Bayesian Analysis (2021) **16**, Number 4, pp. 1085–1112

The Median Probability Model and Correlated Variables

Maria M. Barbieri[*], James O. Berger[†], Edward I. George[‡], and Veronika Ročková[§,¶]

Abstract. The median probability model (MPM) (Barbieri and Berger, 2004) is defined as the model consisting of those variables whose marginal posterior probability of inclusion is at least 0.5. The MPM rule yields the best single model for prediction in orthogonal and nested correlated designs. This result was originally conceived under a specific class of priors, such as the point mass mixtures of non-informative and g-type priors. The MPM rule, however, has become so very popular that it is now being deployed for a wider variety of priors and under correlated designs, where the properties of MPM are not yet completely understood. The main thrust of this work is to shed light on properties of MPM in these contexts by (a) characterizing situations when MPM is still safe under correlated designs, (b) providing significant generalizations of MPM to a broader class of priors (such as continuous spike-and-slab priors). We also provide new supporting evidence for the suitability of g-priors, as opposed to independent product priors, using new predictive matching arguments. Furthermore, we emphasize the importance of prior model probabilities and highlight the merits of non-uniform prior probability assignments using the notion of model aggregates.

Keywords: Bayesian variable selection, median probability model, multicollinearity, spike and slab.

MSC2020 subject classifications: Primary 62C10; secondary 62F15.

1 Introduction

This paper investigates the extent to which the median probability model rule of Barbieri and Berger (2004) can be used for variable selection when the covariates are correlated. To this end, we consider the usual linear model

$$\boldsymbol{Y}_{n\times 1} \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right), \tag{1.1}$$

where $\boldsymbol{Y}$ is the $n \times 1$ vector of responses, $\boldsymbol{X}$ is the $n \times q$ design matrix of covariates, $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown coefficients, and σ^2 is a known or unknown scalar. The equation (1.1) corresponds to the full model and we are interested in selecting a

[*]Department of Economics, Università Roma Tre, Roma, Italy, marilena.barbieri@uniroma3.it

[†]Department of Statistical Science, Duke University, Durham, NC, U.S.A., berger@duke.edu

[‡]Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA, U.S.A., edgeorge@wharton.upenn.edu

[§]Booth School of Business, University of Chicago, Chicago, IL, U.S.A., Veronika.Rockova@chicagobooth.edu

[¶]The author gratefully acknowledges the support from the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business and the National Science Foundation (grant DMS-1944740).

 https://doi.org/10.1214/20-BA1249

submodel indexed by $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$, where $\gamma_i \in \{1, 0\}$ for whether the i^{th} covariate is in or out of the model. We tacitly assume that the response and predictors have been centered and thereby omit the intercept term.

For prediction of a new observation $y^\star$ from $\boldsymbol{x}^\star$ under squared error loss, the optimal model $\boldsymbol{\gamma}^o$ is known to satisfy (Lemma 1 of Barbieri and Berger (2004))

$$\boldsymbol{\gamma}^o = \arg\min_{\boldsymbol{\gamma}} R(\boldsymbol{\gamma}) \quad \text{with} \quad R(\boldsymbol{\gamma}) \equiv \left(\boldsymbol{H}_{\boldsymbol{\gamma}}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} - \bar{\boldsymbol{\beta}}\right)' \boldsymbol{Q} \left(\boldsymbol{H}_{\boldsymbol{\gamma}}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} - \bar{\boldsymbol{\beta}}\right), \tag{1.2}$$

where $\bar{\boldsymbol{\beta}} = \mathsf{E}[\boldsymbol{\beta} \,|\, \boldsymbol{Y}] = \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \,|\, \boldsymbol{Y}) \boldsymbol{H}_{\boldsymbol{\gamma}}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the overall posterior mean of $\boldsymbol{\beta}$ under the hierarchical prior $\pi(\boldsymbol{\gamma})$ and $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma})$; $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the conditional posterior mean under $\boldsymbol{\gamma}$; $\pi(\boldsymbol{\gamma} \,|\, \boldsymbol{Y})$ is the posterior probability of model $M_{\boldsymbol{\gamma}}$; $\boldsymbol{Q} = \mathsf{E}[\boldsymbol{x}^\star \boldsymbol{x}^{\star\prime}] = \boldsymbol{X}'\boldsymbol{X}$, essentially the assumption that random covariates in the future will be like those arising in the data; and $\boldsymbol{H}_{\boldsymbol{\gamma}}$ the $q \times |\boldsymbol{\gamma}|$ stretching matrix (defined in Section 2.2 of Barbieri and Berger (2004)) which satisfies

$$\boldsymbol{H}_{\boldsymbol{\gamma}} = (h_{ij})_{ij=1}^{q,|\boldsymbol{\gamma}|} \quad \text{where } h_{ij} \text{ equals 1 if } \gamma_i = 1 \text{ and } j = \sum_{r=1}^{i} \gamma_r, \text{ and 0 otherwise.} \tag{1.3}$$

As (1.2) reveals, $\boldsymbol{\gamma}^o$ can be regarded as the best single-model approximation to model averaging.

Contrary to what might be commonly conceived as an optimal predictive model, $\boldsymbol{\gamma}^o$ is *not* necessarily the modal highest posterior probability model. In orthogonal and nested correlated designs, Barbieri and Berger (2004) show that the optimal model $\boldsymbol{\gamma}^o$ is also the *median probability model* $\boldsymbol{\gamma}^{MP}$, namely the model consisting of variables whose marginal inclusion probability $\pi(\gamma_i = 1 \,|\, \boldsymbol{Y})$ is at least 0.5. Furthermore, compared to the maximum-a-posteriori model (MAP), a major attraction of the median probability model (MPM) is the speed with which it can be well approximated via Markov chain Monte Carlo (MCMC) methods. Whereas the MAP must be approximated by a single high probability model among the 2^p possible models, approximating the MPM only requires estimates of the p marginal inclusion probabilities, each of which can be quickly and more accurately estimated from MCMC sampled binary inclusion indicators. Thus even when the MAP and MPM are identical, which may often be the case, the MPM offers a much faster route to computing them both.

The MPM is now routinely used for distilling posterior evidence towards variable selection; Clyde et al. (2011); Feldkircher (2012); Garcia-Donato and Martinez-Beneito (2013); Ghosh (2015); Piironen and Vehtari (2017) and Drachal (2018) are some of the articles that have used and discussed the performance of the MPM. Despite its widespread use in practice, however, the optimality of MPM has so far been shown under comparatively limited circumstances. In particular, the priors $\pi(\boldsymbol{\beta} \,|\boldsymbol{\gamma})$ are required to be such that the MPM estimator $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is proportional to the maximum likelihood estimator (MLE) under $\boldsymbol{\gamma}$. This property will be satisfied by e.g. the *point-mass* spike-and-slab g-type priors (Zellner, 1986; Liang et al., 2008). However, the also very popular *continuous* spike-and-slab mixtures (George and McCulloch, 1993; Ishwaran and Rao,

2003; Ročková and George, 2014; Ročková, 2018) will fail to satisfy this requirement. Here, we will show that this condition is not necessary for MPM to be predictive optimal. In particular, we provide significant generalizations of the existing MPM optimality results for a wider range of priors such as the continuous spike-and-slab mixtures and, more generally, independent product priors.

Barbieri and Berger (2004) presented a situation with correlated covariates (due to Merlise Clyde) in which the MPM was clearly not optimal. Thus there has been a concern that correlated covariates (reality) might make the MPM practically irrelevant. Hence another purpose of this paper is to explore the extent to which correlated covariates can degrade the performance of the MPM. We address this with theoretical studies concerning the impact of correlated covariates, and numerical studies; the magnitude of the scientific domain here limits us (in the numerical studies) to consider a relatively exhaustive study of the two variable case, made possible by geometric considerations. The overall conclusion is that (in reality) there can be a small degradation of performance, but the degradation is less than that experienced by the MAP in correlated scenarios.

First, using predictive matching arguments (Berger and Pericchi, 2001; Bayarri et al., 2012; Fouskakis et al., 2018), we provide new arguments for the suitability of g-type priors as opposed to independent product priors. Going further, we highlight the importance of prior model probabilities assignments and discuss their "dilution" issues (George, 2010) in highly collinear designs. Introducing the notion of model aggregates, we showcase the somewhat peculiar behavior of separable model priors obtained with a fixed prior inclusion probability. We show that the beta-binomial prior copes far better with variable redundancy. We also characterize the optimal predictive model and relate it to the MPM through relative risk comparisons. We also provide several "mini-theorems" showing predictive (sub)optimality of the MPM when $q = 2$.

The paper is structured as follows. Section 2 introduces the notion of model collectives and looks into some interesting limiting behaviors of the MPM when the predictors are correlated. Section 3 delves into a special case with 2 collinear predictors. Section 4 generalizes the optimality of the MPM to other priors and Section 5 wraps up with a discussion.

2 Highly Correlated Variables, g-Priors and MPM

2.1 The Marginal Likelihood Under Many Highly Correlated Variables

One reasonable requirement for objective model selection priors is that they be properly matched across models that are indistinguishable from a predictive point of view. Recall that two models are regarded as *predictive matching* (Bayarri et al., 2012) if their marginal likelihoods are close in terms of some distance. In this section, we take a closer look at the marginal likelihood for the model (1.1) under the celebrated g-priors (Zellner, 1986), assuming that the $n \times (p + k)$ design matrix $\boldsymbol{X}_\epsilon$ satisfies

$$\boldsymbol{X}_\epsilon = [\boldsymbol{B}_{n\times p},\ \boldsymbol{x} + \epsilon\,\boldsymbol{\delta}_1,\ \cdots,\ \boldsymbol{x} + \epsilon\,\boldsymbol{\delta}_k] \tag{2.1}$$

for some $\epsilon > 0$, where $\boldsymbol{B}$ consists of p possibly correlated regressors and where $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k$ are $(n \times 1)$ perturbation vectors. For clarity of exposition, we first assume that $\sigma^2 > 0$ is fixed and later extend our considerations to the random case. We assume that $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k$ are orthonormal and orthogonal to an $(n \times 1)$ vector $\boldsymbol{x}$ and $\boldsymbol{B}$,[1] while $\boldsymbol{x}$ and $\boldsymbol{B}$ are not necessarily orthogonal. We will be letting ϵ be very small to model the situation of having k highly correlated variables. For the full model (1.1), the g-prior is

$$\boldsymbol{\beta}_{(p+k)\times 1} \sim \mathcal{N}_{p+k}\left(\boldsymbol{0}, g\,\sigma^2(\boldsymbol{X}_\epsilon'\boldsymbol{X}_\epsilon)^{-1}\right) \tag{2.2}$$

for some $g > 0$ (typically n). Assuming, for now, that σ^2 is fixed the corresponding marginal likelihood is

$$\boldsymbol{Y} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2\boldsymbol{I} + g\,\sigma^2\boldsymbol{X}_\epsilon(\boldsymbol{X}_\epsilon'\boldsymbol{X}_\epsilon)^{-1}\boldsymbol{X}_\epsilon'\right).$$

Note that $\boldsymbol{X}_\epsilon(\boldsymbol{X}_\epsilon'\boldsymbol{X}_\epsilon)^{-1}\boldsymbol{X}_\epsilon'$ is the projection matrix onto the column space of $\boldsymbol{X}_\epsilon$. Hence, having near duplicate columns in $\boldsymbol{X}_\epsilon$ should not change this matrix much at all. Indeed, the following Lemma shows that, as $\epsilon \to 0$, this is a fixed matrix (depending only on $\boldsymbol{B}$ and $\boldsymbol{x}$).

Lemma 1. *Denote with* $\boldsymbol{P} = \lim_{\epsilon\to 0}\ \boldsymbol{X}_\epsilon(\boldsymbol{X}_\epsilon'\boldsymbol{X}_\epsilon)^{-1}\boldsymbol{X}_\epsilon'$. *Then*

$$\boldsymbol{P} = \boldsymbol{P}_B + \frac{(\boldsymbol{I} - \boldsymbol{P}_B)\boldsymbol{x}\boldsymbol{x}'(\boldsymbol{I} - \boldsymbol{P}_B)}{\boldsymbol{x}'(\boldsymbol{I} - \boldsymbol{P}_B)\boldsymbol{x}}, \tag{2.3}$$

where $\boldsymbol{P}_{\boldsymbol{B}} = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'$.

