University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
Main Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2019

The quantile probability model

Heyard, Rachel ; Held, Leonhard

Abstract: There is now a large literature on optimal predictive model selection. Bayesian methodology based on the -prior has been developed for the linear model where the median probability model (MPM) has certain optimality features. However, it is unclear if these properties also hold in the generalised linear model (GLM) framework, frequently used in clinical prediction models. In an application to the GUSTO-I trial based on logistic regression where the goal was the development of a clinical prediction model for 30-day mortality, sensitivity of the MPM with respect to commonly used prior choices on the model space and the regression coefficients was encountered. This makes a decision on a final model difficult. Therefore an extension of the MPM has been developed, the quantile probability model (QPM), that uses posterior inclusion probabilities to define a drastically reduced set of candidate models. Predictive model selection criteria are then applied to identify the model with best predictive performance. In the application the QPM turns out to be independent of the prior choices considered and gives better predictive performance than the MPM. In addition, a novel batching method is presented to efficiently estimate the Monte Carlo standard error of the predictive model selection criterion.

DOI: https://doi.org/10.1016/j.csda.2018.08.022

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-165882
Journal Article
Accepted Version

Originally published at:
Heyard, Rachel; Held, Leonhard (2019). The quantile probability model. Computational Statistics Data Analysis, 132:84-99.
DOI: https://doi.org/10.1016/j.csda.2018.08.022

The quantile probability model

Rachel Heyard and Leonhard Held

Department of Biostatistics at the Institute of Epidemiology, Biostatistics and Prevention, University of Zurich, Switzerland

Abstract

There is now a large literature on optimal predictive model selection. Bayesian methodology based on the g-prior has been developed for the linear model where the *median probability model* (MPM) has certain optimality features. However, it is unclear if these properties also hold in the generalised linear model (GLM) framework, frequently used in clinical prediction models. In an application to the GUSTO-I trial based on logistic regression where the goal was the development of a clinical prediction model for 30-day mortality, sensitivity of the MPM with respect to commonly used prior choices on the model space and the regression coefficients was encountered. This makes a decision on a final model difficult. Therefore an extension of the MPM has been developed, the *quantile probability model* (QPM), that uses posterior inclusion probabilities to define a drastically reduced set of candidate models. Predictive model selection criteria are then applied to identify the model with best predictive performance. In the application the QPM turns out to be independent of the prior choices considered and gives better predictive performance than the MPM. In addition, a novel batching method is presented to efficiently estimate the Monte Carlo standard error of the predictive model selection criterion.

Key words: Bayesian variable selection, inclusion probability, quantile probability model, median probability model, deviance information criterion, Monte Carlo standard error

*Corresponding author: R. Heyard (Hirschengraben 84, 8001 Zurich, rachel.heyard@uzh.ch, +41 44 634 49 73)

1. Introduction

Developing good prediction models by selecting the most relevant covariates is highly important in clinical research (Steyerberg, 2009). A frequently used Bayesian variable selection method, the median probability model (MPM) presented by Barbieri and Berger (2004), was proven to be optimal for continuous outcomes in the linear regression model. However, clinical prediction models often deal with binary outcomes and logistic regression where these optimality features may not hold. To obtain optimal prediction models in generalised linear models (GLMs) we generalise the MPM to the quantile probability model (QPM).

Standard Bayesian model and variable selection uses posterior model probabilities $\mathsf{Pr}(\mathcal{M}_j \mid \text{data})$ in order to select a single best model out of a set of different candidates $\mathcal{M}_j$, where $j \in \mathcal{J}$. If there are p potential variables, the selection is made among 2^p models and $\mathcal{J} = \{0, \dots 2^p - 1\}$. The posterior model probabilities depend on the Bayes factor (Kass and Raftery, 1995) which compares two models by taking the ratio of their marginal likelihoods. The marginal likelihood

$$f(\text{data} \mid \mathcal{M}_j) = \int f(\text{data} \mid \boldsymbol{\theta}_j, \mathcal{M}_j) f(\boldsymbol{\theta}_j \mid \mathcal{M}_j) \, \mathrm{d}\boldsymbol{\theta}_j,$$

depends on the prior distribution $f(\boldsymbol{\theta}_j \mid \mathcal{M}_j)$ of the model-specific parameter vector $\boldsymbol{\theta}_j$. Eliciting those prior distributions is a tedious task. Objective Bayesian methods unburden the statistician from choosing the parameter priors for all the models if no subjective prior information is available. Those methods have been well developed in the Gaussian linear model where Zellner's g-prior (Zellner, 1986) is usually chosen on the regression coefficients $\boldsymbol{\beta}_j$. This g-prior is defined as a multivariate normal distribution with mean zero and covariance matrix $g\sigma^2(\mathbf{X}_j^\top \mathbf{X}_j)^{-1}$, where g is a multiplicative factor, σ^2 is the residual variance and $\mathbf{X}_j$ is the design matrix (typically excluding the column for the intercept) of model $\mathcal{M}_j$. The multiplicative factor g can be estimated with an empirical Bayes (EB) (George and Foster, 2000) or a full Bayesian approach with a prior on g, such as the hyper-g, hyper-g/n (Liang et al., 2008) or Zellner-Siow (ZS) prior (Zellner and Siow, 1980).

The posterior model probabilities also depend on the prior model probabilities $\mathsf{Pr}(\mathcal{M}_j)$. Different prior settings on the model space have been proposed. However, in

practice there is often sensitivity of the MPM with respect to such prior choices on the model space and on the regression coefficients. We therefore propose the QPM, which is less sensitive to these prior settings.

This paper is structured as follows: Section 2 first describes how objective Bayesian variable selection methodology has been extended to generalised linear models. Then, Section 3 defines the median probability model, reviews different objective prior settings on the model and parameter priors in Section 3.1, and illustrates in a case study described in Section 3.2, that the MPM is sensitive to these prior choices. This motivates the need for a generalisation which we introduce in Section 4.1 as the quantile probability model. Section 4.2 presents a novel, computationally fast approach to efficiently compute the Monte Carlo standard error of the deviance information criterion (DIC), which is used to determine the QPM. Returning to our case study, we show in Section 4.3 that the QPM is independent of prior choices. Alternative information criteria, such as the Watanabe-Akaike information criterion (WAIC) or the leave-one-out cross-validation information criterion (LOO IC) are considered in Section 4.4 and lead, in our application, to the very same QPM. In Section 4.5 we discuss a potential quantile probability model average and close with some discussion in Section 5.

2. Objective Bayesian variable selection in generalised linear models

Consider a generalised linear model $\mathcal{M}_j$ with linear predictor $\eta_{ij} = \alpha + \mathbf{x}_{ij}^\top \boldsymbol{\beta}_j$ for observation $i \in \{1, \ldots, n\}$. The intercept α, the regression coefficient vector $\boldsymbol{\beta}_j$ and possible additional parameters are collected in $\boldsymbol{\theta}_j$. The extension of the standard g-prior on the regression coefficients in the linear model is the generalised g-prior (Sabanés Bové and Held, 2011)

$$\boldsymbol{\beta}_j \mid \mathcal{M}_j \sim \mathrm{N}_{p_j}\left(\mathbf{0}, gc\left(\mathbf{X}_j^\top \mathbf{W} \mathbf{X}_j\right)^{-1}\right),$$

where p_j is the number of variables in $\mathcal{M}_j$, c is a constant depending on the variance function, the response function $h(\cdot)$ and the intercept, $\mathbf{W}$ is a diagonal matrix with the weights of the observations and g is a multiplicative factor. Specifically, in logistic regression used in the application in Section 3.2 and 4.3, the response variable $Y_i \sim$

$\text{Bin}(1, \pi_i)$ is binary. The probabilities π_i are linked to the linear predictor η_i through the response function $h(\cdot)$; the expit function $\pi_i = h(\eta_i) = \exp(\eta_i)/(1 + \exp(\eta_i))$.

To calculate the posterior model probabilities, the marginal likelihood of each model needs to be computed numerically which slows down computation. A solution to this computational problem has been proposed by Held et al. (2015) using test-based Bayes factors (TBF) based on the deviance statistic (Johnson, 2008; Hu and Johnson, 2009). Let z_j be the deviance statistic of model $\mathcal{M}_j$ with d_j degrees of freedom, then the TBF of model $\mathcal{M}_j$ *versus* the null model $\mathcal{M}_0$ can be written in a closed form (Johnson, 2008):

$$\text{TBF}_{j,0} = (g+1)^{-d_j/2} \exp\left(\frac{g}{g+1}\frac{z_j}{2}\right). \tag{1}$$

Again, g needs to be estimated, empirically or using a full Bayesian approach. In order to calculate the posterior model probabilities

$$\mathsf{Pr}(\mathcal{M}_j \mid \text{data}) = \frac{\text{TBF}_{j,0}\,\mathsf{Pr}(\mathcal{M}_j)}{\sum_{k\in\mathcal{J}} \text{TBF}_{k,0}\,\mathsf{Pr}(\mathcal{M}_k)}, \text{ for } j \in \mathcal{J}, \tag{2}$$

prior probabilities $\mathsf{Pr}(\mathcal{M}_j)$ need to be defined on all the models in the model space $\mathcal{J}$. Possible prior choices will be discussed later in Section 3.1. The TBF-methodology is implemented in the `R`-package `glmBfp` available on `R-CRAN` (Held et al., 2015).

3. The median probability model for the generalised linear model

In a scenario, where a single model needs to be selected, the maximum a posteriori (MAP) model is commonly considered to be the best choice. It is defined as the model whose posterior probability in equation (2) is the highest. However, the MAP model has not necessarily the best prediction performance and the median probability model has been proposed by Barbieri and Berger (2004) as an alternative. This MPM includes all the variables with posterior inclusion probability (PIP) higher or equal to 0.5. The PIP of variable $x_k,\ k \in \{1, ..., p\}$, is defined as

$$\mathsf{Pr}(x_k \text{ included} \mid \text{data}) = \sum_{j\in\mathcal{J}} \mathsf{Pr}(\mathcal{M}_j \mid \text{data})\ \mathbb{1}_{[x_k\in\mathcal{M}_j]}. \tag{3}$$

Barbieri and Berger showed that the MPM is, under certain conditions such as the orthogonality of predictors, the optimal prediction model when selecting among normal

linear models. The MPM is frequently used in practice, for example in Ding et al. (2014), Ghosh (2015), Piironen and Vehtari (2016) and Held et al. (2016). However Barbieri and Berger's optimality theory for the MPM relies on quite strong conditions which will often not apply. In particular, the theory is considered only for the normal linear model and it is unclear whether the optimality also applies to GLMs.

