailed versions of LCL and LCE procedures. One point to note is that the setup of the simulations in Section 6 ($n = 500$, $p = 100$) was only briefly studied in Li and Clyde (2018). Indeed, most of the

		AUC	CS	LS	BRIER
$\delta = n$	LPEPE	**0.8324***	**0.9971***	**0.1824**	0.0496
	LPEPL	**0.8324***	1.0082	**0.1824**	**0.0495**
	LCL	0.8300	0.9931	0.1831	0.0497
	CRPEP	0.7789	1.0578	0.1965	0.0521
	DRPEP	0.7790	1.0569	0.1963	0.0521
$\delta \sim$ robust	LPEPE	**0.8322**	**1.0129**	0.1822	**0.0495**
	LPEPL	0.8320	1.0239	**0.1820**	**0.0495**
	LCL	0.8316	0.9804	0.1822	**0.0495**
$\delta \sim$ hyper g/n	LPEPE	0.8319	**1.0074**	0.1823	0.0495
	LPEPL	**0.8322**	1.0197	0.1821	0.0495
	LCL	0.8311	1.0109	**0.1818**	**0.0493**
	CRPEP	0.7956	1.1677	0.1951	0.0522
	DRPEP	0.7800	1.0571	0.1961	0.0520
	LASSO	**0.8305**	**1.0369**	**0.1816***	**0.0492***
	SCAD	0.8243	0.9135	0.1838	0.0496
	MCP	0.8250	0.9196	0.1838	0.0496

Table 6: Average prediction accuracy measures in a 10-fold cross validation study for `GUSTO-I` dataset; Bold represents group maximum for AUC, for CS closest to one, and group minimum for LS and Brier score; * represents the best score among all methods.

simulation studies in Li and Clyde (2018) focus on settings involving fewer covariates ($p = 20$). In this setting, LCL and LCE behave quite well (see supplementary Section I). Our analyses suggest that these results are driven by a combination of sensitivity to the choice of hyperprior for δ and issues with the way `BAS` integrates over δ. Interestingly, the sensitivity to the hyperprior does not seem to be present for the LPEP procedures. We believe that this stability represents a key advantage of our method.

This paper focuses on developing the LPEP for logistic regression. However, the formulation can be easily extended to many other generalized linear models. Many of the computational advantages of our procedure extend to binomial, negative binomial and multinomial logic models where the data augmentation approach of Polson et al. (2013) can be readily applied. This is also true for probit models, where computation can rely on the data augmentation approach of Albert and Chib (1993), as well as for loglinear regression using the approach of Frühwirth-Schnatter et al. (2009).

Supplementary Material

Supporting Information for "Laplace Power-expected-posterior priors for logistic regression" (DOI: 10.1214/23-BA1389SUPP; .pdf). Supplementary materials include detailed proofs of Theorems 1, 2, 3 and 4, details of the MCMC algorithms for LPEP prior, and results from additional simulation studies and three additional real datasets. Code implementing our algorithms along with all real and simulated data sets are available

here. Code to replicate the results in the paper is available on Github.

References

Albert, A. and Anderson, J. A. (1984). "On the existence of maximum likelihood estimates in logistic regression models." *Biometrika*, 71(1): 1–10. MR0738319. doi: https://doi.org/10.1093/biomet/71.1.1. 2, 6

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88(422): 669–679. MR1224394. 19

Andrade, J. A. A. and O'Hagan, A. (2006). "Bayesian robustness modeling using regularly varying distributions." *Bayesian Analysis*, 1(1): 169–188. MR2227369. doi: https://doi.org/10.1214/06-BA106. 7

Andrade, J. A. A. and O'Hagan, A. (2011). "Bayesian robustness modelling of location and scale parameters." *Scandinavian Journal of Statistics*, 38(4): 691–711. MR2859745. doi: https://doi.org/10.1111/j.1467-9469.2011.00750.x. 7

Barber, R. F., Drton, M., and Tan, K. M. (2016). "Laplace approximation in high-dimensional Bayesian regression." In *Statistical Analysis for High-Dimensional Data*, 15–36. Springer. MR3616262. 8

Bayarri, M., Berger, J. O., Jang, W., Ray, S., Pericchi, L. R., and Visser, I. (2019). "Prior-based Bayesian information criterion." *Statistical Theory and Related Fields*, 3(1): 2–13. MR3963141. doi: https://doi.org/10.1080/24754269.2019.1582126. 1, 2

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of statistics*, 40(3): 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 1, 4, 7, 8

