

## THE MEDIAN PROBABILITY MODEL AND CORRELATED VARIABLES

BY MARIA M. BARBIERI<sup>†</sup>, JAMES O. BERGER<sup>‡</sup>,  
EDWARD I. GEORGE AND<sup>§</sup>, VERONIKA ROČKOVA<sup>\*,¶</sup>,  
17 August 2018

*Università Roma Tre* <sup>†</sup>

*Duke University* <sup>‡</sup>

*University of Pennsylvania* <sup>§</sup>

*University of Chicago* <sup>¶</sup>

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.

**1. Introduction.** This paper investigates to which extent 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

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

---

\* Veronika.Rockova@ChicagoBooth.edu;

This work was supported by the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business.

*Keywords and phrases:* Bayesian variable selection, Median probability model, Multicollinearity, Spike and slab

where  $\mathbf{Y}$  is the  $n \times 1$  vector of responses,  $\mathbf{X}$  is the  $n \times q$  design matrix of covariates,  $\boldsymbol{\beta}$  is a  $q \times 1$  vector of unknown coefficients, and  $\sigma^2$  is a known scalar. The equation (1.1) corresponds to the full model and we are interested in selecting a submodel indexed by  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_q)'$ , where  $\gamma_i \in \{1, 0\}$  for whether the  $i^{\text{th}}$  covariate is in or out of the model. We tacitly assume that the response has been centered and thereby omit the intercept term.

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

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

where  $\bar{\boldsymbol{\beta}} = \mathbb{E}[\boldsymbol{\beta} | \mathbf{Y}]$  is the overall posterior mean of  $\boldsymbol{\beta}$  under the hierarchical prior  $\pi(\boldsymbol{\gamma})$  and  $\pi(\boldsymbol{\beta} | \boldsymbol{\gamma})$ ,  $\mathbf{H}_{\boldsymbol{\gamma}}$  is a  $q \times |\boldsymbol{\gamma}|$  stretching matrix (defined in Section 2.2 of Barbieri and Berger (2004)) whose  $(i, j)$  entry is 1 if  $\gamma_i = 1$  and  $j = \sum_{r=1}^i \gamma_r$  and 0 otherwise, and  $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$  is the conditional posterior mean under  $\boldsymbol{\gamma}$  and where  $\mathbf{Q} = \mathbb{E}[\mathbf{x}^* \mathbf{x}^{*'}] = \mathbf{X}' \mathbf{X}$ .

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 the *median probability model*  $\boldsymbol{\gamma}^{MP}$ . This is defined to be the model consisting of variables whose marginal inclusion probability  $\pi(\gamma_i = 1 | \mathbf{Y})$  is at least 0.5. This model can be regarded as the best single-model approximation to model averaging.

Compared to the often targeted highest posterior model (HPM), a major attraction of the median probability model (MPM) is the relative ease with which it can be found via MCMC. Whereas finding the HPM requires identification of the largest of the  $2^p$  distinct model probabilities, finding the MPM is much less costly, requiring identification only of those of the  $p$  marginal covariate probabilities which are greater 0.5. And when the HPM 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; Scott and Carvalho (2008); Clyde, Ghosh and Littman (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

$\widehat{\beta}_\gamma$  is proportional to the MLE estimator under  $\gamma$ . 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 HPM in correlated scenarios.

First, using predictive matching arguments (Berger and Pericchi, 2001; Bayarri et al., 2012; Fouskakis, Ntzoufras and Perrakis, 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. The effect of many highly correlated variables and $g$ -priors on the median probability model.