3.1. Prior choices

Bayesian variable selection methods require prior choices. In the absence of subjective prior information, several "objective" prior choices have been proposed (Held et al., 2015). Specifically, a prior needs to be defined on the model space and, given that we use the generalised g-prior on the regression coefficients, we need to decide further how the multiplicative factor g in (1) is selected. The latter can, for example, be estimated using local empirical Bayes (LEB) by choosing g maximising equation (1) for a particular model $\mathcal{M}_j$, which leads to

$$\hat{g}_{\text{LEB}} = \max\left\{z_j/d_j - 1, 0\right\}.$$

This is a local approach because the prior parameter g is separately estimated for each model. For a full Bayesian approach, we can set a hyper prior on g such as the hyper-g/n prior

$$\frac{g/n}{g/n+1} \sim \text{U}(0,1),$$

the Zellner-Siow (ZS) prior

$$g \sim \text{IG}\left(\alpha = \frac{1}{2}, \beta = \frac{n}{2}\right),$$

or the ZS adapted prior:

$$g \sim \text{IncIG}\left(\alpha = \frac{1}{2}, \beta = \frac{n+3}{2}\right),$$

where U is the uniform, IG is the inverse-gamma and IncIG is the incomplete inverse-gamma distribution. Of course, there are other possibilities to define g, but we will restrict ourselves to the ones introduced above.

A commonly used model prior $\mathsf{Pr}(\mathcal{M}_j)$ for variable selection uses independent and identical Bernoulli priors for the inclusion indicators $\gamma_{jk} \sim \text{B}(q)$ of variable x_k in

model $\mathcal{M}_j$:

$$\mathsf{Pr}(\mathcal{M}_j \mid q) = \prod_{k=1}^{p} q^{\gamma_{jk}} (1-q)^{(1-\gamma_{jk})},$$

where p is the number of potential variables. If we fix the prior inclusion probability to $q = 1/2$, we obtain a uniform prior on the model space with $\mathsf{Pr}(\mathcal{M}_j) = 2^{-p}$ for each model $\mathcal{M}_j$. This prior assures that each candidate has the same *a priori* chance of being selected as the final model. The number of predictors included then follows *a priori* a binomial $\mathrm{Bin}(p, q)$ distribution with expectation $p \cdot q$.

A more general prior specifies a beta distribution on the prior inclusion probability $q \sim \mathrm{Be}(a, b)$ with $\mathsf{E}(q) = a/(a+b)$, $\mathrm{Var}(q) = ab/\{(a+b+1)(a+b)^2\}$ and density $f(q \mid a, b) = \mathrm{B}(a, b)^{-1} q^{a-1}(1-q)^{b-1}$, here $\mathrm{B}(x, y)$ denotes the beta function. Thus, if p_j is the number of predictors in $\mathcal{M}_j$, we can derive the prior model probability as a function of the hyperparameters a and b:

$$\begin{aligned}
\mathsf{Pr}(\mathcal{M}_j \mid a, b) &= \int_0^1 \mathsf{Pr}(\mathcal{M}_j \mid q)\, f(q \mid a, b)\, \mathrm{d}q \\
&= \int_0^1 q^{p_j} (1-q)^{p-p_j} \frac{1}{\mathrm{B}(a, b)} q^{a-1}(1-q)^{b-1}\, \mathrm{d}q \\
&= \frac{1}{\mathrm{B}(a, b)} \int_0^1 q^{a+p_j-1}(1-q)^{b+p-p_j-1}\, \mathrm{d}q \\
&= \frac{\mathrm{B}(a+p_j, b+p-p_j)}{\mathrm{B}(a, b)}. \qquad (4)
\end{aligned}$$

According to equation (4) changes in a and b affect the prior on the model space and in the same time the posterior inclusion probabilities and consequently influence which variables are included in the MPM. With $a = 1$ and $b = 1$ we obtain the multiplicity-corrected model prior (Scott and Berger, 2010):

$$\mathsf{Pr}(\mathcal{M}_j \mid a = 1, b = 1) = \frac{1}{p+1} \left(\begin{array}{c} p \\ p_j \end{array} \right)^{-1}. \qquad (5)$$

The prior probability (5) depends on the number of variables p_j included in model $\mathcal{M}_j$ and the prior distribution of the number of included predictors is discrete uniform on $\{0, 1, \ldots, p\}$. The next section will illustrate how the selection of variables in the MPM is affected by all the prior choices introduced above.

3.2. Case Study

In our case study we reanalyse data from the GUSTO-I trial, previously analysed in Ennis et al. (1998), Steyerberg (2009), Held et al. (2015) and Li and Clyde (2016). This is a large randomised study comparing four different treatments in over 40,000 acute myocardial infarction patients (Lee et al., 1995). More specifically we will use a publicly available subgroup from the Western region of the USA with $n = 2{,}188$ patients and focus on the binary endpoint 30-day mortality (Steyerberg, 2009). The variable selection techniques which will be presented in this paper should select among the 17 covariates $x_1, \ldots, x_{17}$ in Table 1 to predict the outcome y using logistic regression.

Variable	Description
y	Death within 30 days after acute myocardial infarction (Yes = 1, No = 0)
x_1	Gender (Female = 1, Male = 0)
x_2	Age (years)
x_3	Killip class (4 categories)
x_4	Diabetes (Yes = 1, No = 0)
x_5	Hypotension (Yes = 1, No = 0)
x_6	Tachycardia (Yes = 1, No = 0)
x_7	Anterior infarct location (Yes = 1, No = 0)
x_8	Previous myocardial infarction (Yes = 1, No = 0)
x_9	Height (cm)
x_{10}	Weight (kg)
x_{11}	Hypertension history (Yes = 1, No = 0)
x_{12}	Smoking (3 categories: Never / Ex / Current)
x_{13}	Hypercholesterolaemia (Yes = 1, No = 0)
x_{14}	Previous angina pectoris (Yes = 1, No = 0)
x_{15}	Family history of myocardial infarctions (Yes = 1, No = 0)
x_{16}	ST elevation on ECG: Number of leads (0-11)
x_{17}	Time to relief of chest pain more than 1 hour (Yes = 1, No = 0)

Table 1: Definition of the variables in the GUSTO-I trial data.

In the application on the GUSTO-I trial data all $p = 17$ potential variables are considered so that a total of $2^{17} = 131{,}072$ candidate models can be defined, depending on the set of variables included. We use multiplicity-corrected and uniform priors on the model space. Both model priors are available in the `glmBfp` package in `R` which will be used for the analysis. To estimate g we use local empirical Bayes (LEB), the hyper-g/n and the Zellner-Siow (ZS) adapted priors, also available in the `glmBfp` package. In total we thus consider six different combinations of model and parameter priors. For each combination we compute the inclusion probabilities exactly through exhaustive computation of the posterior probabilities of all models considered.

For most prior combinations, the MAP model includes seven variables, $x_2, x_3, x_5, x_6, x_8, x_{10}, x_{16}$. Only for the ZS adapted method combined with a multiplicity-corrected model prior, the model with five variables $x_2, x_3, x_5, x_6, x_{16}$ has highest posterior probability. Figure 1 now gives the PIPs of all potential variables in an increasing order depending on the prior on the model space (in columns) and on the estimation method for g (in rows). This figure illustrates how the MPM depends on these prior choices. If we choose to estimate g using LEB, the choice of the prior model probability does not influence the selection of variables included in the MPM; the MPM always contains the eight most probable variables. The same is true using a hyper-g/n prior on g, even though the ordering of the variables by their inclusion probabilities changes slightly from multiplicity-corrected to uniform model prior (see x_1 and x_{10}). However, if we use a ZS adapted prior on g, the MPM drops two more variables. Furthermore, for a uniform model prior and a hyper-g/n approach for g, the PIPs of the variables x_1 and x_{10} get very close to the MPM variable inclusion threshold of 0.5 questioning whether this sharp cut-off is really appropriate.

It may also be of interest to study sensitivity of the MPM with respect to the hyperparameters a and b in equation (4). Suppose the posterior model probabilities $\mathsf{Pr}(\mathcal{M}_j \mid a, b)$ are already computed for specific values a and b, for example for the multiplicity-corrected model prior with $a = b = 1$. In order to compute the PIPs for different hyperparameters a^* and b^* we only need to calculate the new prior model

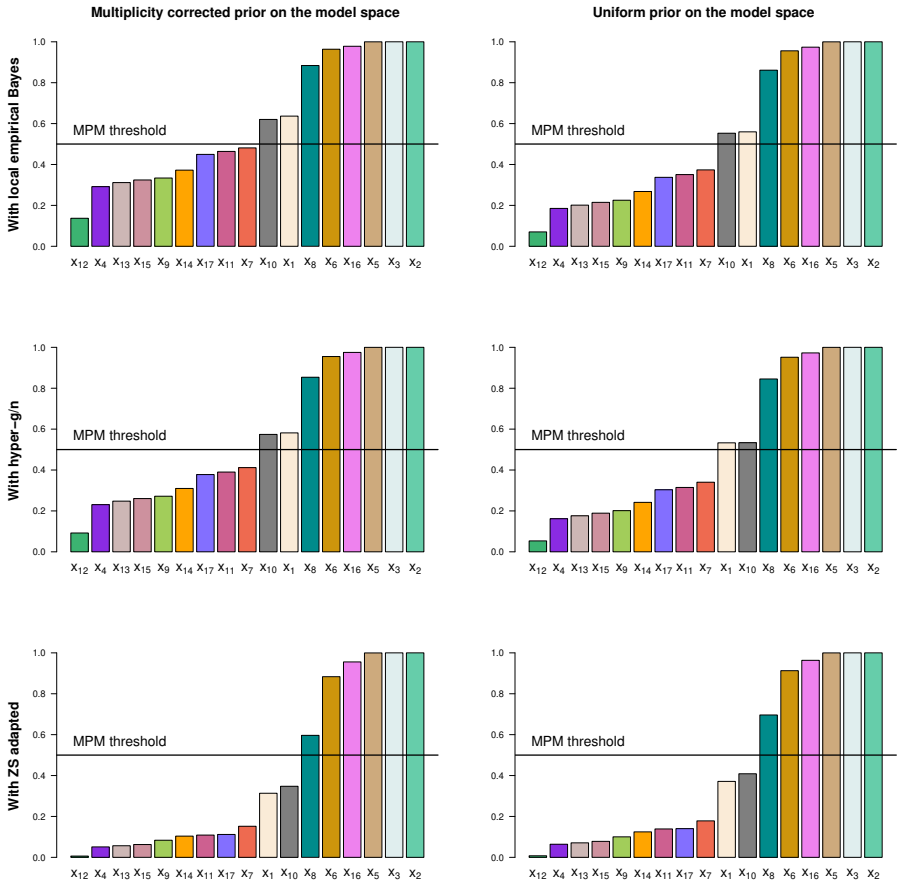

Figure 1: The posterior inclusion probabilities of the 17 variables for all prior combinations. A different color is attributed to each variable. In the two columns we choose between a multiplicity-corrected and a uniform model prior. The difference among the rows shows the influence of the estimation method for g on the inclusion probabilities and the final MPM. Using a ZS adapted prior on g leads to a MPM with 6 variables whereas the other prior choices lead to a larger MPM with 8 variables.