Berger, J. O. and Pericchi, L. R. (1996a). "The intrinsic Bayes factor for linear models." In J. M. Bernardo, A. P. D., J. O. Berger and Smith, A. F. M. (eds.), *Bayesian Statistics 5*, 25–44. Oxford Univ. Press. MR1425398. 1

Berger, J. O. and Pericchi, L. R. (1996b). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91(433): 109–122. 2

Berger, J. O., Pericchi, L. R., Ghosh, J., Samanta, T., and De Santis, F. (2001). "Objective Bayesian methods for model selection: Introduction and comparison." *Lecture Notes-Monograph Series*, 135–207. MR2000753. doi: https://doi.org/10.1214/lnms/1215540968. 1

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). "Lasso Meets Horseshoe." *Statistical Science*, 34(3): 405–427. MR4017521. doi: https://doi.org/10.1214/19-STS700. 2

Bové, D. S. and Held, L. (2011). "Hyper-g priors for generalized linear models." *Bayesian Analysis*, 6(3): 387–410. MR2843537. doi: https://doi.org/10.1214/ba/1339616469. 2, 5, 8

Breheny, P. and Huang, J. (2011). "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection." *Annals of Applied Statistics*, 5(1): 232–253. MR2810396. doi: https://doi.org/10.1214/10-AOAS388. 12

Califf, R. M., White, H. D., Van de Werf, F., Sadowski, Z., Armstrong, P. W., Vahanian, A., Simoons, M. L., Simes, R. J., Lee, K. L., and Topol, E. J. (1996). "One-year results from the Global Utilization of Streptokinase and TPA for Occluded Coronary Arteries (GUSTO-I) trial." *Circulation*, 94(6): 1233–1238. 17

Chen, J. and Chen, Z. (2012). "Extended BIC for small-n-large-P sparse GLM." *Statistica Sinica*, 555–574. MR2954352. doi: https://doi.org/10.5705/ss.2010.216. 9

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2000). "Power prior distributions for generalized linear models." *Journal of Statistical Planning and Inference*, 84(1-2): 121–137. MR1747500. doi: https://doi.org/10.1016/S0378-3758(99)00140-8. 3

Clyde, M. (2020). *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*. R package version 1.5.5. 12

Consonni, G., Fouskakis, D., Liseo, B., Ntzoufras, I., et al. (2018). "Prior distributions for objective Bayesian analysis." *Bayesian Analysis*, 13(2): 627–679. MR3807861. doi: https://doi.org/10.1214/18-BA1103. 1

De Santis, F. and Spezzaferri, F. (2001). "Consistent fractional Bayes factor for nested normal linear models." *Journal of statistical planning and inference*, 97(2): 305–321. MR1861156. doi: https://doi.org/10.1016/S0378-3758(00)00240-8. 1

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). "On Bayesian model and variable selection using MCMC." *Statistics and Computing*, 12(1): 27–36. MR1877577. doi: https://doi.org/10.1023/A:1013164120801. 2, 4

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, 96(456): 1348–1360. MR1946581. doi: https://doi.org/10.1198/016214501753382273. 12

Fouskakis, D., Ntzoufras, I., and Draper, D. (2015). "Power-expected-posterior priors for variable selection in Gaussian linear models." *Bayesian Analysis*, 10(1): 75–107. MR3420898. doi: https://doi.org/10.1214/14-BA887. 1, 3, 4

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2018). "Power-expected-posterior priors for generalized linear models." *Bayesian Analysis*, 13(3): 721–748. MR3807864. doi: https://doi.org/10.1214/17-BA1066. 2, 4, 5, 6, 12, 16, 18

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software*, 33(1): 1. MR1082147. 12

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009). "Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data." *Statistics and Computing*, 19(4): 479–492. MR2565319. doi: https://doi.org/10.1007/s11222-008-9109-4. 19

George, E. I. and McCulloch, R. E. (1997). "Approaches for Bayesian variable selection." *Statistica sinica*, 339–373. 10

Geweke, J. (1996). "Variable selection and model comparison in regression." *In Bayesian Statistics 5*. MR1425430. 1

Ghosh, J. (2019). "Cauchy and other shrinkage priors for logistic regression in the presence of separation." *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(6): e1478. MR4021448. doi: https://doi.org/10.1002/wics.1478. 2

Ghosh, J., Li, Y., and Mitra, R. (2018). "On the use of Cauchy prior distributions for Bayesian logistic regression." *Bayesian Analysis*, 13(2): 359–383. MR3780427. doi: https://doi.org/10.1214/17-BA1051. 7