2.1. *The marginal likelihood in the presence of 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 design matrix satisfies

$$(2.1) \quad \mathbf{X}_{n \times (p+k)} = [\mathbf{B}_{n \times p}, \mathbf{x} + \epsilon \boldsymbol{\delta}_1, \dots, \mathbf{x} + \epsilon \boldsymbol{\delta}_k]$$

for some  $\epsilon > 0$ , where  $\mathbf{B}$  consists of  $p$  possibly correlated regressors and where  $\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_k$  are  $(n \times 1)$  perturbation vectors. We assume that  $\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_k$  are orthonormal and orthogonal to  $\mathbf{x}$  and  $\mathbf{B}$ ,<sup>1</sup> while  $\mathbf{x}$  and  $\mathbf{B}$  are not necessarily orthogonal. We will be letting  $\epsilon$  be very small to model the situation of having  $k$  highly correlated variables. For the full model (1.1), the  $g$ -prior is

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

for some  $g > 0$  (typically  $n$ ) and the corresponding marginal likelihood is

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

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

LEMMA 2.1. *Denote with  $\mathbf{P} = \lim_{\epsilon \rightarrow 0} \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$ . Then*

$$\mathbf{P} = \mathbf{P}_B + \frac{1}{(\|\mathbf{x}\|^2 - \mathbf{x}' \mathbf{P}_B \mathbf{x})} [\mathbf{P}_B \mathbf{x} - \mathbf{x}] [\mathbf{x}' \mathbf{P}_B - \mathbf{x}'],$$

where  $\mathbf{P}_B = \mathbf{B} (\mathbf{B}' \mathbf{B})^{-1} \mathbf{B}'$ .

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

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

---

<sup>1</sup>This assumption is not necessary, but greatly simplifies the illustration.

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

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

The result follows by multiplying this matrix with  $\mathbf{X}$  and  $\mathbf{X}'$ , and taking the limit as  $\epsilon \rightarrow 0$ .  $\square$

One important conclusion from Lemma 2.1 is that no matter how many columns of highly correlated variables are present in the model, the marginal likelihood will essentially be

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

as  $\epsilon \rightarrow 0$ . Thereby all models including all predictors in  $\mathbf{B}$  and at least one replicate of  $\mathbf{x}$  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  $\mathbf{B}$  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  $\mathbf{B}_{\boldsymbol{\gamma}_1}$  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 2.1.** (*A model collective*) Let  $\boldsymbol{\gamma}_1 \in \{0, 1\}^p$  be a vector of inclusion indicators associated with the  $p$  variables in  $\mathbf{B}$ . Denote by  $M_{\boldsymbol{\gamma}_1, \mathbf{x}}$  the model collective comprising all models consisting of the  $\boldsymbol{\gamma}_1$  variables together with one or more of the (near) duplicates of  $\mathbf{x}$ .

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

$$(2.2) \quad m(\mathbf{y} \mid \boldsymbol{\gamma}_1, \mathbf{x}) = \phi(\mathbf{y} \mid \mathbf{0}, \sigma^2\mathbf{I} + g\sigma^2\mathbf{P}_{\boldsymbol{\gamma}_1}),$$

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

**LEMMA 2.2.** Let  $m(\mathbf{y} \mid \boldsymbol{\gamma}_1)$  denote the marginal likelihood under the model  $\boldsymbol{\gamma}_1$ . Then we have  $m(\mathbf{y} \mid \boldsymbol{\gamma}_1, \mathbf{x}) = m(\mathbf{y} \mid \boldsymbol{\gamma}_1) \times m(\mathbf{y} \mid \mathbf{x})$ , where

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

and  $\mathbf{P}_{\mathbf{B}_{\gamma_1}} = \mathbf{B}_{\gamma_1}(\mathbf{B}'_{\gamma_1}\mathbf{B}_{\gamma_1})^{-1}\mathbf{B}'_{\gamma_1}$ . Note that, if  $\mathbf{x}$  is orthogonal to  $\mathbf{B}$ , then

$$(2.3) \quad m(\mathbf{y} | \mathbf{x}) = \frac{1}{\sqrt{1+g}} \exp \left\{ \frac{g}{2\sigma^2(1+g)} \times \left[ \mathbf{y}' \frac{\mathbf{x}}{\|\mathbf{x}\|} \right]^2 \right\}.$$