probabilities $\mathsf{Pr}(\mathcal{M}_j \mid a^*, b^*)$ with equation (4) and then use

$$\begin{aligned}
\mathsf{Pr}(\mathcal{M}_j \mid \text{data}, a^*, b^*) &\propto f(\text{data} \mid \mathcal{M}_j)\ \mathsf{Pr}(\mathcal{M}_j \mid a^*, b^*) \\
&= f(\text{data} \mid \mathcal{M}_j)\ \mathsf{Pr}(\mathcal{M}_j \mid a, b) \times \frac{\mathsf{Pr}(\mathcal{M}_j \mid a^*, b^*)}{\mathsf{Pr}(\mathcal{M}_j \mid a, b)} \\
&\propto \mathsf{Pr}(\mathcal{M}_j \mid \text{data}, a, b)\ \frac{\mathsf{Pr}(\mathcal{M}_j \mid a^*, b^*)}{\mathsf{Pr}(\mathcal{M}_j \mid a, b)},
\end{aligned}$$

with subsequent normalisation. Finally, to find the inclusion probability of a particular

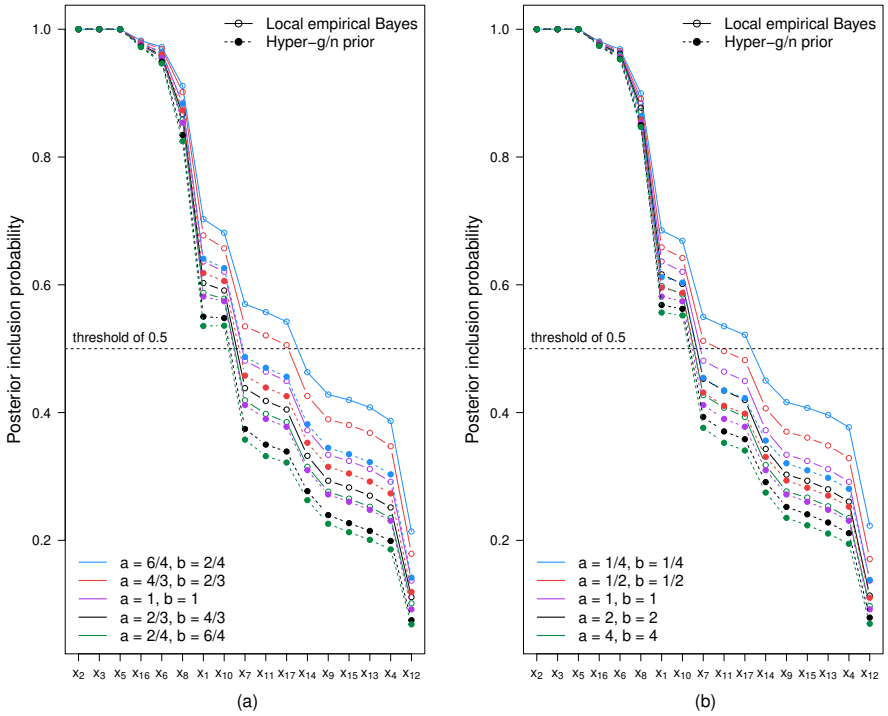

Figure 2: Posterior inclusion probabilities depending on the hyperparameters a and b of the beta prior on the inclusion probability. The influence of the choice of the hyperparameters a and b on the MPM is shown for the LEB method and the hyper-g/n prior.

variable x_k under the new prior with a^* and b^*, we use equation (3).

Figure 2 demonstrates how the PIPs in the application change by just varying the hyperparameters a and b of the beta distribution on the prior inclusion probability. In subfigure (a) we choose a and b such that the prior sample size $a + b$ of the beta distribution is kept fixed to 2. The prior sample size informs us about the weight attached to the prior distribution. In subfigure (b) the expected prior inclusion probability $a/(a+b)$ is fixed to 0.5 by selecting $a = b$. Selecting equal hyperparameters is of interest because the MPM inclusion threshold of 0.5 could be interpreted as selecting the variables whose posterior inclusion probability is larger than their prior expectation, suggesting the data strengthens the evidence that a specific variable belongs in the model. This

justification of the MPM suggests that the inclusion threshold should change with the prior inclusion probability. This is however not discussed in the literature. The generalisation of the MPM presented in the next section tries to fill this gap.

Figure 2 shows that the MPM is very sensitive to changes in the hyperparameters. The MPM with LEB and hyperparameters $a = 6/4$ and $b = 2/4$ would include eleven variables whereas other prior combinations would only select eight variables. This sensitivity of the MPM is also illustrated in a short simulation study on logistic regression (see Appendix A). The fact that the ordering of the variables by their inclusion probabilities is quite stable further motivates a generalisation of the MPM that enables us to find the best predictive model based on a different threshold than 0.5. We call this generalisation the *quantile probability model*.

4. Methodology

4.1. The quantile probability model

Without loss of generality, suppose the PIPs are sorted and further assume that they are mutually distinct, so that $\pi_1 < \cdots < \pi_p < 1$. Now, instead of setting a fixed threshold of inclusion such as 0.5 for the MPM, we vary the threshold π_{T} in the set of the observed PIPs, so $\pi_{\text{T}} \in \{\pi_1, ..., \pi_p\}$. At each π_{T} we include the variables having a PIP higher or equal to the inclusion threshold, so the larger π_{T} the simpler the candidate model. We also include the null model (with threshold $\pi_{\text{T}} = 1$) and thus obtain $p + 1$ candidate models.

The QPM is the candidate model that has best predictive performance. To assess the performance of these $p + 1$ candidate models we first use the deviance information criterion (DIC) (Spiegelhalter et al., 2002), but will consider alternative predictive criteria in Section 4.4. The DIC is defined as a classical estimate of fit plus twice the effective number of parameters p_D:

$$\text{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D = 2\bar{D} - D(\bar{\boldsymbol{\theta}}). \tag{6}$$

Here $\boldsymbol{\theta}$ is the the parameter vector of the model considered and $\bar{\boldsymbol{\theta}}$ its expectation, $D(\boldsymbol{\theta}) = -2\log\{f(\mathbf{y} \mid \boldsymbol{\theta})\} + 2\log\{f(\mathbf{y})\}$ is the 'Bayesian' deviance, with $f(\mathbf{y} \mid \boldsymbol{\theta})$

being the likelihood function and $f(\mathbf{y})$ being some standardising function of the data. Finally, $\bar{D}$ is the posterior expectation of the deviance and p_D is an estimate of the effective number of parameters in the model. Note that DIC is negatively oriented, so the QPM will be the candidate model with smallest DIC.

Barbieri and Berger (2004) proved that the MPM is optimal for prediction under the squared error loss when we want to select among normal linear models. Likewise, Spiegelhalter et al. (2002) state in their paper that differences in DIC are estimates of differences in expected loss in prediction which correspond to squared error loss in the linear model. Moreover, DIC can be seen as a Bayesian analogue of Akaike information criterion (AIC) (Akaike, 1973) which is best to find the optimal prediction model: AIC $= D(\hat{\boldsymbol{\theta}}) + 2p$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate. Especially for models with weak prior information, we have $\bar{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$ and $p_D \approx p$, hence DIC $\approx$ AIC (Lunn et al., 2012). Stone (1977) showed furthermore that a model comparison based on cross-validation is asymptotically equivalent to a model comparison based on AIC.

Since Barbieri and Berger use a predictive criterion, the squared error, to prove the optimality of the MPM, we therefore suggest to use in our search for the QPM the DIC which generalises the squared error loss to non-normal generalised linear models. The QPM can be seen as a generalisation of the MPM with an inclusion threshold not fixed at 0.5. Further, the DIC, as well as other predictive information criteria which will be presented later, can be expressed using the log predictive density (lpd), $\log f(\mathbf{y} \mid \boldsymbol{\theta})$:

$$\begin{aligned} \text{DIC} &= -2\,\text{elpd}_{\text{DIC}} \\ &= -2 \log f(\mathbf{y} \mid \widehat{\boldsymbol{\theta}}) + 2\text{p}_{\text{DIC}}, \end{aligned}$$

where elpd_{DIC} is the expected lpd, $\widehat{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$ and p_{DIC} is the effective number of parameters (Gelman et al., 2014). The authors also show that the lpd is proportional to the mean squared error if the model is normal, so that the QPM is equivalent to the MPM in the normal linear model framework. Moreover, the lpd is also often referred to as the log score which is the most commonly used proper scoring rule to assess predictive accuracy for probabilistic predictions. Note that the prior on the model space does not affect the DIC of a particular model.

4.2. Monte Carlo error for DIC

To calculate the posterior inclusion probabilities of the p variables we use an exhaustive search on the model space. So, no Monte Carlo (MC) error is created at this stage. However, to compute the DIC of the QPM candidates, we sample from the posterior distribution of the regression coefficients as outlined in Held et al. (2015). This procedure induces MC error.

Efficient calculation of the MC SE of DIC turns out to be rather difficult. Zhu and Carlin (2000) presented an approach based on the multivariate delta method. However, the authors report only "mixed results, with none emerging as sufficiently accurate to merit routine use". They rather suggest to use a "brute force" approach, by replicating the calculation of the DIC a large number N of iterations, resulting in a sequence of N DIC estimates $(\text{DIC}_i, i = 1, \ldots, N)$, with subsequent calculation of the sample standard deviation. This "brute force" approach is very time-consuming. In the following we present an approach that needs significantly less computation time.

Suppose our DIC calculation relies on a sequence $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_S\}$ of parameter vectors generated by MC simulations of size S. Our approach begins by breaking up Θ into a fixed number k of consecutive batches of equal length S/k. Then the DIC within each batch is computed, resulting in k DICs. The standard error between the k DICs is retrieved: $\text{SE}(\text{DIC}^{(k)}) = \left(\frac{1}{k-1}\sum_{i=1}^{k}(\text{DIC}_i - \overline{\text{DIC}})\right)^{1/2}$. These steps are repeated for different $k \in K$. We thus retain a sample of standard errors, $\{\text{SE}(\text{DIC}^{(k)})\}_{k\in K}$. We now make use of the 'square-root law' which states that the accuracy of an estimator is inversely proportional to the square root of the sample size, here S/k:

$$\text{SE}(\text{DIC}^{(k)}) \propto \sqrt{\frac{1}{S/k}} \propto \sqrt{k}, \text{ so that } \text{SE}(\text{DIC}^{(k)}) = c \cdot \sqrt{k}\,. \tag{7}$$

In order to find an estimate $\hat{c}$ for c, we use a weighted linear regression with outcome $\text{SE}(\text{DIC}^{(k)})$, explanatory variable $\sqrt{k}$ and weight equal to k [`lm(SE ~ -1 + sqrt(k), weight = k)` in `R`]. We finally obtain the MC SE of the DIC of interest based on the original sample size S from (7) with $k = 1$: $\text{SE}(\text{DIC}^{(1)}) = \hat{c}$.