Proof. Let $\boldsymbol{1}$ be the k-column vector of ones, so that $\boldsymbol{1}\boldsymbol{1}'$ is the $k \times k$ matrix of ones, and let $\boldsymbol{v} = \boldsymbol{B}'\boldsymbol{x}$. Note first that

$$\boldsymbol{X}_\epsilon'\boldsymbol{X}_\epsilon = \begin{pmatrix} \boldsymbol{B}'\boldsymbol{B} & \boldsymbol{v}\boldsymbol{1}' \\ \boldsymbol{1}\boldsymbol{v}' & \|\boldsymbol{x}\|^2\boldsymbol{1}\boldsymbol{1}' + \epsilon^2\boldsymbol{I} \end{pmatrix}$$

and, letting $C = (\|\boldsymbol{x}\|^2 - \boldsymbol{v}'(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{v})$,

$$(\boldsymbol{X}_\epsilon'\boldsymbol{X}_\epsilon)^{-1} = \begin{pmatrix} \left((\boldsymbol{B}'\boldsymbol{B})^{-1} + \frac{k}{\epsilon^2+kC}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{v}\boldsymbol{v}'(\boldsymbol{B}'\boldsymbol{B})^{-1}\right) & -(\epsilon^2+kC)^{-1}(\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{v}\boldsymbol{1}' \\ -(\epsilon^2+kC)^{-1}\boldsymbol{1}\boldsymbol{v}'(\boldsymbol{B}'\boldsymbol{B})^{-1} & \frac{1}{\epsilon^2}\left(\boldsymbol{I} - \frac{C}{\epsilon^2+kC}\boldsymbol{1}\boldsymbol{1}'\right) \end{pmatrix}.$$

The result follows by multiplying this matrix with $\boldsymbol{X}_\epsilon$ and $\boldsymbol{X}_\epsilon'$, and taking the limit as $\epsilon \to 0$. □

Lemma 1 can perhaps be more readily understood using the following intuitive explanation. With $k = 1$, the limiting design matrix $\lim_{\varepsilon\to 0}\boldsymbol{X}_\varepsilon = [\boldsymbol{B}\,|\,\boldsymbol{x}]$ is full rank and yields the projection matrix $\boldsymbol{P}$ in (2.3). While with $k > 1$ the limiting matrix of $\boldsymbol{X}_\varepsilon$ is no longer full rank, its columns $[\boldsymbol{B}\,|\,\boldsymbol{x}\,|\,\ldots\,|\,\boldsymbol{x}]$ span the same space as $[\boldsymbol{B}\,|\,\boldsymbol{x}]$ and so it is expected that the limiting projection matrix be the same as for the case $[\boldsymbol{B}\,|\,\boldsymbol{x}]$.

[1] This assumption is not necessary, but greatly simplifies the illustration.

One important conclusion from Lemma 1 is that no matter how many columns of highly correlated variables are present in the model, the marginal likelihood under the g-prior will essentially be

$$\boldsymbol{Y} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2\boldsymbol{I} + g\,\sigma^2\boldsymbol{P}\right)$$

as $\epsilon \to 0$. Thereby all models including all predictors in $\boldsymbol{B}$ and at least one replicate of $\boldsymbol{x}$ can be essentially regarded as predictive matching.

We let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')'$ denote the global vector of inclusion indicators, where $\boldsymbol{\gamma}_1$ is associated with $\boldsymbol{B}$ and $\boldsymbol{\gamma}_2$ is associated with the k near duplicates. The same analysis holds for any sub-model $\boldsymbol{\gamma}_1 \in \{0,1\}^p$, defined by the design matrix $\boldsymbol{B}_{\boldsymbol{\gamma}_1}$ consisting of the active variables corresponding to the 1's in $\boldsymbol{\gamma}_1$. Before proceeding, we introduce the notion of a *model collective* which will be useful for characterizing the properties of g-priors and the median probability model in collinear designs.

Definition 1 (A Model Collective)**.** *Let $\boldsymbol{\gamma}_1 \in \{0,1\}^p$ be a vector of inclusion indicators associated with the p variables in $\boldsymbol{B}$. Denote by $M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}$ the model collective comprising all models consisting of the $\boldsymbol{\gamma}_1$ variables together with one or more of the (near) duplicates of $\boldsymbol{x}$.*

Let $\boldsymbol{P}_{\boldsymbol{\gamma}_1}$ be the limiting projection matrix corresponding to any of the models inside the model collective $M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}$. The limiting marginal likelihood of such models under the g-prior is

$$m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1, \boldsymbol{x}) = \phi\left(\boldsymbol{y} \mid \boldsymbol{0}, \sigma^2\boldsymbol{I} + g\sigma^2\boldsymbol{P}_{\boldsymbol{\gamma}_1}\right), \tag{2.4}$$

where $\phi(\boldsymbol{y} \mid \boldsymbol{\mu}, \Sigma)$ denotes a multivariate Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

Lemma 2. *Let $m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1)$ denote the marginal likelihood under the model $\boldsymbol{\gamma}_1$. Then we have*

$$m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1, \boldsymbol{x}) = \psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}_1) \times m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1), \tag{2.5}$$

where

$$\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}_1) = \frac{1}{\sqrt{1+g}} \exp\left\{\frac{g}{2\sigma^2(1+g)} \times \frac{[\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{B}_{\boldsymbol{\gamma}_1}})\boldsymbol{x}]^2}{\boldsymbol{x}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{B}_{\boldsymbol{\gamma}_1}})\boldsymbol{x}}\right\} \tag{2.6}$$

and $\boldsymbol{P}_{\boldsymbol{B}_{\boldsymbol{\gamma}_1}} = \boldsymbol{B}_{\boldsymbol{\gamma}_1}(\boldsymbol{B}_{\boldsymbol{\gamma}_1}'\boldsymbol{B}_{\boldsymbol{\gamma}_1})^{-1}\boldsymbol{B}_{\boldsymbol{\gamma}_1}'$. Note that, if $\boldsymbol{x}$ is orthogonal to $\boldsymbol{B}$, then

$$\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}_1) = \psi(\boldsymbol{y}, \boldsymbol{x}) \equiv \frac{1}{\sqrt{1+g}} \exp\left\{\frac{g}{2\sigma^2(1+g)} \times \left[\boldsymbol{y}'\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right]^2\right\}. \tag{2.7}$$

Proof. We denote with $\boldsymbol{P}$ the projection matrix $\boldsymbol{P}_{\boldsymbol{B}_{\boldsymbol{\gamma}}}$ and $\boldsymbol{z} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}/\sqrt{\boldsymbol{x}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}}$. We use the fact that $(\boldsymbol{I}+g\boldsymbol{P})(\boldsymbol{I}-\boldsymbol{P}) = (\boldsymbol{I}-\boldsymbol{P})$ and therefore $(\boldsymbol{I}+g\boldsymbol{P})^{-1}(\boldsymbol{I}-\boldsymbol{P}) = (\boldsymbol{I}-\boldsymbol{P})$ to find that

$$\begin{aligned}(\boldsymbol{I} + g\boldsymbol{P} + g\boldsymbol{z}\boldsymbol{z}')^{-1} &= (\boldsymbol{I} + g\boldsymbol{P})^{-1} - \frac{(\boldsymbol{I} + g\boldsymbol{P})^{-1}\boldsymbol{z}\boldsymbol{z}'(\boldsymbol{I} + g\boldsymbol{P})^{-1}}{g^{-1} + \boldsymbol{z}'(\boldsymbol{I} + g\boldsymbol{P})^{-1}\boldsymbol{z}} \\ &= (\boldsymbol{I} + g\boldsymbol{P})^{-1} - \frac{g(\boldsymbol{I} - \boldsymbol{P})^{-1}\boldsymbol{x}\boldsymbol{x}'(\boldsymbol{I} - \boldsymbol{P})^{-1}}{1+g}.\end{aligned}$$

The statement then follows from (2.4) using the fact

$$|\boldsymbol{I}+g\boldsymbol{P}+g\boldsymbol{z}\boldsymbol{z}'| = |\boldsymbol{I}+g\boldsymbol{P}|\,(1+g\boldsymbol{z}'(\boldsymbol{I}+g\boldsymbol{P})^{-1}\boldsymbol{z}) = |\boldsymbol{I}+g\boldsymbol{P}|(1+g). \qquad \square$$

Remark 1. *Note that the quantity (2.7) is proportional to the marginal likelihood $m(\boldsymbol{y}\,|\,\boldsymbol{x})$ of a model including only a covariate $\boldsymbol{x}$. Indeed,*

$$\psi(\boldsymbol{y},\boldsymbol{x}) = (2\pi)^{n/2}\sigma^n \mathrm{e}^{\frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{y}} m(\boldsymbol{y}\,|\,\boldsymbol{x}).$$

Remark 2. *If $\boldsymbol{x}$ is orthogonal to $\boldsymbol{B}$, the corresponding Bayes estimates are just the usual g-prior posterior means*

$$\frac{g}{1+g}\left(\widehat{\boldsymbol{\beta}}^{MLE}_{\boldsymbol{\gamma}_1}, \frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|^2}\right). \tag{2.8}$$

Moreover, adding at least one of the near-identical predictors multiplies the limiting marginal likelihood by a constant factor that does not depend on the number of copies. The same conclusion applies also to the case when σ^2 is random (Remark 3).

Remark 3 (The Case of Unknown Variance)**.** *Consider a conjugate prior form (2.2) with σ^2 having an inverse gamma distribution*

$$\sigma^2 \sim IG(\eta/2, \eta\lambda/2). \tag{2.9}$$

The limiting marginal likelihood of each model in the model collective $M_{\boldsymbol{\gamma}_1,x}$ and a model $\boldsymbol{\gamma}_1$ satisfy (Section 4.2 of George and McCulloch (1997))

$$m(\boldsymbol{y}\,|\,\boldsymbol{\gamma}_1,\boldsymbol{x}) \propto \frac{1}{\sqrt{g+1}}(\eta\lambda + S^2_{\boldsymbol{\gamma}_1,x})^{-\frac{n+\eta}{2}} \quad \textit{and} \quad m(\boldsymbol{y}\,|\,\boldsymbol{\gamma}_1) \propto \frac{1}{\sqrt{g+1}}(\eta\lambda + S^2_{\boldsymbol{\gamma}_1})^{-\frac{n+\eta}{2}},$$

where, using Lemma 1,

$$S^2_{\boldsymbol{\gamma}_1,x} = \underbrace{\boldsymbol{y}'\left(\boldsymbol{I} - \frac{g}{g+1}\boldsymbol{P}_{\boldsymbol{B}_{\boldsymbol{\gamma}_1}}\right)\boldsymbol{y}}_{S^2_{\boldsymbol{\gamma}_1}} - \underbrace{\frac{g}{g+1}\frac{[\boldsymbol{y}(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{B}_{\boldsymbol{\gamma}_1}})\boldsymbol{x}]^2}{\boldsymbol{x}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{B}_{\boldsymbol{\gamma}_1}})\boldsymbol{x}}}_{C(\boldsymbol{\gamma}_1,\boldsymbol{y},\boldsymbol{x},g)}.$$

The limiting marginal likelihood satisfies (2.5) with

$$\psi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\gamma}_1) = \left(1 - \frac{C(\boldsymbol{\gamma}_1,\boldsymbol{y},\boldsymbol{x},g)}{\eta\lambda + S^2_{\boldsymbol{\gamma}_1}}\right)^{-\frac{n+\eta}{2}}.$$

For the orthogonal case (when $\boldsymbol{x}$ is orthogonal to $\boldsymbol{B}$), we have

$$\psi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\gamma}_1) = \left(1 - \frac{g/(g+1)[\boldsymbol{y}'\boldsymbol{x}/\|\boldsymbol{x}\|]^2}{\eta\lambda + S^2_{\boldsymbol{\gamma}_1}}\right)^{-\frac{n+\eta}{2}}.$$

Including at least one copy of the covariate $\boldsymbol{x}$ in a model $\boldsymbol{\gamma}_1$ still inflates the marginal likelihood, where now the inflation factor $\psi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\gamma}_1)$ depends on $\boldsymbol{\gamma}_1$. However, $\psi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\gamma}_1)$ still does not depend on the number of copies.

2.2 Dimensional Predictive Matching

As a first application of Lemma 2, we note that the (limiting) marginal likelihood under the g-prior is the same, no matter how many replicates of $\boldsymbol{x}$ are in the model. This property can be regarded as a variant of *dimensional predictive matching*, one of the desiderata relevant for the development of objective model selection priors (Bayarri et al. (2012)). This type of predictive matching across dimensions is, however, new in the sense that the matching holds for all training samples, not only the minimal ones.

Corollary 2.1. *Mixtures of g-priors are dimensional predictive matching in the sense that the limiting marginal likelihood of all models within the model collective is the same, provided that the mixing distribution over g is the same across all models.*

Proof. Follows directly from Lemma 2. □

Remark 4. *According to Remark 3, Corollary 2.1 applies to both fixed and random* σ^2 *with a prior* (2.9). *Note that a similar conclusion also holds for the usual objective prior* $1/\sigma^2$ *which is obtained as the limiting case as* $\eta \to 0$. *Throughout the rest of the paper, we implicitly assume the prior* (2.9) *and/or the objective prior* $1/\sigma^2$ *whenever we refer to random* σ^2.