PROOF. Letting  $\mathbf{P}$  denote  $\mathbf{P}_{\mathbf{B}_{\gamma_1}}$  and  $\mathbf{z} = (\mathbf{I} - \mathbf{P})\mathbf{x} / \sqrt{\mathbf{x}'(\mathbf{I} - \mathbf{P})\mathbf{x}}$ , this follows from the identities

$$\begin{aligned} (\mathbf{I} + g\mathbf{P} + g\mathbf{z}\mathbf{z}')^{-1} &= (\mathbf{I} + g\mathbf{P})^{-1} - \frac{(\mathbf{I} + g\mathbf{P})^{-1}\mathbf{z}\mathbf{z}'(\mathbf{I} + g\mathbf{P})^{-1}}{[g^{-1} + \mathbf{z}'(\mathbf{I} + g\mathbf{P})^{-1}\mathbf{z}]}, \\ |\mathbf{I} + g\mathbf{P} + g\mathbf{z}\mathbf{z}'| &= |\mathbf{I} + g\mathbf{P}| (1 + g\mathbf{z}'(\mathbf{I} + g\mathbf{P})^{-1}\mathbf{z}), \\ (\mathbf{I} + g\mathbf{P})(\mathbf{I} - \mathbf{P}) &= (\mathbf{I} - \mathbf{P}), \text{ so that } (\mathbf{I} + g\mathbf{P})^{-1}(\mathbf{I} - \mathbf{P}) = (\mathbf{I} - \mathbf{P}). \square \end{aligned}$$

REMARK 2.1. If  $\mathbf{x}$  is orthogonal to  $\mathbf{B}$ , the corresponding Bayes estimates are just the usual  $g$ -prior posterior means

$$(2.4) \quad \frac{g}{1+g} \left( \hat{\boldsymbol{\beta}}_{\gamma_1}^{MLE}, \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|^2} \right).$$

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.

2.2. *Dimensional predictive matching.* As a first application of Lemma 2.2, we note that the (limiting) marginal likelihood under the  $g$ -prior is the same, no matter how many replicates of  $\mathbf{x}$  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 2.2.  $\square$

In contrast, it is of interest to look at what happens with an alternative prior for  $\boldsymbol{\beta}$  such as a  $N(\mathbf{0}, \mathbf{I})$  prior. If a model has  $j$  near-replicates of  $\mathbf{x}$ , the effective parameter for  $\mathbf{x}$  in that model is the sum of the  $j$   $\beta$ '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  $\mathbf{B}'\mathbf{B} = \mathbf{I}$  in (2.1). In particular, we will be working with an orthogonal design that has been augmented with multiple copies of one predictor

$$(2.5) \quad \mathbf{X}_{n \times (p+k)} = [\mathbf{x}_1, \dots, \mathbf{x}_p, \underbrace{\mathbf{x}, \dots, \mathbf{x}}_k],$$

where  $\mathbf{x}_1, \dots, \mathbf{x}_p, \mathbf{x}$  are orthonormal. 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  $z = \mathbf{x}'\mathbf{y}$ ,  $z_i = \mathbf{x}'_i\mathbf{y}$  for  $i = 1, \dots, p$  and  $\mathbf{z} = (z'_1, z'_2)'$ , where  $z_1 = (z_1, \dots, z_p)'$  and  $z_2 = z\mathbf{1}_k$ . 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  $\boldsymbol{\gamma}$ ,

$$(2.6) \quad \boldsymbol{\beta}_\boldsymbol{\gamma} \sim \mathcal{N}_{|\boldsymbol{\gamma}|} \left( \mathbf{0}, g \sigma^2 (\mathbf{X}'_{|\boldsymbol{\gamma}|} \mathbf{X}_{|\boldsymbol{\gamma}|})^+ \right),$$

where  $(\mathbf{X}'_{|\boldsymbol{\gamma}|} \mathbf{X}_{|\boldsymbol{\gamma}|})^+$  is the Moore-Penrose pseudo-inverse. The following lemma characterizes the optimal predictive model under (2.6) and (2.5).