This approach is inspired by the batch means method presented in Flegal et al.

(2008) which, however, does not use weighted linear regression to estimate the MC SE.

4.3. Case study revisited

Given the PIPs in Figure 1, we obtain 18 candidate models for each of the different prior combinations. The prior on the model space does not influence the model selection criterion, so that, if the ordering of the inclusion probabilities is independent of prior settings, we only present the results using a multiplicity-corrected model prior. This is the case for the LEB and ZS adapted methods. With a hyper-g/n approach, the ordering of the variables regarding their PIPs is slightly affected by the prior choice on the model space (see x_1 and x_{10} in Figure 1). This leads to one different candidate model, since x_1 is dropped before x_{10} with a uniform model prior. However, the conclusion for both prior model probabilities turned out to be the same so we report only the results using the multiplicity-corrected model prior.

Figure 3 shows the DIC of all QPM candidate models. The minimum DIC for each prior combination, and therefore the QPM, is reached once the nine variables $x_1, x_2, x_3, x_5, x_6, x_7, x_8, x_{10}, x_{16}$ are included in the model. The QPM is independent of the estimation method for g and the prior on the model space: for all combinations, the QPM turns out to include the same nine variables. This is different for the MPM, which depends on the priors being used, as demonstrated before.

In our application, we use an MC sample of size S =10,000 to calculate the DICs. To compute the MC SE we partition our MC samples into $k \in \{20, 50, 100, 200\}$ batches of equal size S/k and apply the methodology introduced above.

To investigate whether the selected QPM with nine variables does indeed have lowest DIC, we check whether the difference in DIC between two subsequent models, $\mathcal{M}_j$ and $\mathcal{M}_{j-1}$, $\forall j \in \{1, \ldots, 17\}$ is significantly different from zero. The MC SE of the difference $\Delta\text{DIC}_j = \text{DIC}_{j-1} - \text{DIC}_j$ is

$$\text{SE}(\Delta\text{DIC}_j) = \sqrt{\text{SE}(\text{DIC}_{j-1})^2 + \text{SE}(\text{DIC}_j)^2},$$

since DIC_j and DIC_{j-1} are independent. Table 2 shows these differences with their respective standard errors. We only present the results using the multiplicity-corrected

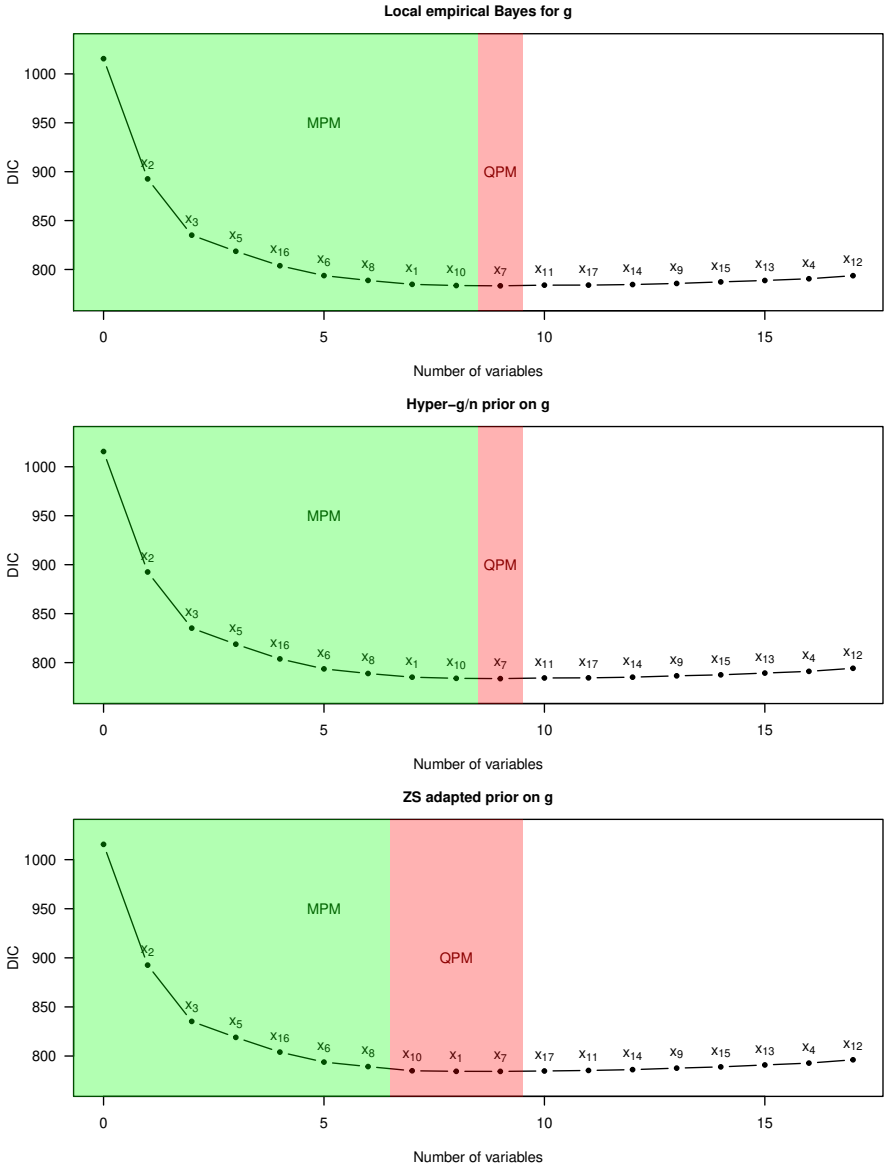

Figure 3: DIC as a function of the number of variables included for different prior choices on g. The prior on the model space is the multiplicity-corrected model prior. The variable name on top of the points refers to the variable that is added at this stage. The MPM is highlighted in green, whereas the QPM is defined by the red and the green shaded parts of the plots. The different MPMs include six to eight variables. All the QPMs include the same nine variables. The size of the posterior samples to calculate the DICs is 10,000.

	LEB		hyper-g/n		ZS adapted	
$\mathcal{M}_0$ vs. $\mathcal{M}_1$	122.952	(0.11)	123.025	(0.11)	123.007	(0.11)
$\mathcal{M}_1$ vs. $\mathcal{M}_2$	57.530	(0.16)	57.335	(0.16)	57.298	(0.16)
$\mathcal{M}_2$ vs. $\mathcal{M}_3$	16.480	(0.19)	16.396	(0.18)	16.277	(0.18)
$\mathcal{M}_3$ vs. $\mathcal{M}_4$	14.783	(0.20)	15.007	(0.19)	14.988	(0.19)
$\mathcal{M}_4$ vs. $\mathcal{M}_5$	10.085	(0.21)	10.200	(0.21)	10.176	(0.21)
$\mathcal{M}_5$ vs. $\mathcal{M}_6$	4.924	(0.21)	4.702	(0.21)	4.572	(0.22)
$\mathcal{M}_6$ vs. $\mathcal{M}_7$	4.016	(0.21)	3.770	(0.22)	4.270	(0.23)
$\mathcal{M}_7$ vs. $\mathcal{M}_8$	1.226	(0.21)	1.173	(0.22)	0.560	(0.23)
$\mathcal{M}_8$ vs. $\mathcal{M}_9$	0.285	(0.20)	0.295	(0.22)	0.084	(0.23)
$\mathcal{M}_9$ vs. $\mathcal{M}_{10}$	-0.693	(0.22)	-0.667	(0.23)	-0.417	(0.24)
$\mathcal{M}_{10}$ vs. $\mathcal{M}_{11}$	-0.042	(0.23)	-0.118	(0.23)	-0.542	(0.24)
$\mathcal{M}_{11}$ vs. $\mathcal{M}_{12}$	-0.575	(0.23)	-0.695	(0.23)	-0.838	(0.25)
$\mathcal{M}_{12}$ vs. $\mathcal{M}_{13}$	-1.111	(0.22)	-1.363	(0.24)	-1.533	(0.25)
$\mathcal{M}_{13}$ vs. $\mathcal{M}_{14}$	-1.578	(0.22)	-1.028	(0.24)	-1.273	(0.25)
$\mathcal{M}_{14}$ vs. $\mathcal{M}_{15}$	-1.462	(0.23)	-1.797	(0.24)	-1.964	(0.26)
$\mathcal{M}_{15}$ vs. $\mathcal{M}_{16}$	-1.768	(0.23)	-1.746	(0.24)	-1.891	(0.25)
$\mathcal{M}_{16}$ vs. $\mathcal{M}_{17}$	-3.141	(0.23)	-3.130	(0.25)	-3.372	(0.27)

Table 2: Differences in DIC between subsequent QPM candidate models and their Monte Carlo (MC) standard error in parentheses for different method to estimate g. A multiplicity-corrected prior is used on the model space. The DICs are calculated with an MC sample size of 10,000.

model prior, since the results for the uniform prior are more or less equal. The model prior does not affect the DIC. Two questions emerge: First, is the QPM model clearly defined? Most of the ΔDIC are large relative to their MC SE, but others like $\mathcal{M}_8$ vs. the QPM model $\mathcal{M}_9$ and $\mathcal{M}_9$ vs. $\mathcal{M}_{10}$ are quite small relative to their MC SE. Secondly, is there evidence that the selected QPM model predicts better than the corresponding MPM model ($\mathcal{M}_8$ for LEB and hyper-g/n and $\mathcal{M}_6$ for ZS adapted)? This is clearly the case for the ZS adapted prior, whereas the comparison of $\mathcal{M}_8$ and $\mathcal{M}_9$ for LEB and hyper-g/n gives rather large MC standard errors compared to the observed differences.

To reduce the MC SE and to guarantee that the QPM is clearly defined, we have recomputed DIC for the three models $\mathcal{M}_8$, $\mathcal{M}_9$ and $\mathcal{M}_{10}$ with a larger sample size $S =$ 500,000 and updated K accordingly to ensure that the larger sample is still partitioned into batches of equal length: $K = \{80, 100, 125, 160, 200\}$. Table 3 shows that all MC SE are now sufficiently small, perhaps except for the DIC difference between $\mathcal{M}_8$ and $\mathcal{M}_9$ under the ZS adapted prior. In particular, we can now conclude in all three cases that the QPM model has better prediction performance than the corresponding MPM model.