In contrast, it is of interest to look at what happens with an alternative prior for $\boldsymbol{\beta}$ such as a $N(\mathbf{0}, \boldsymbol{I})$ prior. If a model has j near-replicates of $\boldsymbol{x}$, the effective parameter for $\boldsymbol{x}$ in that model is the sum of the j β's, which will each have a $N(0, j)$ prior. So the marginal likelihoods will depend strongly on the number of replicates, even though there is no difference in the models.

2.3 When all Non-Duplicated Covariates are Orthogonal

To get insights into the behavior of the median probability model for correlated predictors, we consider an instructive example obtained by setting $\epsilon = 0$ and $\boldsymbol{B}'\boldsymbol{B} = n\,\boldsymbol{I}$ in (2.1). In particular, we will be working with an orthogonal design that has been augmented with multiple copies of one predictor

$$\boldsymbol{X}_{n\times(p+k)} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \underbrace{\boldsymbol{x}, \ldots, \boldsymbol{x}}_{k}], \tag{2.10}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{x}$ are orthogonal and standardized so that $\|\boldsymbol{x}_i\|^2 = \|\boldsymbol{x}\|^2 = n$. There is no qualitative difference between this and the previous assumption of highly correlated covariate vectors; going directly to the limiting case of replicate covariate vectors makes matters pedagogically easier to understand.

A few points are made with this toy example. First, we want to characterize the optimal predictive model and generalize the MPM rule when the designs have blocks of (nearly) identical predictors. Second, we want to understand how close to the optimal predictive model the MPM actually is in this limiting case. Third, we want to highlight the benefits of the g-prior correlation structure. We denote by $z = \boldsymbol{x}'\boldsymbol{y}$, $z_i = \boldsymbol{x}_i'\boldsymbol{y}$ for $i = 1, \ldots, p$ and $\boldsymbol{z} = (\boldsymbol{z}_1', \boldsymbol{z}_2')'$, where $\boldsymbol{z}_1 = (z_1, \ldots, z_p)'$ and $\boldsymbol{z}_2 = z\mathbf{1}_k$. We will again

split the variable inclusion indicators into two groups $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')' \in \{0,1\}^{p+k}$, where $\boldsymbol{\gamma}_1$ is attached to the first p and $\boldsymbol{\gamma}_2$ to the last k predictors. To begin, we assume the generalized g-prior on regression coefficients, given the model $\boldsymbol{\gamma}$,

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim \mathcal{N}_{|\boldsymbol{\gamma}|}\left(\mathbf{0}, g\,\sigma^2(\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{X}_{\boldsymbol{\gamma}})^{+}\right), \tag{2.11}$$

where $(\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{X}_{\boldsymbol{\gamma}})^{+}$ is the Moore-Penrose pseudo-inverse and where $\boldsymbol{X}_{\boldsymbol{\gamma}}$ denotes the subset of $\boldsymbol{X}$ with active indicators $\boldsymbol{\gamma}$. The following lemma characterizes the optimal predictive model under (2.11) and (2.10).

Lemma 3. *Consider the model* (1.1) *where* $\boldsymbol{X}$ *satisfies* (2.10) *and where* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{x}$ *are orthogonal with* $\|\boldsymbol{x}_i\|^2 = \|\boldsymbol{x}\|^2 = n$. *Under the prior* (2.11) *and fixed or random* σ^2 *any model* $\boldsymbol{\gamma}^o = (\boldsymbol{\gamma}_1^{o\prime}, \boldsymbol{\gamma}_2^{o\prime})'$ *that satisfies*

$$\gamma_{1i}^o = 1 \quad \textit{iff} \quad \pi(\gamma_{1i} = 1 \,|\, \boldsymbol{Y}) > 0.5, \quad i = 1, \ldots, p, \tag{2.12}$$

$$|\boldsymbol{\gamma}_2^o| \geq 1 \quad \textit{iff} \quad \pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y}) > 0.5 \tag{2.13}$$

is predictive optimal.

Proof. Due to the block-diagonal nature of the matrix $\boldsymbol{X}'\boldsymbol{X} = n\begin{pmatrix} \mathrm{I}_p & \mathbf{0}_{p\times k} \\ \mathbf{0}_{k\times p} & \mathbf{1}_k\mathbf{1}_k' \end{pmatrix}$, the posterior mean under the non-null model $\boldsymbol{\gamma}$ satisfies

$$\boldsymbol{H}_{\boldsymbol{\gamma}}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \frac{g/n}{1+g}\begin{pmatrix} \mathrm{diag}\{\boldsymbol{\gamma}_1\} & \mathbf{0} \\ \mathbf{0} & \frac{1}{|\boldsymbol{\gamma}_2|}\mathrm{diag}\{\boldsymbol{\gamma}_2\} \end{pmatrix}\boldsymbol{z}\,.$$

The overall posterior mean $\bar{\boldsymbol{\beta}} = \mathsf{E}\,(\boldsymbol{\beta} \,|\, \boldsymbol{Y}) = \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \,|\, \boldsymbol{Y})\boldsymbol{H}_{\boldsymbol{\gamma}}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ then satisfies

$$\begin{aligned}\boldsymbol{H}_{\boldsymbol{\gamma}}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} - \bar{\boldsymbol{\beta}} &= \\ &= \frac{g/n}{(1+g)}\begin{pmatrix} \mathrm{diag}\{\boldsymbol{\gamma}_1 - \mathsf{E}[\boldsymbol{\gamma}_1 \,|\, \boldsymbol{Y}]\} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}\left\{\frac{\boldsymbol{\gamma}_2}{|\boldsymbol{\gamma}_2|} - \mathsf{E}\left[\frac{\boldsymbol{\gamma}_2}{|\boldsymbol{\gamma}_2|} \,|\, \boldsymbol{Y}, \boldsymbol{\gamma}_2 \neq \mathbf{0}\right]\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y})\right\} \end{pmatrix}\boldsymbol{z}\,.\end{aligned}$$

The optimal predictive model minimizes $R(\boldsymbol{\gamma})$ defined in (1.2). Due to the fact that $\boldsymbol{Q}$ is block-diagonal, the criterion $R(\boldsymbol{\gamma})$ separates into two parts, one involving the first p independent variables and the second involving the k identical copies. In particular, $R(\boldsymbol{\gamma}) = R_1(\boldsymbol{\gamma}_1) + R_2(\boldsymbol{\gamma}_2)$ where

$$R_1(\boldsymbol{\gamma}_1) = \frac{g^2/n}{(1+g)^2}\sum_{i=1}^{p} z_i^2(\gamma_{1i} - \mathsf{E}\,[\gamma_{1i} \,|\, \boldsymbol{Y}])^2, \tag{2.14}$$

$$\begin{aligned}R_2(\boldsymbol{\gamma}_2) = \frac{g^2 z^2/n}{(1+g)^2}\Big\{&[1 - \pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y})]^2\,\mathbb{I}(\boldsymbol{\gamma}_2 \neq \mathbf{0}) \\ &+ \pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y})^2\mathbb{I}(\boldsymbol{\gamma}_2 = \mathbf{0})\Big\}\,.\end{aligned} \tag{2.15}$$

The statement then follows from (2.14) and (2.15). With duplicate columns, the optimal predictive model $\boldsymbol{\gamma}^o \equiv \arg\min_{\boldsymbol{\gamma}} R(\boldsymbol{\gamma})$ is not unique. Any model $\boldsymbol{\gamma}^o = (\boldsymbol{\gamma}_1^{o\prime}, \boldsymbol{\gamma}_2^{o\prime})'$ defined through (2.12) and (2.13) will minimize the criterion $R(\boldsymbol{\gamma})$. □

The last k variables in the optimal predictive model thus act jointly as one variable, where the decision to include $\boldsymbol{x}$ is based on a *joint* posterior probability $\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y})$. This intuitively appealing treatment of $\boldsymbol{x}$ is an elegant byproduct of the g-prior. We

will see in the next section that such clustered inclusion no longer occurs in the optimal predictive model under independent product priors. The risk of the optimal model is

$$R(\boldsymbol{\gamma}^o) = \frac{g^2/n}{(1+g)^2} \sum_{i=1}^{p} z_i^2 \min\{\mathsf{E}\left[\gamma_{1i} \,|\, \boldsymbol{Y}\right], 1 - \mathsf{E}\left[\gamma_{1i} \,|\, \boldsymbol{Y}\right]\}^2 + \frac{g^2 z^2/n}{(1+g)^2} \min\{\pi[\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y}], 1 - \pi[\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y}]\}^2.$$

Contrastingly, recall that the median probability model $\boldsymbol{\gamma}^{MP} = (\boldsymbol{\gamma}_1^{MP\prime}, \boldsymbol{\gamma}_2^{MP\prime})$ is defined through

$$\gamma_i^{MP} = 1 \quad \text{iff} \quad \pi(\gamma_i = 1 \,|\, \boldsymbol{Y}) > 0.5 \quad \text{for} \quad i = 1, \ldots, p+k.$$

The median probability model $\boldsymbol{\gamma}^{MP}$ thus behaves as the optimal model $\boldsymbol{\gamma}^o$ for the first p variables. For the k duplicate copies, however, $\boldsymbol{\gamma}_2^{MP}$ consists of either *all ones or all zeros*. The MP rule correctly recognizes that the decision to include $\boldsymbol{x}$ is ultimately dichotomous: either all $\boldsymbol{x}$'s in or all $\boldsymbol{x}$'s out. Moreover, when the median model decides "all in", it *will be predictive optimal*. Indeed, $\pi(\gamma_{2i}^{MP} = 1 \,|\, \boldsymbol{Y}) > 1/2$ for $i \in \{1, \ldots, k\}$ implies $\pi(\boldsymbol{\gamma}_2 = \mathbf{0} \,|\, \boldsymbol{Y}) < 1/2$. The MP model will deviate from the optimal model only when $\pi(\gamma_{2i}^{MP} = 1 \,|\, \boldsymbol{Y}) < 1/2$ and $\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y}) > 1/2$ in which case

$$\frac{R(\boldsymbol{\gamma}^{MP}) - R(\boldsymbol{\gamma}^o)}{R(\boldsymbol{\gamma}^o)} = \frac{R_2(\boldsymbol{\gamma}_2^{MP}) - R_2(\boldsymbol{\gamma}_2^o)}{R(\boldsymbol{\gamma}^o)} = \frac{g^2 z^2 [1 - 2\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y})]}{(1+g)^2 [R_1(\boldsymbol{\gamma}_1^o) + R_2(\boldsymbol{\gamma}_2^o)]}.$$

The term $R_1(\boldsymbol{\gamma}_1^o)$ in (2.14) can be quite large when p is large, implying that the relative risk can be quite small. The MP model is thus not too far away from the optimal predictive model in this scenario.

Several conclusions can be drawn from our analysis of this toy example. First, Lemma 3 shows that, in the presence of perfect correlation, it is the *joint inclusion* rather than marginal inclusion probabilities that guide the optimal predictive model selection. Second, the clone variables ultimately act collectively as one variable, which has important implications on the assignment of prior model probabilities. We will elaborate on this important issue in Section 2.4, 2.5 and 2.6. Third, purely from a predictive point of view, all models in the model collective (including at least one $\boldsymbol{x}$) are equivalent. The g-prior here appears to be egalitarian in the sense that it (rightly) treats all these models equally. This property is not retained under independent product priors, as shown below.