LEMMA 2.3. *Consider the model (1.1) where  $\mathbf{X}$  satisfies (2.5) and where  $\mathbf{x}_1, \dots, \mathbf{x}_p, \mathbf{x}$  are orthonormal. Under the prior (2.6), any model  $\boldsymbol{\gamma}^o = (\boldsymbol{\gamma}_1^o, \boldsymbol{\gamma}_2^o)'$  that satisfies*

$$(2.7) \quad \gamma_{1i}^o = 1 \quad \text{iff} \quad \pi(\gamma_{1i} = 1 \mid \mathbf{Y}) > 0.5, \quad i = 1, \dots, p$$

$$(2.8) \quad |\boldsymbol{\gamma}_2^o| \geq 1 \quad \text{iff} \quad \pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} \mid \mathbf{Y}) > 0.5$$

*is predictive optimal.*

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

$$\mathbf{H}_\boldsymbol{\gamma} \widehat{\boldsymbol{\beta}}_\boldsymbol{\gamma} = \frac{g}{(1+g)} \begin{pmatrix} \text{diag}\{\boldsymbol{\gamma}_1\} & \mathbf{0} \\ \mathbf{0} & \frac{1}{|\boldsymbol{\gamma}_2|} \text{diag}\{\boldsymbol{\gamma}_2\} \end{pmatrix} \mathbf{z}$$

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

$$\mathbf{H}_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} - \bar{\boldsymbol{\beta}} = \frac{g}{(1+g)} \begin{pmatrix} \text{diag}\{\boldsymbol{\gamma}_1 - \mathbb{E}[\boldsymbol{\gamma}_1 | \mathbf{Y}]\} & \mathbf{0} \\ \mathbf{0} & \text{diag}\left\{\frac{\boldsymbol{\gamma}_2}{|\boldsymbol{\gamma}_2|} - \mathbb{E}\left[\frac{\boldsymbol{\gamma}_2}{|\boldsymbol{\gamma}_2|} \mid \mathbf{Y}, \boldsymbol{\gamma}_2 \neq \mathbf{0}\right] \pi(\boldsymbol{\gamma}_2 \neq \mathbf{0})\right\} \end{pmatrix} \mathbf{z}$$

The optimal predictive model minimizes  $R(\boldsymbol{\gamma})$  defined in (1.2). Due to the fact that  $\mathbf{Q}$  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

$$(2.9) \quad R_1(\boldsymbol{\gamma}_1) = \frac{g^2}{(1+g)^2} \sum_{i=1}^p z_i^2 (\gamma_{1i} - \mathbb{E}[\gamma_{1i} | \mathbf{Y}])^2$$

$$(2.10) \quad R_2(\boldsymbol{\gamma}_2) = \frac{g^2 z^2}{(1+g)^2} \left\{ [1 - \pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} | \mathbf{Y})]^2 \mathbb{I}(\boldsymbol{\gamma}_2 \neq \mathbf{0}) + \pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} | \mathbf{Y})^2 \mathbb{I}(\boldsymbol{\gamma}_2 = \mathbf{0}) \right\}.$$

The statement then follows from (2.9) and (2.10). With duplicate columns, the optimal predictive model  $\boldsymbol{\gamma}^o \equiv \arg \min_{\boldsymbol{\gamma}} R(\boldsymbol{\gamma})$  is not unique. Any model  $\boldsymbol{\gamma}^o = (\boldsymbol{\gamma}_1^{o'}, \boldsymbol{\gamma}_2^{o'})'$  defined through (2.7) and (2.8) will minimize the criterion  $R(\boldsymbol{\gamma})$ .  $\square$

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

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

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

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



The median probability model  $\gamma^{MP}$  thus behaves as the optimal model  $\gamma^o$  for the first  $p$  variables. For the  $k$  duplicate copies, however,  $\gamma_2^{MP}$  consists of either *all ones* or *all zeros*. The MP rule correctly recognizes that the decision to include  $\mathbf{x}$  is ultimately dichotomous: either all  $\mathbf{x}$ 's in or all  $\mathbf{x}$ 's out. Moreover, when the median model decides ‘‘all in’’, it *will be predictive optimal*. Indeed,  $\pi(\gamma_{2i}^{MP} = 1 \mid \mathbf{Y}) > 1/2$  for  $i \in \{1, \dots, k\}$  implies  $\pi(\gamma_2 = \mathbf{0} \mid \mathbf{Y}) < 1/2$ . The MP model will deviate from the optimal model only when  $\pi(\gamma_{2i}^{MP} = 1 \mid \mathbf{Y}) < 1/2$  and  $\pi(\gamma_2 \neq \mathbf{0} \mid \mathbf{Y}) > 1/2$  in which case