	LEB		hyper-g/n		ZS adapted	
$\mathcal{M}_8$ vs. $\mathcal{M}_9$	0.217	(0.031)	0.166	(0.034)	0.057	(0.033)
$\mathcal{M}_9$ vs. $\mathcal{M}_{10}$	-0.295	(0.031)	-0.342	(0.034)	-0.170	(0.032)

Table 3: Differences in DIC between the QPM $\mathcal{M}_9$ and its neighboring models $\mathcal{M}_8$ and $\mathcal{M}_{10}$ as well as the Monte Carlo (MC) standard error of these differences in parentheses. The DICs are calculated with an MC sample size of 500,000.

The posterior model probabilities (PMP) of the MPM and QPM may also be of interest. If Barbieri and Berger's results also apply in GLMs, then the MPM should be close to the MAP model. In contrast, the QPM model may be less likely *a priori*, as it is selected based on predictive performance. Figure 4 gives the cumulative PMP of the 100 best ranked models for the six different prior combinations. The top 100 models have already around 30-50% posterior probability combined, for the ZS adapted prior even close to 80% posterior mass. The QPM $\mathcal{M}_9$ as well as the different MPMs with eight or six variables are always among the top 100 models for all possible prior combinations. It is interesting that the model with highest PMP, the MAP model, is not always among the QPM candidate models. This is the case for three out of the six prior combinations. We comment on this further in the discussion.

Figure 5 shows how the rankings of the QPM and the MPM are influenced by changing the hyperparameters a and b of the prior inclusion probability. The same hyperparameter choices are used as in Figure 2. The rank of the MPM is somewhat dependent on the hyperparameters and is particularly high (around 40) under LEB for

some specific prior settings on a and b. In contrast, the QPM always ranks more or less at the same position for LEB and hyper-g/n (between 10 and 20). The rank of the QPM model under the ZS adapted prior is, however, considerably larger than the rank of the MPM model. This reflects the fact that the search for the QPM model is based on the inclusion probabilities, but does not directly take the actual posterior model probabilities into account. We note that for LEB combined with $a = 1/2$ and $b = 1/2$ the QPM is identical to the MPM.

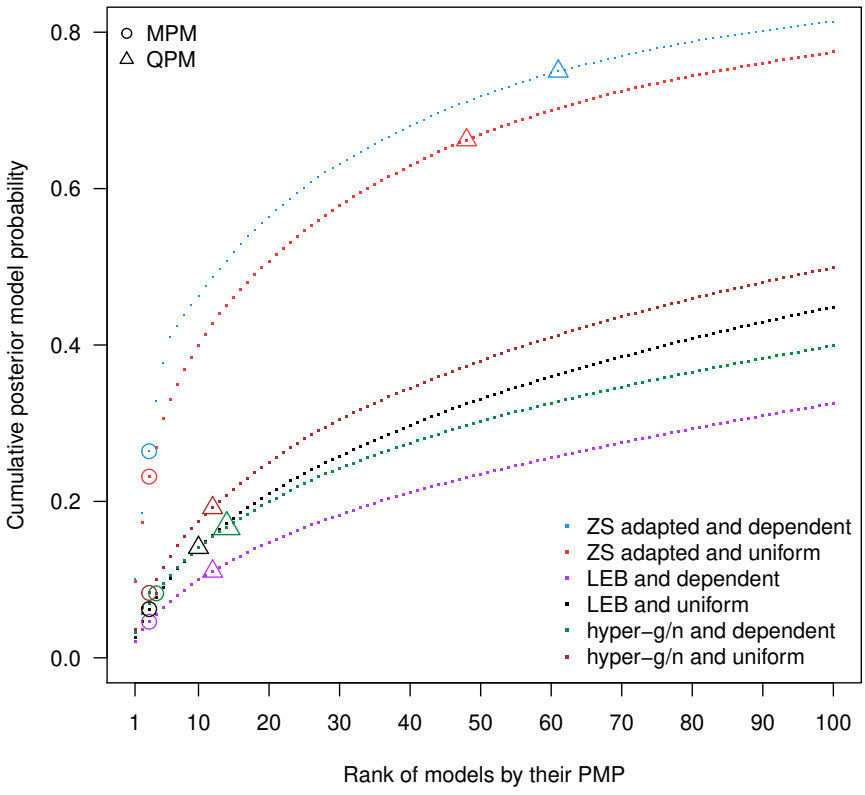

Figure 4: Cumulative posterior model probabilities of the top 100 models for different prior combinations. The QPM and the MPM are highlighted. The MPM has larger PMP (and smaller rank) than the QPM.

In order to compare the MAP model to the QPM, we computed the difference in

DIC between both models as well as the corresponding MC SE. Table 4 shows these quantities for the prior choices considered. The DIC of the QPM is always significantly lower than the DIC of the MAP, especially with the ZS adapted prior on g.

	LEB		hyper-g/n		ZS adapted	
MAP vs. QPM	0.853	(0.21)	0.745	(0.22)	9.386	(0.21)

Table 4: Difference in DIC between the MAP model and the selected QPM as well as its Monte Carlo (MC) standard error in parentheses for different estimation methods for g. A multiplicity-corrected prior is used on the model space. The DICs are calculated with an MC sample size of 10'000.

4.4. Alternative information criteria

Instead of DIC, the Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010) or the leave-one-out cross-validation information criterion (LOO IC) (Vehtari et al., 2017) can be used to identify the QPM among the $p + 1$ models of interest.

Vehtari et al. (2017) present an efficient approach to evaluate models using LOO IC and WAIC. They also discuss their `R`-package `loo` that has been implemented to evaluate models fitted with STAN but whose `waic()` and `loo()` functions can easily be used to evaluate other models. The QPM candidate minimising WAIC as well as LOO IC for all prior choice combinations is again model $\mathcal{M}_9$, including the nine most relevant variables according to their PIPs. Table 5 shows the difference in these alternative information criteria for subsequent candidate models and their respective MC SE for S =10,000. The MC SE for WAIC and LOO IC can be computed as described in section 4.2 for the DIC. Again, most of the differences are large compared to the corresponding standard errors and we still retrieve the same QPM with nine variables as with DIC. However, the differences around $\mathcal{M}_9$ ($\mathcal{M}_8$ vs. $\mathcal{M}_9$ and $\mathcal{M}_9$ vs. $\mathcal{M}_{10}$) are quite small relative to their standard errors, so we used a larger MC sample size (S =500,000), see Table 6 for the differences in WAIC. There is still uncertainty whether $\mathcal{M}_8$ or $\mathcal{M}_9$ is the QPM model for the hyper-g/n prior and the ZS adapted prior whereas for LEB the difference in WAIC is nearly three times as large as the corresponding MC SE, so $\mathcal{M}_9$ can again be identified as the QPM model for LEB. In

principle we could decrease the MC SE further by increasing S but the `loo` package ran into memory problems in the calculation of LOO IC already for S =500,000.

4.5. A quantile probability model average

In order to further improve the prediction performance of our model we finally considered a quantile probability model average (QPMA): a model average of the candidate models $\mathcal{M}_0$ to $\mathcal{M}_9$ with the QPM $\mathcal{M}_9$ being the most complex one. With this approach we hoped to improve the predictions because the model average will imply covariate-specific shrinkage with more shrinkage towards zero for less relevant variables. In the model average we used DIC-based weights $w_0, \ldots, w_9$, defined in analogy to AIC-based weights (Buckland et al., 1997; Wagenmakers and Farrell, 2004):

$$w_j = \frac{\exp(-\frac{1}{2}\Delta\text{DIC}_j)}{\sum_{k=0}^{9}\exp(-\frac{1}{2}\Delta\text{DIC}_k)},$$

where $\Delta\text{DIC}_j = \text{DIC}_j - \min(\text{DIC}_0, \ldots \text{DIC}_9)$. To sample from the QPMA we can simply sample from the reduced candidate model space $\{\mathcal{M}_0, \ldots, \mathcal{M}_9\}$ with respective weights $\{w_0, \ldots, w_9\}$. We computed the DIC, WAIC and LOO IC based on the QPMA. However, the prediction performance of the QPMA was slightly worse than the one of the QPM. This may be due to the small number of models in the model average. A full model average of all the 2^9 models with a subset of the nine QPM predictors might improve the predictive performance, but will also increase the computational burden considerably.