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4): 711–732. MR1380810. doi: https://doi.org/10.1093/biomet/82.4.711. 2, 4

Heinze, G. and Schemper, M. (2002). "A solution to the problem of separation in logistic regression." *Statistics in medicine*, 21(16): 2409–2419. 2

Held, L., Bové, D. S., and Gravestock, I. (2015). "Approximate Bayesian model selection with the deviance statistic." *Statistical Science*, 242–257. MR3353106. doi: https://doi.org/10.1214/14-STS510. 17

Ibrahim, J. G. and Chen, M.-H. (2000). "Power prior distributions for regression models." *Statistical Science*, 15(1): 46–60. MR1842236. doi: https://doi.org/10.1214/ss/1009212673. 3

Ibrahim, J. G. and Laud, P. W. (1991). "On Bayesian analysis of generalized linear models using Jeffreys's prior." *Journal of the American Statistical Association*, 86(416): 981–986. MR1146346. 5

Johnson, V. E. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2): 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 1

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107(498): 649–660. MR2949346. doi: https://doi.org/10.1198/jasa.2011.ap10446. 1

Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the American Statistical Association*, 90(431): 928–934. MR1354008. 1, 4

Konis, K. (2007). "Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models." Ph.D. thesis, University of Oxford. 6

Kosmidis, I. and Schumacher, D. (2020). *detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates*. R package version 0.1. URL https://CRAN.R-project.org/package=detectseparation 6

Li, H. and Pati, D. (2017). "Variable selection using shrinkage priors." *Computational Statistics & Data Analysis*, 107: 107–119. MR3575062. doi: https://doi.org/10.1016/j.csda.2016.10.008. 2

Li, Y. and Clyde, M. A. (2018). "Mixtures of g-priors in generalized linear models." *Journal of the American Statistical Association*, 113(524): 1828–1845. MR3902249. doi: https://doi.org/10.1080/01621459.2018.1469992. 2, 10, 12, 13, 15, 17, 18, 19

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g-priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 1, 4, 7

Luo, S. and Chen, Z. (2013). "Selection consistency of EBIC for GLIM with non-canonical links and diverging number of parameters." *Statistics and its Interface*, 275–284. MR3066691. doi: https://doi.org/10.4310/SII.2013.v6.n2.a10. 9

Mansournia, M. A., Geroldinger, A., Greenland, S., and Heinze, G. (2018). "Separation in logistic regression: causes, consequences, and control." *American journal of epidemiology*, 187(4): 864–870. 15

McCullagh, P. and Nelder, J. A. (2019). *Generalized linear models*. Routledge. MR3223057. doi: https://doi.org/10.1007/978-1-4899-3242-6. 1

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 1

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 99–118. MR1325379. 1

Pérez, J. M. and Berger, J. O. (2002). "Expected-posterior prior distributions for model selection." *Biometrika*, 89(3): 491–512. MR1929158. doi: https://doi.org/10.1093/biomet/89.3.491. 2, 3, 5

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American Statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 2, 9, 10, 19

Porwal, A. and Rodríguez, A. (2023). "Supplementary Material for "Laplace Power-Expected-Posterior Priors for Logistic Regression"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1389SUPP. 6

Potter, D. M. (2005). "A permutation test for inference in logistic regression with small- and moderate-sized data sets." *Statistics in medicine*, 24(5): 693–708. MR2134534. doi: https://doi.org/10.1002/sim.1931. 15

Rossell, D., Abril, O., and Bhattacharya, A. (2021). "Approximate Laplace approximations for scalable model selection." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4): 853–879. MR4320004. 9

Schwarz, G. (1978). "Estimating the dimension of a model." *The annals of statistics*, 461–464. MR0468014. 5

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288. MR1379242. 12

Wedderburn, R. W. (1976). "On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models." *Biometrika*, 63(1): 27–32. MR0408092. doi: https://doi.org/10.1093/biomet/63.1.27. 9

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions." In Goel, P. K. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. Amsterdam: North-Holland/Elsevier. MR0881437. 1

Zellner, A. and Siow, A. (1980). "Posterior odds ratios for selected regression hypotheses." *Trabajos de estadística y de investigación operativa*, 31(1): 585–603. 1

Zhang, C.-H. (2010). "Nearly unbiased variable selection under minimax concave penalty." *The Annals of statistics*, 38(2): 894–942. MR2604701. doi: https://doi.org/10.1214/09-AOS729. 12

Acknowledgments

We would like to thank Dimitris Fouskakis and Ioannis Ntzourfras for sharing their code, and Merlise Clyde for answering our queries related to the `BAS` package.