$$\frac{R(\gamma^{MP}) - R(\gamma^o)}{R(\gamma^o)} = \frac{R_2(\gamma_2^{MP}) - R_2(\gamma_2^o)}{R(\gamma^o)} = \frac{g^2 z^2 [1 - 2\pi(\gamma_2 \neq \mathbf{0} \mid \mathbf{Y})]}{(1 + g)^2 [R_1(\gamma_1^o) + R_2(\gamma_2^o)]}.$$

The term  $R_1(\gamma_1^o)$  in (2.9) 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 2.3 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. We will elaborate on this property in Section 4.3, showing that optimal predictive model can be characterized using both posterior means and covariances of  $\gamma$  (in equicorrelated designs). 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  $\mathbf{x}$ ) 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 2.2. (*Independent Product Priors*) Let us replace (2.6) with an independent prior covariance structure

$$(2.11) \quad \beta_\gamma \sim \mathcal{N}_{|\gamma|}(\mathbf{0}, \sigma^2 \mathbf{I}_{|\gamma|}),$$

The posterior mean  $\bar{\beta}$  then satisfies  $H_\gamma \hat{\beta}_\gamma - \bar{\beta} =$

$$\begin{pmatrix} \text{diag}\{\gamma_1 - \mathbb{E}[\gamma_1 \mid \mathbf{Y}]\} & \mathbf{0} \\ \mathbf{0} & \text{diag}\{\gamma_2 - \mathbb{E}[\gamma_2 \mid \mathbf{Y}]\} - \frac{1}{1+|\gamma_2|} \gamma_2 \gamma_2' + \mathbb{E}\left[\frac{1}{1+|\gamma_2|} \gamma_2 \gamma_2' \mid \mathbf{Y}\right] \end{pmatrix} \mathbf{z}$$

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

$$(2.12) \quad R_1^*(\boldsymbol{\gamma}_1) = \sum_{i=1}^p z_i^2 (\gamma_{1i} - \mathbb{E}[\gamma_{1i} | \mathbf{Y}])^2$$

$$(2.13) \quad R_2^*(\boldsymbol{\gamma}_2) = z^2 \left[ \frac{|\boldsymbol{\gamma}_2|}{1 + |\boldsymbol{\gamma}_2|} - \mathbb{E} \left( \frac{|\boldsymbol{\gamma}_2|}{1 + |\boldsymbol{\gamma}_2|} \middle| \mathbf{Y} \right) \right]^2.$$

The optimal predictive model for the last  $k$  variables now has a bit less intuitive explanation. It consists of any collection of variables of size  $|\boldsymbol{\gamma}_2^o|$  for which  $|\boldsymbol{\gamma}_2^o|/[1 + |\boldsymbol{\gamma}_2^o|]$  is as close as possible to the posterior mean of  $|\boldsymbol{\gamma}_2|/[1 + |\boldsymbol{\gamma}_2|]$ . It is worthwhile to note that this does not need to be the null or the full model. For instance, one can show that  $\boldsymbol{\gamma}_2^o = \mathbf{0}$  when  $\mathbb{E}[|\boldsymbol{\gamma}_2| | \mathbf{Y}] < 1/3$  (or more generally when  $\pi(\boldsymbol{\gamma}_2 = \mathbf{0} | \mathbf{Y}) > 2/3$ ). The full model  $\boldsymbol{\gamma}_2^o = \mathbf{1}$  will be optimal when  $\mathbb{E}[|\boldsymbol{\gamma}_2| | \mathbf{Y}] > k - \frac{k+1}{2k+1}$ . 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

$$(2.14) \quad \pi(\boldsymbol{\gamma}) = \pi(|\boldsymbol{\gamma}|) / \binom{p+k}{|\boldsymbol{\gamma}|},$$