Remark 5 (Independent Product Priors)**.** *Let us replace* (2.11) *with an independent prior covariance structure*

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim \mathcal{N}_{|\boldsymbol{\gamma}|}\left(\mathbf{0}, \sigma^2 g/n \; \mathrm{I}_{|\boldsymbol{\gamma}|}\right) \tag{2.16}$$

for $g = n$, which corresponds to the usual scaling when the predictors are centered and standardized so that $\|\boldsymbol{x}_i\|^2 = n$. The posterior mean $\bar{\boldsymbol{\beta}}$ then satisfies

$$H_{\boldsymbol{\gamma}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} - \bar{\boldsymbol{\beta}} = \begin{pmatrix} \frac{1}{n+1}\mathrm{diag}\{\boldsymbol{\gamma}_1 - \mathsf{E}[\boldsymbol{\gamma}_1 | \boldsymbol{Y}]\} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}\{\boldsymbol{\gamma}_2 - \mathsf{E}[\boldsymbol{\gamma}_2 | \boldsymbol{Y}]\} - \frac{n}{1+n|\boldsymbol{\gamma}_2|}\boldsymbol{\gamma}_2\boldsymbol{\gamma}_2' + \mathsf{E}\left[\frac{n}{1+n|\boldsymbol{\gamma}_2|}\boldsymbol{\gamma}_2\boldsymbol{\gamma}_2' \middle| \boldsymbol{Y}\right] \end{pmatrix} \boldsymbol{z}\,.$$

The criterion $R(\boldsymbol{\gamma}) = R_1^\star(\boldsymbol{\gamma}_1) + R_2^\star(\boldsymbol{\gamma}_2)$ *again separates into two parts, where*

$$R_1^\star(\boldsymbol{\gamma}_1) = \frac{1}{n+1}\sum_{i=1}^{p} z_i^2(\gamma_{1i} - \mathsf{E}\left[\gamma_{1i} \,|\, \boldsymbol{Y}\right])^2, \tag{2.17}$$

$$R_2^\star(\boldsymbol{\gamma}_2) = z^2\left[|\boldsymbol{\gamma}_2| - \mathsf{E}\left(|\boldsymbol{\gamma}_2| \,|\, \boldsymbol{Y}\right) - \frac{n|\boldsymbol{\gamma}_2|}{1+n|\boldsymbol{\gamma}_2|} + \mathsf{E}\left(\frac{n|\boldsymbol{\gamma}_2|}{1+n|\boldsymbol{\gamma}_2|}\,\middle|\,\boldsymbol{Y}\right)\right]^2. \tag{2.18}$$

The optimal predictive model for the last k *variables now has a bit less intuitive explanation. Denote with* $c(|\boldsymbol{\gamma}|) = |\boldsymbol{\gamma}| - \frac{n|\boldsymbol{\gamma}|}{1+n|\boldsymbol{\gamma}|}$*. The optimal predictive model now consists of any collection of variables of size* $|\boldsymbol{\gamma}_2^o|$ *for which* $c(|\boldsymbol{\gamma}_2^o|)$ *is as close as possible to the posterior mean of* $c(|\boldsymbol{\gamma}_2|)$*. It is worthwhile to note that this does not need to be the null or the full model. For instance,* $\boldsymbol{\gamma}_2^o = \mathbf{0}$ *when* $\mathsf{E}\left[c(|\boldsymbol{\gamma}_2|) \,|\, \boldsymbol{Y}\right] < 0.5/(1+n)$ *and* $\boldsymbol{\gamma}_2^o = \mathbf{1}$ *when* $c(|\boldsymbol{\gamma}_2|) > 0.5\mathsf{E}\left[c(k-1) + c(k) \,|\, \boldsymbol{Y}\right]$*. Besides these narrow situations, the optimal model* $\boldsymbol{\gamma}_2^o$ *will have a nontrivial size (other than* 0 *or* k*). The median probability model will still maintain the dichotomy by either including all or none of the* x*'s. However, contrary to the g-prior it is not guaranteed to be "optimal" when, for instance,* $\boldsymbol{\gamma}_2^{MP} = \mathbf{1}$*. It seems that the mission of the optimal model under the independent prior is a bit obscured. It is not obvious why models in the same model collective should be treated differentially and ranked based on their size. The independence prior correlation structure thus induces what seems as an arbitrary identifiability constraint.*

2.4 Prior Probabilities on Model Collectives

It has been now standard to assume that each model of dimension $|\boldsymbol{\gamma}|$ has an equal prior probability

$$\pi(\boldsymbol{\gamma}) = \pi(|\boldsymbol{\gamma}|)/\binom{p+k}{|\boldsymbol{\gamma}|}, \tag{2.19}$$

with $\pi(|\boldsymbol{\gamma}|)$ being the prior probability (usually $1/(p+k+1)$) of the collection of models of dimension $|\boldsymbol{\gamma}|$. One of the observations from Lemma 3 is that it is the *aggregate* posterior probability $\pi(\boldsymbol{\gamma}_2 \neq 0 \,|\, \boldsymbol{Y})$ rather than individual inclusion probabilities $\pi(\gamma_{2i} = 1 \,|\, \boldsymbol{Y})$ that drive the optimal predictive model $\boldsymbol{\gamma}_2^o$ in our collinear design. Thereby, it is natural to inspect the aggregate prior probability $\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0})$. We will be using the notion of model collectives introduced earlier in Definition 1. The number of models of size $j > |\boldsymbol{\gamma}_1|$ in the model collective $M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}$ is $\binom{k}{j-|\boldsymbol{\gamma}_1|}$, so that the prior probability of the model collective $M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}$ is

$$\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}) = \sum_{j=|\boldsymbol{\gamma}_1|+1}^{|\boldsymbol{\gamma}_1|+k} \frac{\pi(j)}{\binom{p+k}{j}}\binom{k}{j-|\boldsymbol{\gamma}_1|}. \tag{2.20}$$

We investigate the prior probability of the model collective under two usual choices: fixed prior inclusion probability (the separable case) and the random (non-separable) case.

The Separable Case Suppose that all variables have a known and equal prior inclusion probability $\theta = \pi(\gamma_j = 1 \,|\, \theta)$ for $j = 1, \ldots, p+k$. Then the probability of the model aggregate, given θ, is

$$\begin{aligned}\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}; \theta) &= \theta^{|\boldsymbol{\gamma}_1|}(1-\theta)^{p-|\boldsymbol{\gamma}_1|} \sum_{j=1}^{k} \theta^j (1-\theta)^{k-j} \binom{k}{j} \\ &= \theta^{|\boldsymbol{\gamma}_1|}(1-\theta)^{p-|\boldsymbol{\gamma}_1|} \left[1-(1-\theta)^k\right]\end{aligned}$$

and the prior probability of the "null model" $M_{\boldsymbol{\gamma}_1,\mathbf{0}}$ (not including any of the correlated variables) is

$$\pi(M_{\boldsymbol{\gamma}_1,\mathbf{0}}; \theta) = \theta^{|\boldsymbol{\gamma}_1|}(1-\theta)^{p-|\boldsymbol{\gamma}_1|}(1-\theta)^k.$$

The ratio satisfies

$$\frac{\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}; \theta)}{\pi(M_{\boldsymbol{\gamma}_1,\mathbf{0}}; \theta)} = \left[\left(\frac{1}{1-\theta}\right)^k - 1\right]. \tag{2.21}$$

This analysis reveals a rather spurious property of the separable prior: regardless of the choice $\theta \in (0,1)$, the model aggregate $M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}$ will always have a higher prior probability than the model $M_{\boldsymbol{\gamma}_1,\mathbf{0}}$ without any $\boldsymbol{x}$ in it. Such a preferential treatment for $\boldsymbol{x}$ is generally unwanted. We illustrate this issue with the uniform model prior (obtained with $\theta = 0.5$) which is still widely used in practice.

With fixed $\theta = 0.5$, all models have an equal prior probability of $2^{-(p+k)}$. The number of models in the collective $M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}$ is $2^k - 1$, and so

$$\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}; 1/2) = (2^k-1)2^{-(p+k)} = (2^k-1)\,\pi(M_{\boldsymbol{\gamma}_1,\mathbf{0}}; 1/2). \tag{2.22}$$

The collective can thus have much more prior probability than $M_{\boldsymbol{\gamma}_1,\mathbf{0}}$. Furthermore, the marginal prior probability of inclusion of $\boldsymbol{x}$ is $\sum_{\boldsymbol{\gamma}_1} \pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}; 1/2) = 1 - 2^{-k}$. Hence, if k is even moderately large, the prior mass is concentrated on the models which include $\boldsymbol{x}$ as a covariate, and the posterior mass will almost certainly also be concentrated on those models. The model-averaged $\bar{\boldsymbol{\beta}}$ will reflect this, essentially only including models that have $\boldsymbol{x}$ as a covariate.

Beta-Binomial Prior It is generally acknowledged (Cui and George, 2008; Ley and Steel, 2009; Scott and Berger, 2010) that assigning equal prior probability to all models is a poor choice, since it does not adjust for the multiple testing that is effectively being done in variable selection. The common alternative (which does adjust for multiple testing), is replace the separable prior with $\theta \sim \mathcal{B}(a,b)$. Then the prior probability of the model aggregate satisfies

$$\begin{aligned}\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}) &= \int_0^1 \pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}; \theta)\mathrm{d}\,\pi(\theta) = \int_0^1 \theta^{|\boldsymbol{\gamma}_1|+a-1}(1-\theta)^{p-|\boldsymbol{\gamma}_1|+b-1}\left[1-(1-\theta)^k\right]\mathrm{d}\,\theta \\ &= \mathcal{B}(|\boldsymbol{\gamma}_1|+a, p-|\boldsymbol{\gamma}_1|+b) - \mathcal{B}(|\boldsymbol{\gamma}_1|+a, p+k-|\boldsymbol{\gamma}_1|+b) \\ &= \mathcal{B}(|\boldsymbol{\gamma}_1|+a, p+k-|\boldsymbol{\gamma}_1|+b)\left[\prod_{j=1}^{k}\frac{a+b+p+j-1}{p+b-|\boldsymbol{\gamma}_1|+j-1} - 1\right]\end{aligned}$$

and

$$\pi(M_{\boldsymbol{\gamma}_1,\mathbf{0}}) = \mathcal{B}(|\boldsymbol{\gamma}_1| + a, p + k - |\boldsymbol{\gamma}_1| + b). \tag{2.23}$$

Then

$$\frac{\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}})}{\pi(M_{\boldsymbol{\gamma}_1,\mathbf{0}})} = \left[\prod_{j=1}^{k}\left(1 + \frac{|\boldsymbol{\gamma}_1| + a}{p + b - |\boldsymbol{\gamma}_1| + j - 1}\right) - 1\right]. \tag{2.24}$$

This ratio is guaranteed to be smaller than under the separable case with a fixed θ when $|\boldsymbol{\gamma}_1| < (p + a + b)\big(\theta - \frac{a}{a+b+p}\big)$. With the usual choice $a = b = 1$, the ratio in (2.24) will be smaller than the one in (2.21) when $|\boldsymbol{\gamma}_1|$ is smaller than $(p + 2)\theta - 1$, i.e. when the number $|\boldsymbol{\gamma}_1|$ of non-duplicated variables is roughly smaller than its expectation under the fixed prior case with a probability θ. This suggests that the beta-binomial prior can potentially cope better with variable redundancy. We elaborate on this point in the next section. In the forthcoming Lemma 5, we provide an approximation to (2.24) as p gets large.

2.5 Posterior Inclusion Probabilities

In the previous section, we have shown that equal prior model probabilities can be problematic because each model collective $M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}$ receives much more prior mass relative to $M_{\boldsymbol{\gamma}_1,\mathbf{0}}$, essentially forcing the inclusion of $\boldsymbol{x}$. Going further, we show how this is reflected in the posterior inclusion probabilities. We first focus on the case when σ^2 is known.