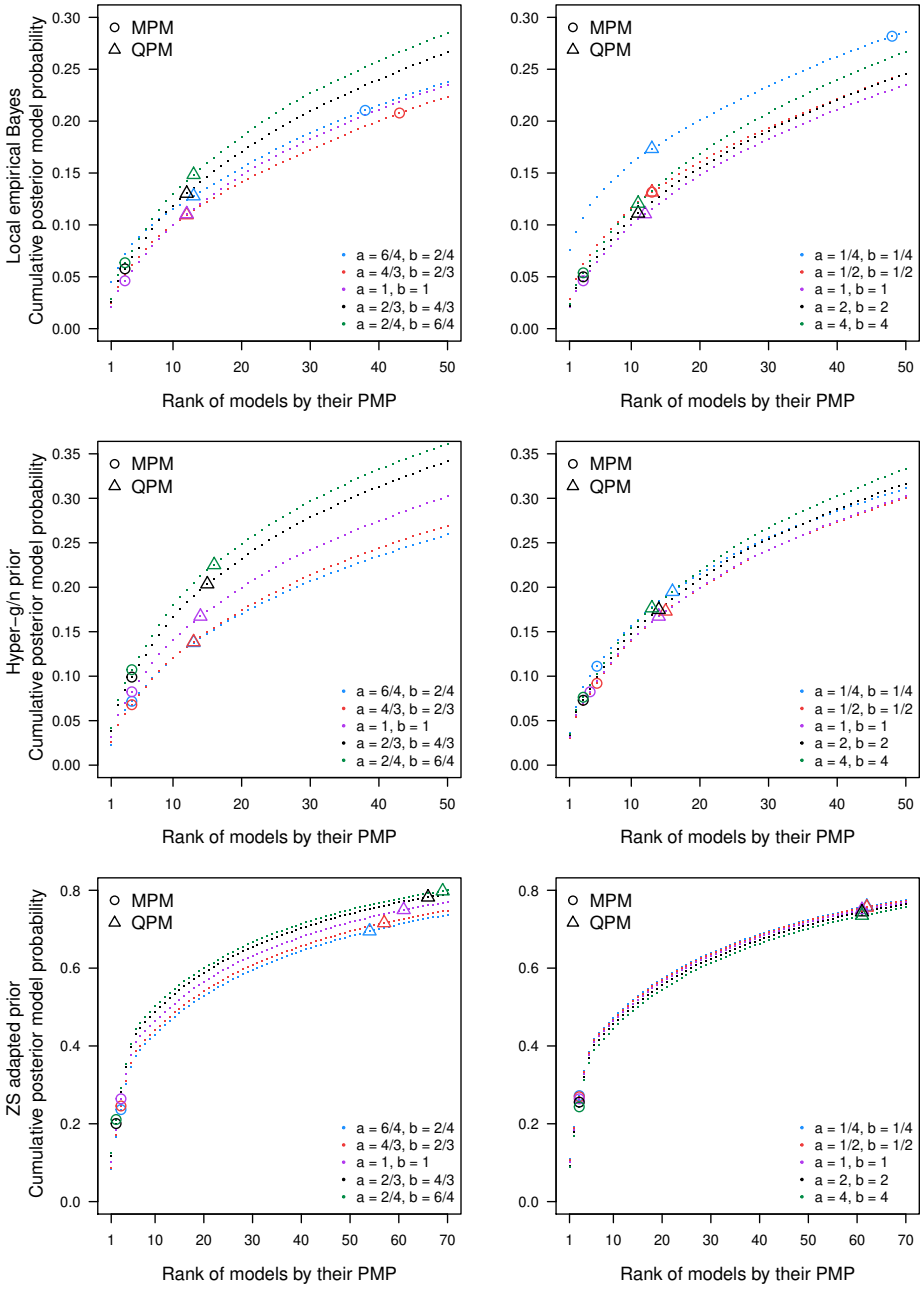

Figure 5: Cumulative posterior model probabilities depending on the estimation method for g (LEB, hyper-g/n or ZS-adapted) as well as on the choice of the hyperparameters a and b. The QPM and the MPM for each prior combination are marked. The ranks of the QPM are between 10 and 20 for LEB and hyper-g/n but considerably larger under the ZS adapted prior. The MPM has smaller ranks, with the exception of three hyperparameter settings under LEB.

	LEB				hyper-g/n				ZS adapted			
	WAIC		LOO		WAIC		LOO		WAIC		LOO	
$\mathcal{M}_0$ vs. $\mathcal{M}_1$	122.755	(0.11)	122.750	(0.11)	122.840	(0.12)	122.835	(0.11)	122.832	(0.12)	122.826	(0.11)
$\mathcal{M}_1$ vs. $\mathcal{M}_2$	56.055	(0.18)	56.036	(0.17)	55.820	(0.17)	55.794	(0.17)	55.782	(0.17)	55.756	(0.17)
$\mathcal{M}_2$ vs. $\mathcal{M}_3$	16.085	(0.21)	16.084	(0.20)	16.006	(0.20)	16.003	(0.19)	15.902	(0.20)	15.899	(0.19)
$\mathcal{M}_3$ vs. $\mathcal{M}_4$	14.625	(0.22)	14.632	(0.22)	14.938	(0.21)	14.948	(0.21)	14.959	(0.21)	14.970	(0.21)
$\mathcal{M}_4$ vs. $\mathcal{M}_5$	10.339	(0.23)	10.323	(0.22)	10.407	(0.23)	10.398	(0.23)	10.407	(0.23)	10.396	(0.23)
$\mathcal{M}_5$ vs. $\mathcal{M}_6$	5.085	(0.24)	5.092	(0.23)	4.849	(0.24)	4.851	(0.23)	4.769	(0.24)	4.774	(0.23)
$\mathcal{M}_6$ vs. $\mathcal{M}_7$	3.823	(0.23)	3.799	(0.23)	3.599	(0.25)	3.593	(0.24)	3.855	(0.25)	3.843	(0.24)
$\mathcal{M}_7$ vs. $\mathcal{M}_8$	0.999	(0.23)	1.003	(0.22)	0.924	(0.25)	0.916	(0.24)	0.644	(0.25)	0.642	(0.25)
$\mathcal{M}_8$ vs. $\mathcal{M}_9$	0.204	(0.23)	0.208	(0.22)	0.216	(0.24)	0.210	(0.24)	0.039	(0.25)	0.029	(0.24)
$\mathcal{M}_9$ vs. $\mathcal{M}_{10}$	-0.827	(0.25)	-0.830	(0.24)	-0.817	(0.26)	-0.822	(0.25)	-0.483	(0.26)	-0.477	(0.25)
$\mathcal{M}_{10}$ vs. $\mathcal{M}_{11}$	-0.016	(0.26)	-0.024	(0.26)	-0.041	(0.25)	-0.028	(0.25)	-0.436	(0.26)	-0.431	(0.25)
$\mathcal{M}_{11}$ vs. $\mathcal{M}_{12}$	-0.733	(0.25)	-0.735	(0.25)	-0.869	(0.26)	-0.873	(0.26)	-0.966	(0.27)	-0.973	(0.26)
$\mathcal{M}_{12}$ vs. $\mathcal{M}_{13}$	-1.259	(0.24)	-1.264	(0.24)	-1.591	(0.27)	-1.598	(0.26)	-1.710	(0.27)	-1.719	(0.26)
$\mathcal{M}_{13}$ vs. $\mathcal{M}_{14}$	-1.636	(0.25)	-1.635	(0.24)	-1.028	(0.27)	-1.028	(0.26)	-1.209	(0.27)	-1.206	(0.27)
$\mathcal{M}_{14}$ vs. $\mathcal{M}_{15}$	-1.573	(0.26)	-1.587	(0.26)	-1.863	(0.27)	-1.879	(0.27)	-1.935	(0.28)	-1.954	(0.27)
$\mathcal{M}_{15}$ vs. $\mathcal{M}_{16}$	-1.766	(0.27)	-1.759	(0.26)	-1.756	(0.27)	-1.755	(0.27)	-1.804	(0.27)	-1.807	(0.26)
$\mathcal{M}_{16}$ vs. $\mathcal{M}_{17}$	-3.412	(0.26)	-3.416	(0.26)	-3.505	(0.29)	-3.520	(0.28)	-3.572	(0.29)	-3.587	(0.28)

Table 5: Differences in information criteria (IC), WAIC and LOO IC, between subsequent QPM candidate models and their Monte Carlo (MC) standard error in parentheses. The ICs are calculated with an MC sample size of 10,000.

	LEB		hyper-g/n		ZS adapted	
$\mathcal{M}_8$ vs. $\mathcal{M}_9$	0.099	(0.035)	-0.017	(0.037)	0.052	(0.036)
$\mathcal{M}_9$ vs. $\mathcal{M}_{10}$	-0.345	(0.035)	-0.108	(0.037)	-0.393	(0.035)

Table 6: Differences in WAIC between the QPM $\mathcal{M}_9$ and its neighboring models $\mathcal{M}_8$ and $\mathcal{M}_{10}$ as well as the Monte Carlo (MC) standard error of these differences in parentheses. The WAICs are calculated with an MC sample size of 500,000.

5. Discussion

In this paper, we propose a new approach to Bayesian variable selection, a generalisation of Barbieri and Berger's MPM. In our application to data from the GUSTO-I study, the presented quantile probability model turns out to be invariant to the prior choices on the model space and on the regression coefficients. An important feature of the proposed methodology is that we drastically reduce a setting with 2^p candidate models to only $p + 1$ models of interest. To find the QPM we assess the prediction performance of these $p+1$ models using the deviance information criterion or an alternative such as Watanabe-Akaike or leave-one out cross-validated information criteria. While searching for the QPM, we also developed an efficient, computationally fast and easily implemented method to find the Monte Carlo standard error of these model selection criteria. We also discussed a potential quantile probability model average which however yields only poor results.

Since the QPM uses DIC which is a Bayesian analogue of AIC, it is an efficient variable selection method. Claeskens and Hjort (2008) define a criterion like AIC as efficient if it selects a model so that the ratio of the expected predictive loss function at the selected model and the expected loss function at its theoretical minimiser converges to one in probability. In other words, an efficient model selection criterion always selects the optimal prediction model.

In the application we have seen that the MAP model is not always among the QPM candidate models. This can be due to the fact that some potential variables are strongly correlated which is the case in our application for x_1, x_9 and x_{10}. Ghosh and Ghattas

(2015) pointed out, that Bayesian variable selection under collinearity demands special caution and the MPM might perform poorly. The authors also demonstrated using simulation studies that the MPM variable inclusion threshold of 0.5 may not be appropriate for highly correlated covariates.

Vehtari and Lampinen (2004) have used the Kullback-Leibler divergence to find the simplest submodel which has similar predictive distribution as the full model. When nested models are considered, the submodels are defined using the marginal posterior probabilities to find the k most probable covariates. They also apply their approach to non-nested models in which case the "full model" is the Bayesian model average. In order to compute the expected predictive discrepancy between models they use k-fold cross-validation. The main difference between their and our approach is that they do not seek to optimise their criterion but try to graphically find an "elbow" in predictive explanatory power. Furthermore, they use cross-validated estimates of the expected criterion, but they do not compute standard errors.

Barbieri and Berger proved that the MPM is often optimal when we are selecting among linear models. We did not investigate the prior choice sensitivity in the linear model case, but it will probably be very similar to the one in GLMs. We note that, if the ordering of the variables is fixed regarding their inclusion probabilities, the QPM is independent of the prior on the model space, since the predictive model selection criteria are not influenced by the model prior. This is not the case for the MPM.

The presented methodology can easily be extended to derive optimal prediction models for survival outcomes. Of particular interest are dynamic prediction models for time-to-event data using a landmark approach as described in Section 3.3 of Held et al. (2016). We plan to do this in the context of the COMBACTE-MAGNET project to develop better prediction models for the risk of acquiring a ventilator associated pneumonia attributed to *P. aeruginosa*.

Acknowledgment

R.H. has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115737-2 (Combatting bacterial resistance in Europe

Scenario	α_{true}	$\boldsymbol{\beta}_{\text{true},1:5}$	$\boldsymbol{\beta}_{\text{true},6:10}$	$\boldsymbol{\beta}_{\text{true},11:15}$
Null	0	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$
Sparse	0	$\mathbf{b}$	$\mathbf{0}$	$\mathbf{0}$
Medium	0	$\mathbf{b}$	$\mathbf{0}$	$\mathbf{b}$
Full	0	$\mathbf{b}$	$\mathbf{b}$	$\mathbf{b}$

Table 7: Values of the coefficients (α_{true}, $\boldsymbol{\beta}_{\text{true}}$) in our logistic regression simulation study with $p = 15$, where $\mathbf{b} = (2, -1, -1, 0.5, -0.5)^{\text{T}}$.

- molecules against gram negative infections [COMBACTE-MAGNET]). The funders had no role in data collection and analysis, decision to publish, or preparation of the manuscript. We acknowledge helpful comments by an Associate Editor and two referees on an earlier version of this article.