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 2.3 is that it is the aggregate posterior probability  $\pi(\boldsymbol{\gamma}_2 \neq \mathbf{0} | \mathbf{Y})$  rather than individual inclusion probabilities  $\pi(\gamma_{2i} = 1 | \mathbf{Y})$  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 2.1. The number of models of size  $j > |\boldsymbol{\gamma}_1|$  in the model collective  $M_{\boldsymbol{\gamma}_1, \mathbf{x}}$  is  $\binom{k}{j - |\boldsymbol{\gamma}_1|}$ , so that the prior probability of the model collective  $M_{\boldsymbol{\gamma}_1, \mathbf{x}}$  is

$$(2.15) \quad \pi(M_{\boldsymbol{\gamma}_1, \mathbf{x}}) = \sum_{j=|\boldsymbol{\gamma}_1|+1}^{|\boldsymbol{\gamma}_1|+k} \frac{\pi(j)}{\binom{p+k}{j}} \binom{k}{j - |\boldsymbol{\gamma}_1|}.$$

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  $\theta = \pi(\gamma_j = 1 \mid \theta)$  for  $j = 1, \dots, p + k$ . Then the probability of the model aggregate, given  $\theta$ , is

$$(2.16) \quad \pi(M_{\gamma_1, \mathbf{x}}; \theta) = \theta^{|\gamma_1|} (1 - \theta)^{p - |\gamma_1|} \sum_{j=1}^k \theta^j (1 - \theta)^{k-j} \binom{k}{j}$$

$$(2.17) \quad = \theta^{|\gamma_1|} (1 - \theta)^{p - |\gamma_1|} \left[ 1 - (1 - \theta)^k \right]$$

and the prior probability of the “null model”  $M_{\gamma_1, \mathbf{0}}$  (not including any of the correlated variables) is

$$(2.18) \quad \pi(M_{\gamma_1, \mathbf{0}}; \theta) = \theta^{|\gamma_1|} (1 - \theta)^{p - |\gamma_1|} (1 - \theta)^k.$$

The ratio satisfies

$$\frac{\pi(M_{\gamma_1, \mathbf{x}}; \theta)}{\pi(M_{\gamma_1, \mathbf{0}}; \theta)} = \left[ \left( \frac{1}{1 - \theta} \right)^k - 1 \right].$$

This analysis reveals a rather spurious property of the separable prior: regardless of the choice  $\theta \in (0, 1)$ , the model aggregate  $M_{\gamma_1, \mathbf{x}}$  will always have a higher prior probability than the model  $M_{\gamma_1, \mathbf{0}}$  without any  $\mathbf{x}$  in it. Such a preferential treatment for  $\mathbf{x}$  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  $M_{\gamma_1, \mathbf{x}}$  is  $2^k - 1$ , and so

$$\pi(M_{\gamma_1, \mathbf{x}}; 1/2) = (2^k - 1)2^{-(p+k)} = (2^k - 1) \pi(M_{\gamma_1, \mathbf{0}}; 1/2).$$

The collective can thus have much more prior probability than  $M_{\gamma_1, \mathbf{0}}$ . Furthermore, the marginal prior probability of inclusion of  $\mathbf{x}$  is  $\sum_{\gamma_1} \pi(M_{\gamma_1, \mathbf{x}}; 1/2) = 1 - 2^{-k}$ . Hence, if  $k$  is even moderately large, the prior mass is concentrated on the models which include  $\mathbf{x}$  as a covariate, and the posterior mass will almost certainly also be concentrated on those models. The model-averaged  $\bar{\beta}$  will reflect this, essentially only including models that have  $\mathbf{x}$  as a covariate.