Lemma 4. *Consider the model (1.1), where $\boldsymbol{X}$ satisfies (2.10) and where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{x}$ are orthogonal with $\|\boldsymbol{x}_i\|^2 = \|\boldsymbol{x}\|^2 = n$. Denote by $z = \boldsymbol{y}'\boldsymbol{x}/\sqrt{n}$ and consider the prior (2.11) with $g = n$, known σ^2 and equal prior model probabilities $\pi(\boldsymbol{\gamma}) = 1/2^{p+k}$. Then we have*

$$\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y}) > 1/2 \quad \textit{iff} \quad z^2 > \log\left(\frac{\sqrt{1+n}}{2^k - 1}\right) 2\sigma^2 \left(1 + \frac{1}{n}\right). \tag{2.25}$$

Proof. We will be relying on the notation $m(\boldsymbol{y} \,|\, \boldsymbol{\gamma}_1, \boldsymbol{x}), m(\boldsymbol{y} \,|\, \boldsymbol{\gamma}_1)$ and $\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}_1)$ introduced in (2.4) and Lemma 2. Using the fact that $m(\boldsymbol{y} \,|\, \boldsymbol{\gamma}_1, \boldsymbol{x}) = m(\boldsymbol{y} \,|\, \boldsymbol{\gamma}_1)\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}_1)$ (from Lemma 2), the posterior probability of joint inclusion $\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y})$ (noting that $\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}})$ and $\pi(M_{\boldsymbol{\gamma}_1,\mathbf{0}})$ depend only on $|\boldsymbol{\gamma}_1|$) equals

$$\begin{aligned}
\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y}) &= \frac{\sum_{\boldsymbol{\gamma}_1} \pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}) m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1, \boldsymbol{x})}{\sum_{\boldsymbol{\gamma}_1} \pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}) m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1, \boldsymbol{x}) + \sum_{\boldsymbol{\gamma}_1} \pi(M_{\boldsymbol{\gamma}_1,\mathbf{0}}) m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1)} \\
&= \frac{\sum_{\boldsymbol{\gamma}_1} \pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}) m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1)\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}_1)}{\sum_{\boldsymbol{\gamma}_1} \pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}) m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1)\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}_1) + \sum_{\boldsymbol{\gamma}_1} \pi(M_{\boldsymbol{\gamma}_1,\mathbf{0}}) m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1)} \\
&= \frac{\psi(\boldsymbol{y}, \boldsymbol{x}) \sum_{i=0}^{p} \pi^{\star}_{i,\boldsymbol{x}}}{\sum_{i=0}^{p} \left[\psi(\boldsymbol{y}, \boldsymbol{x})\pi^{\star}_{i,\boldsymbol{x}} + \pi^{\star}_{i,\mathbf{0}}\right]},
\end{aligned} \tag{2.26}$$

with

$$\pi^{\star}_{i,\boldsymbol{x}} = \sum_{\boldsymbol{\gamma}_1 : |\boldsymbol{\gamma}_1| = i} m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1)\ \pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}) \quad \text{and} \quad \pi^{\star}_{i,\mathbf{0}} = \sum_{\boldsymbol{\gamma}_1 : |\boldsymbol{\gamma}_1| = i} m(\boldsymbol{y} \mid \boldsymbol{\gamma}_1)\ \pi(M_{\boldsymbol{\gamma}_1,\mathbf{0}}), \tag{2.27}$$

and where

$$\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\gamma}_1) = \psi(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{\sqrt{1+g}} \exp\left\{ \frac{g}{2\sigma^2(1+g)} \times \left[\boldsymbol{y}'\boldsymbol{x}\right]^2 / n \right\}$$

was introduced earlier in (2.7). Note that this quantity does not depend on $\boldsymbol{\gamma}_1$ since we assumed that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ and $\boldsymbol{x}$ are orthogonal. When all models have equal prior probabilities, we can evoke the identity (2.22) to find that

$$\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y}) = \frac{(2^k - 1)\psi(\boldsymbol{y}, \boldsymbol{x})}{1 + (2^k - 1)\psi(\boldsymbol{y}, \boldsymbol{x})} \,.$$

With the usual choice $g = n$, it follows that $\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y}) > 0.5$ if and only if $z^2 > \log\left(\frac{\sqrt{1+n}}{2^k - 1}\right) 2\sigma^2(1 + \frac{1}{n})$. □

From Lemma 4 it follows that the optimal predictive model (characterized in Lemma 3) will include $\boldsymbol{x}$ if the number of duplicates k is large enough, even when $\boldsymbol{x}$ has a small effect (z is small). Thus, the choice of equal prior model probabilities for optimal predictive model, in the face of replicate covariates, is potentially quite problematic. If one is only developing a model for prediction in such a situation, such forced inclusion of $\boldsymbol{x}$ is probably suboptimal, but it is only one covariate and so will not typically have a large effect, unless only very small models have significant posterior probability. For prediction, one could presumably do somewhat better by only considering the first $p+1$ variables in the model uncertainty problem, finding the model averaged $\bar{\boldsymbol{\beta}}$ for this subset of variables.

This statement at first seems odd, because we 'know' the model averaged answer in the original problem is optimal from a Bayesian perspective. But that optimality is from the internal Bayesian perspective, assuming we believe that the original model space and assignment of prior probabilities is correct. If we really believed – e.g., that any of k highly correlated genes could be in the model with prior inclusion probabilities each equal to 1/2 (equivalent to the assumption that all models have equal prior probability) – then the original model averaged answer would be correct and we should include $\boldsymbol{x}$ in the prediction. At the other extreme, if we felt that only the collection of all k genes has prior inclusion probability of 1/2, then the result will be like the model averaged answer for the first $p + 1$ variables.

To get some feel for things in the general case (non-uniform model prior), suppose $\pi^\star_{i,\boldsymbol{x}}$ for some $0 \leq i \leq p$ is much bigger than the others, so that (2.26) becomes

$$\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \,|\, \boldsymbol{Y}) \approx \frac{\pi^\star_{i,\boldsymbol{x}} \psi(\boldsymbol{y}, \boldsymbol{x})}{\pi^\star_{i,\boldsymbol{x}} \psi(\boldsymbol{y}, \boldsymbol{x}) + \pi^\star_{i,\mathbf{0}}} \,.$$

Using (2.24), it is immediate that this is bigger than 0.5 if

$$1 < \frac{\pi^\star_{i,\boldsymbol{x}}}{\pi^\star_{i,\mathbf{0}}} \psi(\boldsymbol{y}, \boldsymbol{x}) = \left[\prod_{j=1}^{k} \left(1 + \frac{i + a}{p + b - i + j - 1} \right) - 1 \right] \psi(\boldsymbol{y}, \boldsymbol{x}) \,. \tag{2.28}$$

The following Lemma characterizes the behavior of $\frac{\pi^\star_{i,\boldsymbol{x}}}{\pi^\star_{i,\mathbf{0}}}$ when p gets large.

Lemma 5. *Suppose a and b are integers. As p gets large with i fixed,*

$$\prod_{j=1}^{k}\left(1+\frac{i+a}{p+b-i+j-1}\right)=\left(1+\frac{k}{p}\right)^{i+a}\left(1+\frac{Ck}{p(p+k)}\right)\left(1+O\left(\frac{1}{p^2}\right)\right),$$

where $C=-(i+a)[b-1-i+(i+a+1)/2]$ ($C=(i^2-i-2)/2$ if $a=b=1$). To first order,

$$\prod_{j=1}^{k}\left(1+\frac{i+a}{p+b-i+j-1}\right)=\left(1+\frac{k}{p}\right)^{i+a}\left(1+O\left(\frac{1}{p}\right)\right).$$

Proof. Defining $d=b-1$ and $c=d+a$,

$$\begin{aligned}\prod_{j=1}^{k}\left(1+\frac{i+a}{p+b-i+j-1}\right) &= \prod_{j=1}^{k}\frac{p+c+j}{p+d-i+j}=\frac{(p+c+k)!/(p+c)!}{(p+d+k-i)!/(p+d-i)!}\\ &= \frac{(p+c+k)!/(p+d+k-i)!}{(p+c)!/(p+d-i)!}=\prod_{j=1}^{i+a}\frac{p+d+k-i+j}{p+d-i+j}\\ &= \left(\frac{p+k}{p}\right)^{i+a}\prod_{j=1}^{i+a}\frac{\left(1+\frac{d-i+j}{p+k}\right)}{\left(1+\frac{d-i+j}{p}\right)}.\end{aligned}$$

The first order result follows immediately and the second order result follows from expanding the products in the last term above. □

Utilization of the first order term in (2.28), and again choosing $g=n$ and assuming $\|\boldsymbol{x}\|=\sqrt{n}$, yields that the collective has posterior inclusion probability greater than 0.5 if

$$z^2>\log\left(\frac{\sqrt{1+n}}{(1+k/p)^{[i+a]}-1}\right)2\sigma^2\left(1+\frac{1}{n}\right).$$

Note that this is much less likely to be satisfied than (2.25), when k grows, since $(1+k/p)^{[i+a]}$ is then much smaller than 2^k; thus having many duplicate $\boldsymbol{x}$'s does not ensure that $\boldsymbol{x}$ will be included in the model, as it was in the equal model probability case.

Remark 6 (The Case of Unknown Variance)**.** *Lemma 4 was postulated for the simpler case when the variance σ^2 is known. We have seen in Remark 3 that the inflation factor $\psi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\gamma}_1)$ depends on $\boldsymbol{\gamma}_1$ even when $\boldsymbol{x}$ and $\boldsymbol{B}$ are orthogonal, which complicates the analysis. Sufficient characterizations can be obtained from*

$$\frac{\psi_{min}(2^k-1)}{1+(2^k-1)\psi_{max}}\le\pi(\boldsymbol{\gamma}_2\neq 0\,|\,\boldsymbol{Y})\le\frac{\psi_{max}(2^k-1)}{1+(2^k-1)\psi_{min}},$$

where $\psi_{min}=\min_{\boldsymbol{\gamma}_1}\psi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\gamma}_1)$ and $\psi_{max}=\max_{\boldsymbol{\gamma}_1}\psi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\gamma}_1)$.

While the assumption of having duplicated predictors (used throughout this section) is somewhat theoretical, it sheds light on the most extreme case of correlation where one would expect the median probability model to fail. The performance of the median probability model in the more optimistic and realistic scenarios of near-correlation is shown through simulations in Section 3.4.

2.6 The Dilution Problem

Sets of predictors which are highly correlated with each other become proxies for one another in our linear model (1). This quickly leads to an excess of redundant models, each of which is distinguished only by including a different subset of these. To prevent such redundant models from accumulating too much posterior probability, dilution priors may be considered (George, 2010). Such priors downweigh individual model probabilities based on their proximity to one another, and a variety of strategies to do this may be considered.

When faced with a single identifiable cluster of highly correlated predictors such as our k clones of $\boldsymbol{x}$, a simple dilution strategy would be to first assign a reasonable amount of prior mass to the entire cluster, and then dilute this mass uniformly across all subset models within this cluster. More precisely, to smear out the prior aggregation on $M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}$, one might like to consider different inclusion probabilities. Let $[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p]$ have a prior inclusion probability θ_1 and each of the $\boldsymbol{x}$ clones have a prior inclusion probability θ_2. With

$$\theta_2 = 1-(1-\theta_1)^{1/k} \tag{2.29}$$

we have

$$\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}}) = \theta_1^{|\boldsymbol{\gamma}_1|}(1-\theta_1)^{p-|\boldsymbol{\gamma}_1|}\left[1-(1-\theta_2)^k\right] = \theta_1^{|\boldsymbol{\gamma}_1|+1}(1-\theta_1)^{p-|\boldsymbol{\gamma}_1|} \tag{2.30}$$

and

$$\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{0}}) = \theta_1^{|\boldsymbol{\gamma}_1|}(1-\theta_1)^{p-|\boldsymbol{\gamma}_1|+1}.$$

Assuming (2.29), variables with correlated copies have smaller inclusion probabilities (the more copies, the smaller the probability). This may correct the imbalance between $\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{x}})$ and $\pi(M_{\boldsymbol{\gamma}_1,\boldsymbol{0}})$ by treating the multiple copies of $\boldsymbol{x}$ essentially as one variable. This prior allocation would put $\boldsymbol{x}$ on an equal footing with $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p$ in the optimal predictive model rule (based on $\pi(\boldsymbol{\gamma}_2 \neq \boldsymbol{0}\,|\,\boldsymbol{Y})$), but would disadvantage $\boldsymbol{x}$ in the median probability model. From our considerations above, it would seem that there is a fix to the dilution problem in our synthetic example (with clone $\boldsymbol{x}$'s). However, general recommendations for other correlation patterns are far less clear.

3 The Case of Two Covariates

3.1 The Geometric Representation

The situations analyzed in previous sections may also be considered from a geometric perspective. Using the definition of $\boldsymbol{H}_{\boldsymbol{\gamma}}$ from (1.3), we define

$$\boldsymbol{\alpha}_{\boldsymbol{\gamma}} = \boldsymbol{X}\boldsymbol{H}_{\boldsymbol{\gamma}}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} \tag{3.1}$$

and denote with $\bar{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}) \boldsymbol{X} \boldsymbol{H}_{\boldsymbol{\gamma}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ the overall posterior mean. Under this transformation, the expected posterior loss (1.2) to be minimized may be written as

$$R(\boldsymbol{\gamma}) = (\boldsymbol{\alpha}_{\boldsymbol{\gamma}} - \bar{\boldsymbol{\alpha}})' (\boldsymbol{\alpha}_{\boldsymbol{\gamma}} - \bar{\boldsymbol{\alpha}}) .$$

This implies that the preferred model will be the one whose corresponding $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$ is nearest to $\bar{\boldsymbol{\alpha}}$ in terms of the Euclidean distance.

To geometrically formulate the predictive problem, each model $M_{\boldsymbol{\gamma}}$ may be represented by the point $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$ and the set of models becomes a collection of points in p-dimensional space. The convex hull of these points is a polygon representing the set of possible model averaged estimates $\bar{\boldsymbol{\alpha}}$, as the $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ vary over their range. Any point in this polygon is a possible optimal predictive model, depending on $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$'s. The goal is to geometrically characterize when each single model is optimal, given that a single model must be used.