A. Simulation Study

In order to study the sensitivity of the median probability model (MPM) towards prior settings we performed a simulation study on logistic regression. We closely follow the setup of the logistic regression simulation study in Li and Clyde (2016) using $p = 15$ potential predictors. For each of the $S = 250$ simulated training dataset, we draw the columns of the design matrix $\mathbf{X}$ with $n = 1000$ rows from a standard normal distribution with pairwise correlation $\text{cor}(\mathbf{X}_i, \mathbf{X}_j) = r^{|i-j|}$ for $1 \leq i < j \leq p$, with $r \in \{0, 0.75\}$. As Li and Clyde, we consider four different levels of sparsity in the true model which are summarised in Table 7. A Beta(a, b) distribution is chosen on the prior inclusion probabilities with first, $a = b = 1$ resulting in a multiplicity-corrected model prior and second, with $a = 2/4, b = 6/4$ giving each variable a prior inclusion probability (IP) of $\frac{a}{a+b} = 0.25$ instead of 0.5.

In each of the 250 simulated dataset, the QPM as well as the MPM were retrieved using a sample size of 10'000 for the DIC computation. The model complexity, meaning the number of variables selected, is stored.

Figure 6 illustrates how much more sensitive the MPM is towards prior choices. The average number of variables selected in the 250 MPMs and the 250 QPMs respec-

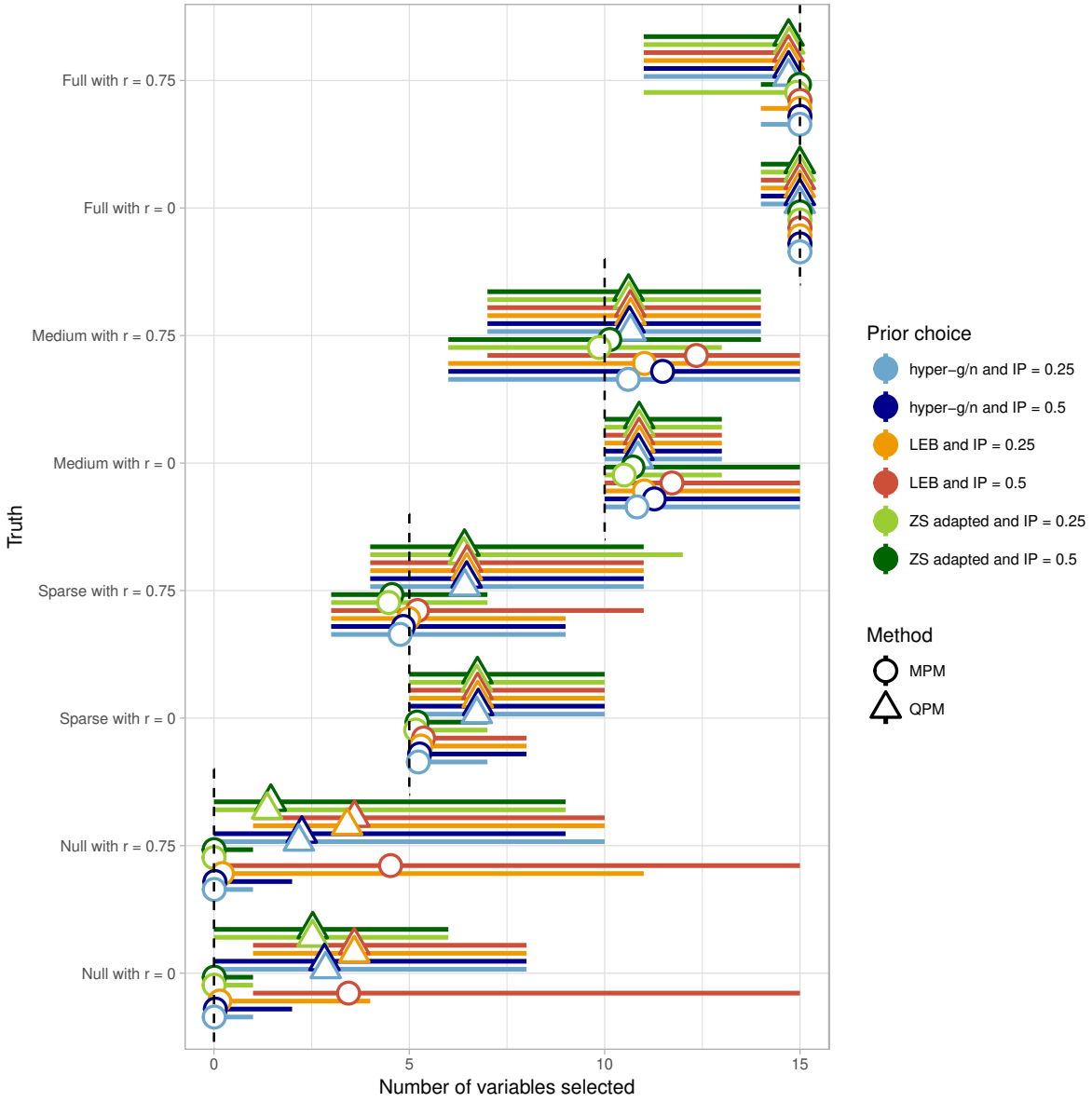

Figure 6: Average number of variables in each model as well as the range of the model complexity in the 250 simulated datasets given the scenario and a specific r. The true number of variables is represented by the dashed vertical lines.

tively is represented by circles (MPM) and triangles (QPM). Furthermore, the range in which the complexity of these models lies for each of the simulated datasets is shown. The average model size as well as its range in the datasets is very similar among prior choices for the QPM, apart from the scenario where the truth is the null model. The model complexity of the MPM, on the other hand depends a lot more on the prior choice. Moreover, as already discussed in the literature (Liang et al., 2008; Johnson, 2008), the LEB approach is not consistent if the null model is true, see the lowest part of Figure 6.

Then, the weighted sum of squared errors SSE($\boldsymbol{\beta}$) $= \sum_{j=0}^{p} \omega_j (\hat{\beta}_j - \beta_{\text{true},j})^2$ is computed, where $\hat{\beta}_j$ is the shrunken estimated coefficient for variables j, with $j = 0$ referring to the unshrunken intercept, $\omega_j = h(\beta_{\text{true},j})(1 - h(\beta_{\text{true},j}))$ and $h(\cdot)$ being the expit response function for logistic regression introduced in Section 2. We used a weighted version of the SSE to account for the difference in magnitude of the coefficients in Table 7, with weights ω_j equal to the derivative of $h(\cdot)$ at $\beta_{\text{true},j}$. Ten times the average weighted SSE($\boldsymbol{\beta}$) of the selected QPM and MPM for each of the 250 realizations of the sparse and medium scenarios is reported in the first two columns of Table 8. To ensure readability of the findings we only show the results for the sparse and medium versions of the true model. The estimation error is lower for the MPM if the true model is sparse and $r = 0$. For the remaining scenarios, QPM usually performs slightly better.

The QPM, as well as the MPM, are designed to be efficient model selection methods. Claeskens and Hjort (2008) discuss the fact that an efficient selection approach cannot be consistent to find the truth. Figure 6 shows that both methods struggle to select the true variables. From this figure, it is already possible to conclude that the MPM performs generally better than the QPM in finding the null as well as the full model. The four last columns of Table 8 show the average number of correctly treated variables (excluded or included) out of 15 variables by the QPM and the MPM selection methods for different prior choices in the 250 simulated data sets for the sparse and medium true model scenarios. The MPM correctly includes/excludes more variables than the QPM if the true model is sparse whereas QPM does better if the truth is medium.

Then, to compare the performance of the MPM with the QPM's performance, we simulate an independent test dataset with $n = 1000$ for each training dataset and compute the area under the ROC curve (AUC, measures discrimination), the calibration slope (CS) (Cox (1958), measures calibration), the logarithmic score (LS) as well as the Brier score (BS) (measure both discrimination and calibration) (Held et al., 2015). See Steyerberg (2009) for a practical review on the methods to validate probabilistic predictions. Both, CS and AUC should be as close as possible to 1, which means perfect calibration, respectively discrimination. The LS and the BS are negatively oriented, meaning that the smaller they are the better the model's performance. Table 9 shows these scores for the sparse and medium scenarios. If the Monte-Carlo error of the difference in scores between QPM and MPM is sufficiently small, the best score is marked in bold. Generally, the MPM performs better if the true model is sparse whereas the QPM shows better calibration and discrimination if the true model is medium.

The most important lesson learned from this simulation study is the sensitivity of the MPM with respect to the prior settings. The QPM tends to include a little bit more variables in the model, but this selection does not get influenced by the priors much. However using for example LEB for the estimation of g and a multiplicity-corrected prior on the model space, leads to a much more complex MPM then using a ZS adapted prior on g and a prior inclusion probability of 0.25 (light green and red circles in Figure 6). Further, if the true model is sparse, MPM scores better with regards to discrimination and calibration whereas QPM does best if the truth is a more complex model.

		$10*$ weighted SSE				Number correctly treated variables			
		sparse		medium		sparse		medium	
		$r = 0$	$r = .75$	$r = 0$	$r = .75$	$r = 0$	$r = .75$	$r = 0$	$r = .75$
LEB	MPM	7.11	7.36	10.59	**10.82**	**12.92**	**12.67**	10.23	9.66
with IP=.5	QPM	**7.09**	**7.30**	**10.59**	10.89	12.05	12.10	**10.78**	**10.70**
LEB	MPM	7.10	7.36	10.59	**10.82**	**13.00**	**12.69**	10.50	10.07
with IP=.25	QPM	**7.09**	**7.30**	**10.59**	10.89	12.02	12.10	**10.78**	**10.70**
hyper-g/n	MPM	**7.09**	7.36	10.61	**10.86**	**13.04**	**12.70**	**10.86**	**10.75**
with IP=.5	QPM	7.10	**7.33**	**10.61**	10.93	12.04	12.16	10.78	10.69
hyper-g/n	MPM	7.09	7.33	10.59	**10.82**	**12.98**	**12.70**	10.67	10.40
with IP=.25	QPM	**7.09**	**7.30**	**10.59**	10.89	12.04	12.12	**10.76**	**10.69**
ZS adapted	MPM	7.09	7.34	10.59	**10.82**	**13.02**	**12.70**	10.79	10.58
with IP=.5	QPM	**7.09**	**7.30**	**10.59**	10.90	12.06	12.15	**10.79**	**10.68**
ZS adapted	MPM	**7.08**	7.38	10.61	**10.87**	**13.07**	**12.66**	**11.00**	**10.78**
with IP=.25	QPM	7.10	**7.33**	**10.61**	10.92	12.07	12.14	10.78	10.69

Table 8: 10 times the average weighted SSE and the average number of correctly included or excluded variables in the different MPMs and QPMs, out of the 250 simulated data sets. The minimum weighted SSE and the maximum number of correctly treated variables for each prior combination are in bold face.