In what follows we will refer to circumstances where $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is equal (up to a constant) to $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\boldsymbol{X}'_{\boldsymbol{\gamma}} \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1} \boldsymbol{X}'_{\boldsymbol{\gamma}} \boldsymbol{Y}$, as it happens under a non-informative prior or a g-prior on model parameters (see Barbieri and Berger (2004) for a discussion on the use of mixed default strategies to determine the posterior probability of each model and the posterior distribution of the parameters). In this context $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$ is (proportional to) the projection of $\boldsymbol{Y}$ onto the space spanned by the columns of $\mathbf{X}_{\boldsymbol{\gamma}}$.

Consider the simple situation in which we have two covariates x_1 and x_2 and four possible models:

$$M_{10} : \{x_1\} \qquad M_{01} : \{x_2\} \qquad M_{11} : \{x_1, x_2\},$$

and the null model M_{00}. These can be represented as four points in the plane.

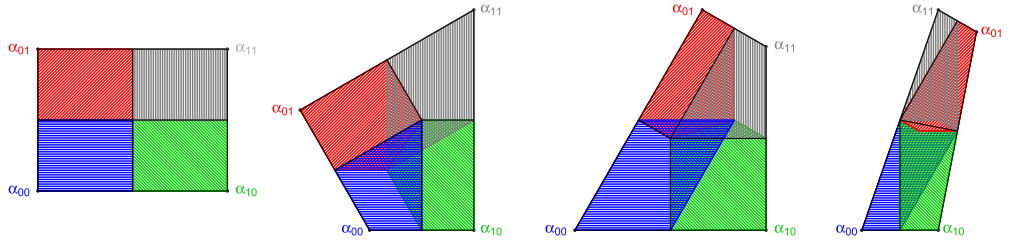
Depending on the sample correlation structure, the polygon region, whose vertices are $\boldsymbol{\alpha}_{00}$, $\boldsymbol{\alpha}_{10}$, $\boldsymbol{\alpha}_{01}$ and $\boldsymbol{\alpha}_{11}$ (i.e. the convex hull of all possible posterior means $\bar{\boldsymbol{\alpha}}$), can have four distinct forms. Each situation may be characterized in terms of the correlations between the variables involved, as summarized in Table 1, where $r_{12} = Corr(x_1, x_2)$, $r_{1y} = Corr(x_1, Y)$ and $r_{2y} = Corr(x_2, Y)$.

<table>
<tr><td>$r_{12} = 0$</td><td>$r_{12}\frac{r_{1y}}{r_{2y}} < 0$</td><td colspan="2">$r_{12}\frac{r_{1y}}{r_{2y}} > 0$</td></tr>
<tr><td rowspan="2">orthogonal</td><td rowspan="2">case 1</td><td>$|r_{12}| < \min\{|\frac{r_{1y}}{r_{2y}}|, |\frac{r_{2y}}{r_{1y}}|\}$</td><td>$|\frac{r_{1y}}{r_{2y}}| < |r_{12}|$</td></tr>
<tr><td>case 2</td><td>case 3</td></tr>
</table>

Table 1: Possible scenarios in terms of r_{12} and the ratio $\frac{r_{1y}}{r_{2y}}$.

In Figure 1 the four forms are plotted for the case $|r_{12}| = 0.5$. (Ignore the colors for now.) In particular, the values of the correlations here are:

Case 1	$r_{12} = -0.5$	$r_{1y} = 0.3$	$r_{2y} = 0.4$
Case 2	$r_{12} = 0.5$	$r_{1y} = 0.3$	$r_{2y} = 0.4$
Case 3	$r_{12} = 0.5$	$r_{1y} = 0.1$	$r_{2y} = 0.3$



(a) Orthogonal (b) Case 1 (c) Case 2 (d) Case 3

Figure 1: Four possible scenarios for the graphical representation of predictive model selection from among $\gamma \in \{(0,0)', (1,0)', (0,1)', (1,1)'\}$. The coordinate axes correspond to $\alpha_{\boldsymbol{\gamma}} \in \mathbb{R}^2$ in (3.1), where $\alpha_{00} = (0,0)'$.

The angles $\alpha_{00}\alpha_{10}\alpha_{11}$ and $\alpha_{00}\alpha_{01}\alpha_{11}$ are always right angles, since $(\alpha_{10}-\alpha_{00})'(\alpha_{11}-\alpha_{10}) = 0$ and $(\alpha_{01}-\alpha_{00})'(\alpha_{11}-\alpha_{01}) = 0$ [the projection of $\boldsymbol{\alpha}_{11}$ on the line spanned by $\boldsymbol{\alpha}_{10}$ is $\boldsymbol{\alpha}_{10}$ itself and similarly for $\boldsymbol{\alpha}_{01}$].

The solid lines divide the figures into the four optimality subregions associated with the four models, namely the sets of those $\bar{\boldsymbol{\alpha}}$ which are closer to one of the $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$.

The colors in Figure 1 indicate the regions where the average model point (the best model averaged answer) could lie if the model with the corresponding color is the median posterior probability. In the orthogonal case, the model averaged answer and model optimality regions always coincide, i.e., the MPM is always optimal. In the other cases, this need not be so. In Case 1, for instance, the red region extends into the (blue) null model's optimality region; thus M_{01} could be the MPM, even when the null model is optimal. Likewise the green region extends into the optimality region of the null model, and the grey region (corresponding to the full model) extends into the optimality regions of all other models. Only the null model is fine here; if the null model is the MPM, it is guaranteed to be optimal.

3.2 Characterizations of the Optimal Model

For the case of two correlated covariates, we obtained partial characterizations of the optimal predictive model and the median probability model. These are summarized by the "mini-theorems" below, whose proofs can be found in the Supplementary Material (Barbieri et al., 2020).

Theorem 1 ("Mini-Theorems"). *Consider the model (1.1) with $q = 2$ and the three cases described in Table 1. Then the following statements hold.*

1. *In Case 1, if M_{00} is the median, it is optimal.*
2. *In Case 2, if M_{11} is the median, it is optimal.*

3. *In Case 1 and 3, if at most one variable has posterior inclusion probability larger than* $1/2$*,* M_{11} *cannot be optimal.*

4. *In Case 2 and 3, if at least one variable has posterior inclusion probability larger than* $1/2$*,* M_{00} *cannot be optimal.*

5. *In Cases 1 and 2, if* M_{00} *or* M_{11} *has posterior probability larger than* 0.5*, it is optimal.*

6. *In any case, if* M_{00} *or* M_{11} *has posterior probability larger than* 0.5*, the other cannot be optimal.*

7. *In Case 3, if* M_{00} *or* M_{11} *has posterior probability smaller than* 0.5 *it cannot be optimal.*

The motivation for developing these mini-theorems was to generate possible theorems that might hold in general. Unfortunately, in going to the three-dimensional problem, we were able to develop counterexamples (not shown here) to each of the mini-theorems.

3.3 Numerical Study of the Performance of the MPM

We present a numerical study that investigates the extent to which the MPM and MAP agree, and how often they differ from the optimal predictive model. The goal was to devise a study that effectively spans the entire range of correlations that are possible and this was easiest to do by limiting the study to the two-dimensional case. We considered (1) equal prior probabilities for the four models, (2) the unit information g-prior for the parameters and (3) the more realistic scenario where the variance σ^2 is unknown and assigned the usual objective prior $1/\sigma^2$.

The study considered the following correlations and sample sizes:

- r_{12} varies over the grid $\{-0.9, -0.8, -0.7, -0.6, -0.5, -0.4, -0.3, -0.2, -0.1, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. ($r_{12} = 0$ was not considered because the MPM is guaranteed to be optimal.)

- r_{1y} and r_{2y} vary over ranges meant to span the range of likely data under either the full model, one-variable model, or null model; the description (and derivation) of the various correlation ranges is given in the Supplementary Material.

- Sample sizes $n = 10, 50$ and 100 are considered.

The reason the numerical study is conducted in this way is to reduce the dimensionality of the problem. In terms of ordinary inputs, one would have to deal with a study over the space of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and the random error vector $\boldsymbol{\varepsilon}$ (or $\boldsymbol{Y}$). But, because the predictive Bayes risks only depend on r_{12}, r_{1y}, r_{2y} and n, we can reduce the study to a three dimensional problem. And, since these are simply correlations, we can

	number of cases	MPM=MAP both=OP (a) %	MPM=MAP both≠OP (b) %	MPM=OP MAP≠OP (c) %	MAP=OP MPM≠OP (d) %	MAP>MPM both≠OP (e) %	MPM>MAP both≠OP (f) %	GM $\frac{R(MPM)}{R(OP)}$	GM $\frac{R(MAP)}{R(OP)}$
Cases combined: Full model scenario									
n=10	534	75.7	17.4	5.1	0.6	0.7*	0.6*	1.078	1.110
n=50	534	94.6	1.5	2.4	0.0	1.5*	0.0	1.017	1.056
n=100	534	95.9	1.5	2.6	0.0	0.0	0.0	1.006	1.058
Overall	1602	88.7	6.8	3.4	0.2	0.7*	0.2*	1.033	1.074
Cases combined: $\beta_1 = 0$ and $\beta_2 \neq 0$ scenario									
n=10	620	68.4	27.3	2.9	0.3	0.5*	0.6*	1.030	1.032
n=50	731	90.4	7.8	0.4	0.1	1.2	0.0	1.019	1.017
n=100	749	91.1	8.7	0.1	0.0	0.0	0.1*	1.016	1.016
Overall	2100	84.1	13.9	1.1	0.1	0.6	0.2*	1.021	1.022
Cases combined: Null model scenario									
n=10	719	65.9	24.8	7.0	0.3	1.1	1.0*	1.032	1.036
n=50	799	85.5	12.3	1.6	0.4	0.0	0.3*	1.013	1.015
n=100	805	91.7	7.5	0.6	0.1	0.0	0.1	1.008	1.008
Overall	2323	81.6	14.5	2.9	0.3	0.3	0.4	1.018	1.020

Table 2: The case of two covariates: performance of MPM and MAP under the full, one-variable and null models.

Legend: columns (a) to (f) contain percentages of cases, over combinations of different values of the correlations among variables; OP denotes the optimal predictive model; MPM>MAP (resp. MAP>MPM) means that MPM (resp. MAP) has a smaller value of risk defined in (1.2) than MAP (resp. MPM); GM is the geometric mean of relative risks (to the optimal model) when MPM or MAP is not optimal. * denotes cases when OP is the *lowest* probability model.

choose a grid of values for each that essentially spans the space of possibilities in the 5-dimensional problem. The details of this are given in the Supplementary Material.

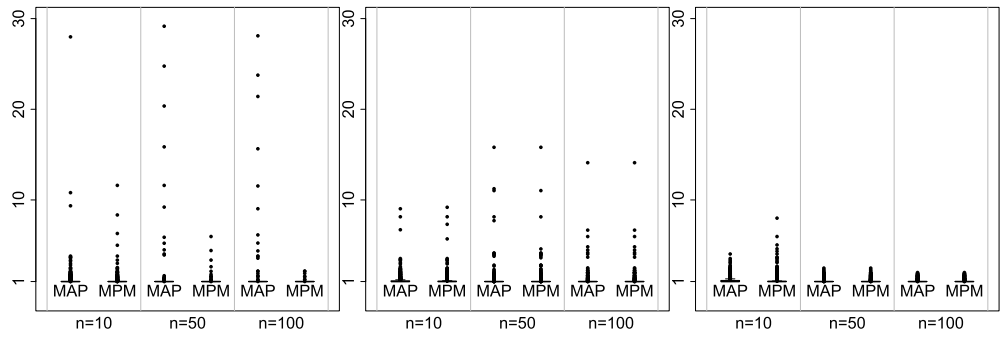
Table 2 summarizes the results, combining the three cases from Table 1, under each of the model correlation scenarios (full, one-variable, and null), i.e. the overall combination of values of r_{12}, r_{1y} and r_{2y} considered. We have 1602, 2100 and 2323 cases for the full, one-variable and null models, respectively. The table reports the percentage of times the MPM and MAP models equal the optimal predictive model (denoted with OP), i.e. the model minimizing (1.2). The last two columns of the table contain the geometric averages of relative risks to the optimal predictive model when MPM or MAP are not optimal, while graphs in Figure 2 depict box-plots of the corresponding distributions.

Here are some observations from Table 2:

- Simpler models are more challenging for the MPM (and MAP); indeed, MPM = OP in 92.1% (1421+54 out of 1602), 85.2% (1767+22 out of 2100) and 84.5% (1895+68 out of 2323) of the cases for the full, one-variable, and null model, respectively; still, these are high success rates, given that correlations vary over the full feasible spectrum.

- As would be expected, both the MPM and MAP do better with larger sample sizes.

- The vast majority of the time, the MPM and MAP are the same model but, when they differ, the MPM is typically better:



(a) Full model scenario (b) $\beta_1 = 0$ and $\beta_2 \neq 0$ scenario (c) Null model scenario

Figure 2: The case of two covariates: boxplots of the relative risks (to the optimal model) when MAP or MPM is not optimal under the full, one-variable and null models (cases combined). Dots correspond to values more extreme than 1.5 times the interquartile range from the box.

  - Combining all three model types (full, one-variable and null), the MPM does better than the MAP (from the MPM = OP $\neq$ MAP and MPM > MAP columns) in 2.7% of the cases (i.e. there are 162 cases MPM = OP $\neq$ MAP or MPM > MAP out of $6025 = 1602 + 2100 + 2323$ total cases); while the MAP does better than the MPM in 44 out of 6025 total cases (0.7%).
  - When the MPM and MAP are not optimal, the MAP may present values of the risk (relative to that of OP) more extreme than the MPM (see Figure 2); the geometric average of the MAP relative risk is higher than the geometric average for the MPM.

- In the cases denoted with a star *, the optimal model OP is, curiously, the *lowest* probability model.

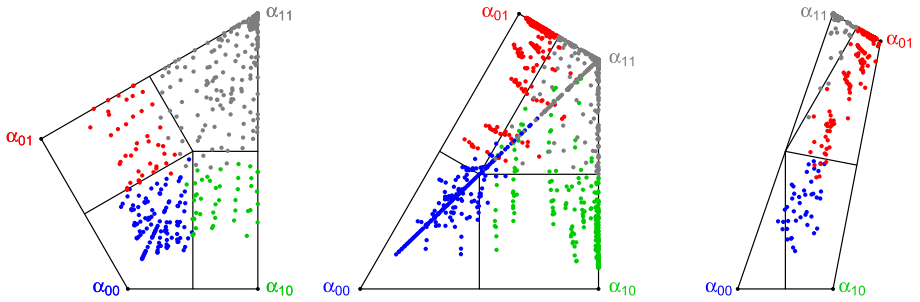
We may have some insight on the role of correlation between covariates through Table 3, which contains summaries under different values of $r_{12} = Corr(x_1, x_2)$, after combining all cases and model scenarios (full, one variable and null). When the correlation between x_1 and x_2 is small, MPM is virtually always optimal (MPM = OP in 98.1% of the cases when $r_{12} = 0.1$). While, as correlation increases, the rate of success degrades (MPM = OP in 75% of the cases when $r_{12} = 0.9$). However the deterioration in the performance of MPM appears slower than that of MAP.

Tables S.2, S.3 and S.4, in the Supplementary material, summarize some additional features of the numerical study. Note that the last sub-tables of Tables S.2–S.4 are included in Table 2. Each of those tables considers one model scenario (either the full model, the one-variable model, or the null model) and presents the results *separately* for the Case 1, Case 2, and Case 3. It is very clear from these tables that the Case 1 scenario is very favorable for the MPM – it is then virtually always the optimal model – while, in Cases 2 and 3, the MPM fails to be the optimal model in roughly 12% of

$\lvert r_{12}\rvert$	number of cases	MPM=MAP both=OP (a) %	MPM=MAP both≠OP (b) %	MPM=OP MAP≠OP (c) %	MAP=OP MPM≠OP (d) %	MAP>MPM both≠OP (e) %	MPM>MAP both≠OP (f) %
0.1	743	96.6	1.6	1.5	0.3	0.0	0.0
0.2	740	93.7	4.3	1.5	0.4	0.1	0.0
0.3	730	91.0	6.9	1.8	0.1	0.1	0.1
0.4	716	87.3	9.6	2.4	0.4	0.1	0.1
0.5	680	82.1	14.0	2.9	0.3	0.4	0.3
0.6	676	80.8	16.1	2.1	0.0	0.7	0.3
0.7	641	75.8	19.8	3.0	0.0	1.1	0.3
0.8	594	73.0	21.9	3.2	0.0	1.2	0.7
0.9	505	71.1	22.2	3.9	0.2	1.4	1.2
Overall	6025	84.4	12.2	2.4	0.2	0.5	0.3

Table 3: The case of two covariates: performance of MPM and MAP models under different values of $r_{12} = Corr(x_1, x_2)$.

Legend: columns (a) to (f) contain the percentage over combinations of all cases (1, 2 and 3), model scenarios (full, one variable and null), sample sizes ($n = 10, 50, 100$) and $r_{iy} = Corr(x_i, Y)$ ($i = 1, 2$); OP denotes the optimal predictive model; MPM>MAP (resp. MAP>MPM) means that MPM (resp. MAP) has a smaller value of risk defined in (1.2) than MAP (resp. MPM).

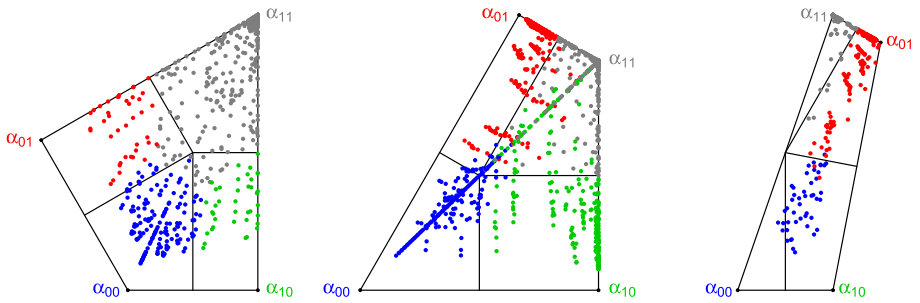


(a) Case 1 (b) Case 2 (c) Case 3

Figure 3: Results from the numerical study under the full model correlation scenario: each dot has the color of the MPM, with the MPM being optimal (or not) if it lies within (or outside) the quadrilateral with external vertex of the same color.

the cases (the proportion of $MPM \neq OP$ cases across the three model classes in Case 2 and 3). This is a useful result if one is in the two-variable situation, since it is easy to determine if one is in Case 1 or not. Alas, it is not known how to generalize this to larger dimensions.

Additional insight can be gained by looking at the nature of the 'failures' of the MPM and MAP. Figure 3, for the MPM, and Figure 4, for the MAP, show the errors being made, in the numerical study, for each of Case 1, Case 2 and Case 3, under the full model correlation scenario. Focusing on the MPM for explanation, the color of the dots in Figure 3 indicates which model was the median probability model; thus a blue



(a) Case 1 (b) Case 2 (c) Case 3

Figure 4: Results from the numerical study under the full model correlation scenario: each dot has the color of the MAP, with the MAP being optimal (or not) if it lies within (or outside) the quadrilateral with external vertex of the same color.

dot indicates that the median probability model was M_{00}, because that is the color of α_{00}. As before, the true optimal model for a dot is the external vertex defining the quadrilateral in which the dot lies; thus, if the blue dot lies within the quadrilateral with α_{00} as the external vertex, the MPM is the optimal model, while if the blue dot lies within the quadrilateral for which α_{10} is the external vertex, the MPM is incorrectly saying that M_{00} is optimal, when actually M_{10} is optimal.

The figures reinforce the earlier messages; Case 1 is nice for the MPM and MAP (almost all the colored dots are in the quadrilateral with the external vertex being of the same color), while Case 2 and, especially, Case 3 here are problematic – in Case 3, the MPM is typically M_{00} when M_{10} is optimal. Careful examination of the figures shows that the MPM is slightly better than the MAP, but the improvement is not dramatic.

The interesting feature revealed by the figures is that, essentially always, when the MPM and MAP fail, they do so by selecting a model of smaller dimension than the optimal model. There are a handful of dots going the other way, but they are hard to find. (This same feature was present in the corresponding figures for the one-variable and null model correlation scenarios, so those figures are omitted.) We highlight this feature because it potentially generalizes; if the MPM and MAP fail, they may typically do so by choosing too-small models.

The numerical study has also been implemented with three possible predictors (x_1, x_2, x_3) using the same ingredients, conveniently adapted. In particular we referred to the usual choice of prior distributions, sample sizes and grid of values of the correlations between covariates. Just as with two variables, we also considered the complete set of feasible correlations between the response and the explicative variables, under each model scenario (full, two variables, one variable and null). Since the results are substantially comparable to those in two dimensions, we only report a concise summary of the conclusions. Indeed more complex models and larger sample sizes are more

	number of cases	MPM=MAP both=OP (a) %	MPM=MAP both≠OP (b) %	MPM=OP MAP≠OP (c) %	MAP=OP MPM≠OP (d) %	MAP>MPM both≠OP (e) %	MPM>MAP both≠OP (f) %	GM $\frac{R(MPM)}{R(OP)}$	GM $\frac{R(MAP)}{R(OP)}$
$\mathbf{\lvert r_{12}\rvert = \lvert r_{13}\rvert = \lvert r_{23}\rvert = 0.1}$									
n=10	5299	86.2	5.8	6.2	1.3	0.1	0.4	1.003	1.010
n=50	5703	97.9	1.2	0.2	0.2	0.6	0.0	1.001	1.001
n=100	5587	98.8	1.0	0.2	0.0	0.0	0.0	1.000	1.000
Overall	16589	94.5	2.6	2.1	0.5	0.2	0.1	1.001	1.003
$\lvert r_{12}\rvert = \lvert r_{13}\rvert = \lvert r_{23}\rvert = 0.5$									
n=10	1908	52.8	36.7	5.8	0.7	2.8	1.2	1.149	1.160
n=50	2186	80.0	12.9	3.4	0.5	2.6	0.6	1.041	1.047
n=100	2203	86.5	9.1	2.0	0.4	1.2	0.8	1.028	1.038
Overall	6297	74.0	18.8	3.6	0.5	2.2	0.9	1.068	1.077
$\lvert r_{12}\rvert = \lvert r_{13}\rvert = \lvert r_{23}\rvert = 0.9$									
n=10	835	43.7	36.4	10.1	1.6	8.0	0.2	1.810	1.668
n=50	1380	63.0	19.6	9.0	0.1	8.0	0.3	1.401	1.365
n=100	1550	71.9	15.2	6.8	0.0	5.7	0.4	1.340	1.317
Overall	3765	62.4	21.5	8.3	0.4	7.1	0.3	1.456	1.406
Correlations combined									
n=10	1271624	50.5	37.4	5.8	0.8	2.5	3	1.190	1.214
n=50	1583299	78.3	15.9	2.3	0.7	1.9	0.9	1.082	1.092
n=100	1628570	84.6	12.3	1.7	0.3	0.6	0.5	1.058	1.075
Overall	4483493	72.7	20.7	3.1	0.6	1.6	1.3	1.103	1.625

Table 4: The case of three covariates: performance of MPM and MAP models under different values of $r_{ij} = Corr(x_i, x_j)$ $(i, j = 1, 2, 3)$.

Legend: columns (a) to (f) contain the percentage over combinations of all model scenarios (full, two variables, one variable and null), sample sizes ($n = 10, 50, 100$) and all feasible values of the correlations between the response variable and each covariate; *Correlations combined* refers to the complete set of the correlations between the covariates; OP denotes the optimal predictive model; MPM>MAP (resp. MAP>MPM) means that MPM (resp. MAP) has a smaller value of risk defined in (1.2) than MAP (resp. MPM); GM is the geometric mean of relative risks (to the optimal model) when MPM or MAP is not optimal.

suitable for MPM and MAP: combining all correlations, under the full model and with $n = 100$ MPM is the optimal model in 93.8% out of the 196 198 cases (MAP in 92%), while under the null model and with $n = 10$ MPM = OP in 52.4% (MAP in 47%) out of the 198 043 cases. Table 4 provides further hints on the effect of correlation between covariates. Combining all other ingredients of the study, when correlation is low MPM and MAP are almost always both equal to the optimal model. Higher correlation is more challenging for MPM and MAP: when covariates are equicorrelated and the common value of the correlation is 0.9, MPM and MAP are the same model in 84% of the cases and both optimal in 62%; when they differ, MPM is the optimal model in almost half of the cases, in the others MAP, although not optimal, is preferable to MPM in terms of risk. However on average the relative risks to the optimal model of MPM and MAP (when neither is optimal) do not differ much.

4 Generalizations of the Median Probability Model Optimality

In orthogonal designs, the primary condition for optimality of the median probability model is that $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$, the conditional posterior mean of $\boldsymbol{\beta}$ under $\boldsymbol{\gamma}$, is obtained by taking

the relevant coordinates of the posterior mean under the full model (condition (17) of Barbieri and Berger (2004)). With $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{D} = \text{diag}\{d_i\}_{i=1}^q$, the likelihood factors into independent likelihoods for each β_i and thereby *any* independent product prior

$$\pi(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma}) = \prod_{i=1}^{q} \pi_i(\beta_i)\,, \tag{4.1}$$

will satisfy the condition (17). This is a very important extension because priors that are fat-tailed are often recommended over sharp-tailed priors, such as the g-prior (for which the optimality results of the MPM were originally conceived).