		AUC				CS				LS				BS			
		sparse		medium		sparse		medium		sparse		medium		sparse		medium	
		$r=0$	$r=.75$	$r=0$	$r=.75$	$r=0$	$r=.75$	$r=0$	$r=.75$	$r=0$	$r=.75$	$r=0$	$r=.75$	$r=0$	$r=.75$	$r=0$	$r=.75$
LEB	MPM	**0.7510**	0.6954	0.7755	0.7310	**0.9790**	**0.9954**	0.9708	0.9711	**0.5108**	0.5914	0.4594	0.5444	**0.1743**	0.2049	0.1564	0.1867
with IP=.5	QPM	0.7503	0.6952	**0.7757**	**0.7312**	0.9654	0.9873	0.9740	**0.9732**	0.5121	0.5917	**0.4591**	**0.5441**	0.1748	0.2050	**0.1563**	**0.1865**
LEB	MPM	**0.7511**	0.6952	0.7756	0.7310	**0.9803**	**0.9961**	0.9733	0.9715	**0.5107**	0.5916	0.4591	0.5443	**0.1743**	0.2050	0.1563	0.1866
with IP=.25	QPM	0.7503	0.6953	0.7757	**0.7312**	0.9653	0.9881	0.9736	**0.9731**	0.5121	0.5916	0.4591	**0.5440**	0.1748	0.2050	0.1563	**0.1865**
hyper-g/n	MPM	**0.7511**	0.6952	0.7756	0.7310	**0.9790**	**0.9927**	0.9702	0.9677	**0.5106**	0.5916	0.4592	0.5444	**0.1743**	0.2050	0.1564	0.1867
with IP=.5	QPM	0.7503	0.6952	**0.7757**	**0.7312**	0.9629	0.9829	0.9719	**0.9698**	0.5122	0.5917	**0.4591**	**0.5441**	0.1748	0.2050	**0.1563**	**0.1865**
hyper-g/n	MPM	**0.7511**	0.6952	0.7757	0.7310	**0.9796**	**0.9934**	0.9719	0.9688	**0.5105**	0.5917	0.4591	0.5443	**0.1742**	0.2050	0.1563	0.1866
with IP=.25	QPM	0.7503	0.6952	0.7757	**0.7312**	0.9634	0.9835	0.9719	0.9696	0.5121	0.5917	0.4591	**0.5441**	0.1748	0.2050	0.1563	**0.1865**
ZS adapted	MPM	**0.7512**	0.6950	**0.7757**	0.7310	**0.9741**	**0.9849**	**0.9612**	0.9532	0.5104	0.5919	**0.4591**	0.5444	**0.1742**	0.2051	0.1563	0.1866
with IP=.5	QPM	0.7503	0.6952	0.7757	**0.7312**	0.9544	0.9695	0.9600	0.9524	0.5122	0.5917	0.4592	**0.5442**	0.1748	0.2050	0.1563	**0.1865**
ZS adapted	MPM	**0.7512**	0.6949	**0.7758**	0.7309	**0.9751**	**0.9854**	**0.9638**	**0.9541**	**0.5103**	0.5920	**0.4589**	0.5446	**0.1742**	0.2051	**0.1562**	0.1867
with IP=.25	QPM	0.7503	0.6952	0.7757	**0.7312**	0.9547	0.9695	0.9600	0.9522	0.5122	0.5917	0.4592	**0.5442**	0.1748	0.2050	0.1563	**0.1865**

Table 9: The average AUC, CS, LS and BS for the sparse and medium scenarios. The best scores for each prior combination are marked in bold if twice the Monte-Carlo error of the difference between MPM and QPM is smaller than the difference.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory, Budapest: Akademiai Kiado, 267–281.

Barbieri, M. M., Berger, J. O., 2004. Optimal predictive model selection. Annals of Statistics 32 (3), 870–897.
URL: http://dx.doi.org/10.1214/009053604000000238
DOI: 10.1214/009053604000000238

Buckland, S. T., Burnham, K. P., Augustin, N. H., 1997. Model selection: an integral part of inference. Biometrics 53 (2), 603–618.

Claeskens, G., Hjort, N. L., 2008. Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
DOI: 10.1017/CBO9780511790485

Cox, D. R., 1958. Two further applications of a model for binary regression. Biometrika 45 (3/4), 562–565.

Ding, L., Kurowski, B. G., He, H., Alexander, E. S., Mersha, T. B., Fardo, D. W., Zhang, X., Pilipenko, V. V., Kottyan, L., Martin, L. J., 2014. Modeling of multivariate longitudinal phenotypes in family genetic studies with Bayesian multiplicity adjustment. BMC Proceedings 8 (Suppl 1).
URL: http://dx.doi.org/10.1186/1753-6561-8-S1-S69
DOI: 10.1186/1753-6561-8-S1-S69

Ennis, M., Hinton, G., Naylor, D., Revow, M., Tibshirani, R., 1998. A comparison of statistical learning methods on the GUSTO database. Statistics in Medicine 17 (21), 2501–2508.
URL: http://dx.doi.org/10.1002/(SICI)1097-0258(19981115)17:21<2501::AID-SIM938>3.0.CO;2-M

Flegal, J. M., Haran, M., Jones, G. L., 2008. Markov chain Monte Carlo: Can we trust the third significant figure? Statististical Science 23 (2), 250–260.

URL: http://dx.doi.org/10.1214/08-STS257
DOI: 10.1214/08-STS257

Gelman, A., Hwang, J., Vehtari, A., Nov. 2014. Understanding predictive information criteria for Bayesian models. Statistics and Computing 24 (6), 997–1016.
URL: http://dx.doi.org/10.1007/s11222-013-9416-2
DOI: 10.1007/s11222-013-9416-2

George, E. I., Foster, D. P., 2000. Calibration and empirical Bayes variable selection. Biometrika 87 (4), 731–747.
URL: http://biomet.oxfordjournals.org/content/87/4/731.abstract
DOI: 10.1093/biomet/87.4.731

Ghosh, J., 2015. Bayesian model selection using the median probability model. Wiley Interdisciplinary Reviews: Computational Statistics 7 (3), 185–193.
URL: http://dx.doi.org/10.1002/wics.1352
DOI: 10.1002/wics.1352

Ghosh, J., Ghattas, A. E., 2015. Bayesian variable selection under collinearity. The American Statistician 69 (3), 165–173.
URL: http://dx.doi.org/10.1080/00031305.2015.1031827
DOI: 10.1080/00031305.2015.1031827

Held, L., Gravestock, I., Sabanés Bové, D., 2016. Objective Bayesian model selection for Cox regression. Statistics in Medicine 35 (29), 5376–5390.
URL: http://dx.doi.org/10.1002/sim.7089

Held, L., Sabanés Bové, D., Gravestock, I., 2015. Approximate Bayesian model selection with the deviance statistic. Statistical Science 30 (2), 242–257.
URL: http://dx.doi.org/10.1214/14-STS510
DOI: 10.1214/14-STS510

Hu, J., Johnson, V. E., 2009. Bayesian model selection using test statistics. Journal of the Royal Statisical Society. Series B 71 (1), 143–158.

Johnson, V. E., 2008. Properties of Bayes factors based on test statistics. Scandinavian Journal of Statistics 35 (2), 354–368.
URL: http://dx.doi.org/10.1111/j.1467-9469.2007.00576.x
DOI: 10.1111/j.1467-9469.2007.00576.x

Kass, R. E., Raftery, A. E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.
URL: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572
DOI: 10.1080/01621459.1995.10476572

Lee, K. L., Woodlief, L., Topol, E. J., Weaver, W. D., Betriu, A., Col, J., Simoons, M., Aylward, P., Van de Werf, F., Califf, R. M., for the GUSTO-I Investigators, 1995. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction: results from an international trial of 41,021 patients. Circulation 91 (6), 1659–1668.

Li, Y., Clyde, M. A., 2016. Mixtures of g-priors in generalized linear models. ArXiv e-prints.
URL: https://arxiv.org/abs/1503.06913v2

Liang, F., Paulo, R., Molina, G., Clyde, M. A., Berger, J. O., 2008. Mixtures of g priors for Bayesian variable selection. Journal of the American Statistical Association 103 (481), 410–423.
URL: http://dx.doi.org/10.1198/016214507000001337
DOI: 10.1198/016214507000001337

Lunn, D., Jackson, C., Best, N., Thomas, A., Spiegelhalter, D., 2012. The BUGS Book - A Practical Introduction to Bayesian Analysis. CRC Press / Chapman and Hall.

Piironen, J., Vehtari, A., 2016. Comparison of Bayesian predictive methods for model selection. Statistics and Computing, 1–25.
URL: http://dx.doi.org/10.1007/s11222-016-9649-y
DOI: 10.1007/s11222-016-9649-y

Sabanés Bové, D., Held, L., 2011. Hyper-g priors for generalized linear models. Bayesian Analysis 6 (3), 387–410.
URL: http://dx.doi.org/10.1214/ba/1339616469
DOI: 10.1214/ba/1339616469

Scott, J. G., Berger, J. O., 2010. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Annals of Statistics 38 (5), 2587–2619.
URL: http://dx.doi.org/10.1214/10-AOS792
DOI: 10.1214/10-AOS792

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B 64 (4), 583–639.

Steyerberg, E., 2009. Clinical Prediction Models. Springer, New York.

Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society. Series B 39 (1), 44–47.
URL: http://www.jstor.org/stable/2984877

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one out cross-validation and WAIC. Statistics and Computing 27, 1413–1432.

Vehtari, A., Lampinen, J., 2004. Model selection via predictive explanatory power. Tech. rep., Laboratory of Computational Engineering, Helsinki University of Technology.

Wagenmakers, E.-J., Farrell, S., 2004. AIC model selection using Akaike weights. Psychonomic Bulletin and Review 11, 192–196.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 11, 3571–3594.

Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel, P. K., Zellner, A. (Eds.), Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Vol. 6 of Studies in Bayesian Econometrics and Statistics. North-Holland, Amsterdam, Ch. 5, pp. 233–243.

Zellner, A., Siow, A., 1980. Posterior odds ratios for selected regression hypotheses. In: Bernardo, J. M., DeGroot, M. H., Lindley, D. V., Smith, A. F. M. (Eds.), Bayesian Statistics: Proceedings of the First International Meeting held in Valencia. University of Valencia Press, Valencia, pp. 585–603.

Zhu, L., Carlin, B. P., 2000. Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. Statistics in Medicine 19 (17-18), 2265–2278.