Example 1 (Point-Mass Spike-and-Slab Priors)**.** *As an example of* (4.1)*, consider the point-mass mixture prior* $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma}) = \prod_{i=1}^{q}[\gamma_i \widetilde{\pi}_i(\beta_i) + (1-\gamma_i)\delta_0(\beta_i)]$*, where* $\widetilde{\pi}_i(\beta_i)$ *could be e.g. the unit-information Cauchy priors, as recommended by Jeffreys, or the non-local spike and slab priors (Rossell and Telesca, 2017).*

Example 2 (Continuous Spike-and-Slab Priors)**.** *The point-mass spike is not needed for the MPM to be optimal. Consider another example of* (4.1)*, the Gaussian mixture prior of George and McCulloch (1993):* $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma}) = \mathcal{N}_q(\boldsymbol{0}_q, \boldsymbol{V}_{\boldsymbol{\gamma}})$*, where* $\boldsymbol{V}_{\boldsymbol{\gamma}} = \text{diag}\{\gamma_i v_1 + (1-\gamma_i)v_0\}_{i=1}^q$ *with* $v_1 >> v_0$*. While the MPM was originally studied for point-mass spike-and-slab mixtures, it is optimal also under the continuous mixture priors. Indeed, to give an alternative argument, note that the posterior mean under a given model* $\boldsymbol{\gamma}$ *satisfies*

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{V}_{\boldsymbol{\gamma}}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \\ &= \text{diag}\left\{\frac{1}{d_i + v_0^{-1}}\right\}\boldsymbol{X}'\boldsymbol{Y} + \text{diag}\left\{\left(\frac{1}{d_i + v_1^{-1}} - \frac{1}{d_i + v_0^{-1}}\right)\gamma_i\right\}\boldsymbol{X}'\boldsymbol{Y}, \end{aligned}$$

where $\boldsymbol{V}_{\boldsymbol{\gamma}}^{-1} = \text{diag}\{\frac{\gamma_i}{v_1} + \frac{(1-\gamma_i)}{v_0}\}_{i=1}^q$*. Then the overall posterior mean vector appears to be*

$$\bar{\boldsymbol{\beta}} = \text{diag}\left\{\frac{1}{d_i + v_0^{-1}}\right\}\boldsymbol{X}'\boldsymbol{Y} + \text{diag}\left\{\left(\frac{1}{d_i + v_1^{-1}} - \frac{1}{d_i + v_0^{-1}}\right)\pi(\gamma_i = 1 \,|\, \boldsymbol{Y})\right\}\boldsymbol{X}'\boldsymbol{Y}.$$

The criterion $R(\boldsymbol{\gamma})$ *in* (1.2) *can be then written as*

$$R(\boldsymbol{\gamma}) = \sum_{i=1}^{q}\left(\frac{1}{d_i + v_1^{-1}} - \frac{1}{d_i + v_0^{-1}}\right)^2 (\gamma_i - \pi(\gamma_i = 1 \,|\, \boldsymbol{Y}))^2\, d_i z_i^2,$$

which easily seen to be minimized by the MPM model.

Barbieri and Berger (2004) show that the MPM is optimal also for correlated regressors, when considering a nested sequence of linear models: $M_{\boldsymbol{\gamma}(j)}$ $(j = 0, \ldots, q)$, where the first j elements of the index set $\boldsymbol{\gamma}(j)$ are equal to one and the last $q - j$ to zero. Again, one of the sufficient conditions is that the posterior mean $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is obtained by taking the relevant coordinates of the posterior mean under the full model, which holds under, e.g., the g-prior. Here, we generalize the class of priors under which MPM

is optimal for nested correlation designs. Assume $q < n$ and denote with $\boldsymbol{A}$ the upper Cholesky triangular matrix such that $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{A}'\boldsymbol{A}$. Then transform the linear model to

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \\ &= (\boldsymbol{X}\boldsymbol{A}^{-1})(\boldsymbol{A}\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \end{aligned}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$. Note first that, since $\boldsymbol{A}^{-1}$ is upper triangular, the nested sequence of models is unchanged; the parameterizations within each model have changed, but only by transforming the variables inside the model. We thus have the same nested model selection problem.

Next note that $(\boldsymbol{X}^*)'\boldsymbol{X}^* = \mathrm{I}_n$, so the likelihood factors into independent likelihoods for the β_i^*; and this independence holds within each of the nested models, since the columns of $\boldsymbol{X}^*$ are orthonormal. Thus, if the prior is chosen to be $\pi(\boldsymbol{\beta}^*) = \prod_{i=1}^{q} \pi_i(\beta_i^*)$, then it follows from the earlier considerations in this section that the median probability model is optimal. For example, assuming $\boldsymbol{\beta}^* \sim \mathcal{N}(0, \boldsymbol{D})$ for some diagonal covariance matrix $\boldsymbol{D}$, we obtain generalizations of the g-prior for which we already know that MPM is optimal.

5 Discussion

The paper consists of two quite different parts. One part (mostly Section 4) focuses on generalizing previous theorems concerning the optimality of the median probability model. In addition to the generalizations therein a number of other generalizations are suggested in the paper, when groups of variables are orthogonal to others. Here are three such results, whose proofs are essentially obvious.

Result 1. If one group of variables is orthogonal to another, then finding the MPM and the optimal procedure can be done separately for each group of variables.

Result 2. If a variable is orthogonal to all others, it can be separated from the problem and handled on its own, and will belong in the optimal model if its inclusion probability is bigger than 1/2.

Result 3. If two groups of orthogonal variables each have a nested structure, then the median probability model is optimal and can be found separately in each group.

In spite of the considerable generalizations of optimality afforded by Section 4 and these related results, the extent to which the median probability model is guaranteed to be optimal is still rather limited. Hence the second goal of the paper was to study the extent to which the MPM failed to be optimal. This was done in two ways. First, by looking at "worst cases," where the number of highly correlated variables grows. The theoretical conclusions (besides Section 2.5) were obtained for the case of fixed and random variance. Note that the extreme case of perfect correlation is where we would expect the performance of the median probability model to be most challenged and several sections of the paper focused on this choice.

The second approach to studying the adequacy of the MPM was through an extensive numerical study to see how often the MPM (and MAP) fail to be optimal. Various degrees of correlated covariates were considered and, indeed, the more extreme the correlation, the worse the MPM performed. However, even with considerable correlation, the MPM was optimal more often than the MAP (when the two models were not the same). These results are encouraging and indicate that the MPM can also be optimal very often, even in non-nested non-orthogonal designs.

The MPM can fail, however, and fail badly, so we finish with a discussion of when this happens, focusing on the situation where there are many highly correlated covariate vectors $\boldsymbol{x}_i, i = 1, \ldots, k$. If k is large, we saw in Section 2.5 that the median probability model may well not include any of these covariates. (The section formally only considered the case of duplicate covariates, but the same conclusions apply qualitatively to the highly correlated case.) Consider four cases.

Case 1. *None of the $\boldsymbol{x}_i$ are useful for prediction:* Now the median probability model might well do better than the model averaged answer for the original problem, since the median probability model will ignore these covariates, while the model averaged answer will include them.

Case 2. *Including one of the $\boldsymbol{x}_i$ is crucial for good prediction:* Now the median probability model does poorly. Unfortunately, the error here, in not including one of the $\boldsymbol{x}_i$, will typically be larger than the gain in Case 1.

Case 3. *One of the $\boldsymbol{x}_i$ is helpful, but not crucial, for good prediction:* This is like the situation in Section 2.1. The harm in the median probability model ignoring the $\boldsymbol{x}_i$ is likely rather small.

Case 4. *Nested Models:* If the above arises in a nested model scenario, the median probability model is, of course, the optimal single model. It can still err, however, through the prior probabilities being inappropriate, assigning too much mass to all the duplicate models. (But this is just saying that the model averaged answer then can also err.)

Supplementary Material

Supplementary Material to "The Median Probability Model and Correlated Variables" (DOI: 10.1214/20-BA1249SUPP; .pdf). The supplementary material contains a proof of Theorem 1, details from the numerical study (including additional tables) and the results of a simulation study with $q = 5$ covariates.

References

Barbieri, M. M., Berger, J. O., George, E. I., and Ročková, V. (2020). "Supplementary Material to "The Median Probability Model and Correlated Variables"." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1249SUPP. 1101

Barbieri, M. M. and Berger, J. O. (2004). "Optimal Predictive Model Selection."

The Annals of Statistics, 32: 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 1085, 1086, 1087, 1100, 1108

Bayarri, S., Berger, J., Forte, A., and Garcia-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40: 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 1087, 1091

Berger, J. O. and Pericchi, R. L. (2001). "Objective Bayesian Methods for Model Selection: Introduction and Comparison (with discussion)." In Lahiri, P. (ed.), *Model Selection*, 135–207. Institute of Mathematical Statistics Lecture Notes- Monograph Series, volume 38. MR2000753. doi: https://doi.org/10.1214/lnms/1215540968. 1087

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). "Bayesian adaptive sampling for variable selection and model averaging." *Journal of Computational and Graphical Statistics*, 20(1): 80–101. MR2816539. doi: https://doi.org/10.1198/jcgs.2010.09049. 1086

Cui, W. and George, E. I. (2008). "Empirical Bayes vs. fully Bayes variable selection." *Journal of Statistical Planning and Inference*, 138(4): 888–900. MR2416869. doi: https://doi.org/10.1016/j.jspi.2007.02.011. 1095

Drachal, K. (2018). "Comparison between Bayesian and information-theoretic model averaging: Fossil fuels prices example." *Energy Economics*, 74: 208–251. doi: https://doi.org/10.1016/j.eneco.2018.04.043. 1086

Feldkircher, M. (2012). "Forecast combination and Bayesian model averaging: A prior sensitivity analysis." *Journal of Forecasting*, 31(4): 361–376. MR2924801. doi: https://doi.org/10.1002/for.1228. 1086

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. 1087

Garcia-Donato, G. and Martinez-Beneito, M. (2013). "On sampling strategies in Bayesian variable selection problems with large model spaces." *Journal of the American Statistical Association*, 108(501): 340–352. MR3174624. doi: https://doi.org/10.1080/01621459.2012.742443. 1086

George, E. I. (2010). "Dilution Priors: Compensating for Model Space Redundancy." *IMS Collections: Borrowing Strength: Theory Powering Applications – A. Festschri for Lawrence D. Brown*, 6: 158–165. MR2798517. 1087, 1099

George, E. I. and McCulloch, R. E. (1993). "Variable Selection Via Gibbs Sampling." *Journal of the American Statistical Association*, 88: 881–889. 1086, 1108

George, E. I. and McCulloch, R. E. (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7: 339–373. 1090

Ghosh, J. (2015). "Bayesian model selection using the median probability model." *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(3): 185–193. MR3349298. doi: https://doi.org/10.1002/wics.1352. 1086

Ishwaran, H. and Rao, J. S. (2003). "Detecting differentially expressed genes in microarrays using Bayesian model selection." *Journal of the American Statistical Association*, 98: 438–455. MR1995720. doi: https://doi.org/10.1198/016214503000224. 1086

Ley, E. and Steel, M. F. (2009). "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression." *Journal of Applied Econometrics*, 24: 651–674. MR2675199. doi: https://doi.org/10.1002/jae.1057. 1095

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. 1086

Piironen, J. and Vehtari, A. (2017). "Comparison of Bayesian predictive methods for model selection." *Statistics and Computing*, 27(3): 711–735. MR3613594. doi: https://doi.org/10.1007/s11222-016-9649-y. 1086

Rossell, D. and Telesca, D. (2017). "Nonlocal Priors for High-Dimensional Estimation." *Journal of the American Statistical Association*, 112: 254–265. MR3646569. doi: https://doi.org/10.1080/01621459.2015.1130634. 1108

Ročková, V. (2018). "Bayesian Estimation of Sparse Signals with a Continuous Spike-and-Slab Prior." *The Annals of Statistics*, 46: 401–437. MR3766957. doi: https://doi.org/10.1214/17-AOS1554. 1086

Ročková, V. and George, E. (2014). "EMVS: The EM approach to Bayesian variable selection." *Journal of the American Statistical Association*, 109: 828–846. MR3223753. doi: https://doi.org/10.1080/01621459.2013.869223. 1086

Scott, J. G. and Berger, J. O. (2010). "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem." *The Annals of Statistics*, 38: 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 1095

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions." *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics*, 6: 233–243. MR0881437. 1086, 1087