*Beta-binomial Prior:* It is generally acknowledged (Cui and George, 2008; 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

$$\begin{aligned} \pi(M_{\gamma_1, \mathbf{x}}) &= \int_0^1 \pi(M_{\gamma_1, \mathbf{x}}; \theta) d\pi(\theta) = \int_0^1 \theta^{|\gamma_1|+a-1} (1-\theta)^{p-|\gamma_1|+b-1} \left[1 - (1-\theta)^k\right] d\theta \\ &= \mathcal{B}(|\gamma_1| + a, p - |\gamma_1| + b) - \mathcal{B}(|\gamma_1| + a, p + k - |\gamma_1| + b) \\ &= \mathcal{B}(|\gamma_1| + a, p + k - |\gamma_1| + b) \left[ \prod_{j=1}^k \frac{a + b + p + j - 1}{p + b - |\gamma_1| + j - 1} - 1 \right] \end{aligned}$$

and

$$(2.19) \quad \pi(M_{\gamma_1, \mathbf{0}}) = \mathcal{B}(|\gamma_1| + a, p + k - |\gamma_1| + b).$$

Then

$$(2.20) \quad \frac{\pi(M_{\gamma_1, \mathbf{x}})}{\pi(M_{\gamma_1, \mathbf{0}})} = \left[ \prod_{j=1}^k \left( 1 + \frac{|\gamma_1| + a}{p + b - |\gamma_1| + j - 1} \right) - 1 \right].$$

This ratio is guaranteed to be smaller than under the separable case with a fixed  $\theta$  when  $|\gamma_1| < (p+b)\theta - a(1-\theta)$ . 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 2.5, we provide an approximation to (2.20) 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  $M_{\gamma_1, \mathbf{x}}$  receives much more prior mass relative to  $M_{\gamma_1, \mathbf{0}}$ , essentially forcing the inclusion of  $\mathbf{x}$ . Going further, we show how this is reflected in the posterior inclusion probabilities.

LEMMA 2.4. *Consider the model (1.1), where  $\mathbf{X}$  satisfies (2.5) and where  $\mathbf{x}_1, \dots, \mathbf{x}_p, \mathbf{x}$  are orthonormal. Denote by  $z = \mathbf{y}'\mathbf{x}$  and consider the prior (2.6) with  $g = n$  and equal prior model probabilities  $\pi(\gamma) = 1/2^{p+k}$ . Then we have*

$$(2.21) \quad \pi(\gamma_2 \neq \mathbf{0} | \mathbf{Y}) > 1/2 \quad \text{iff} \quad z^2 > \log \left( \frac{\sqrt{1+n}}{2^k - 1} \right) 2\sigma^2 \left( 1 + \frac{1}{n} \right).$$

PROOF. The posterior probability of joint inclusion  $\pi(\gamma_2 \neq \mathbf{0} | \mathbf{Y})$  (noting that  $\pi(M_{\gamma_1, \mathbf{x}})$  and  $\pi(M_{\gamma_1, \mathbf{0}})$  depend only on  $|\gamma_1|$ ) equals

$$\begin{aligned}
 \pi(\gamma_2 \neq \mathbf{0} | \mathbf{Y}) &= \frac{\sum_{\gamma_1} \pi(M_{\gamma_1, \mathbf{x}}) m(\mathbf{y} | \gamma_1, \mathbf{x})}{\sum_{\gamma_1} \pi(M_{\gamma_1, \mathbf{x}}) m(\mathbf{y} | \gamma_1, \mathbf{x}) + \sum_{\gamma_1} \pi(M_{\gamma_1, \mathbf{0}}) m(\mathbf{y} | \gamma_1, \mathbf{0})} \\
 &= \frac{\sum_{\gamma_1} \pi(M_{\gamma_1, \mathbf{x}}) m(\mathbf{y} | \gamma_1) m(\mathbf{y} | \mathbf{x})}{\sum_{\gamma_1} \pi(M_{\gamma_1, \mathbf{x}}) m(\mathbf{y} | \gamma_1) m(\mathbf{y} | \mathbf{x}) + \sum_{\gamma_1} \pi(M_{\gamma_1, \mathbf{0}}) m(\mathbf{y} | \gamma_1)} \\
 (2.22) \quad &= \frac{m(\mathbf{y} | \mathbf{x}) \sum_{i=0}^p \pi_{i, \mathbf{x}}^*}{\sum_{i=0}^p [m(\mathbf{y} | \mathbf{x}) \pi_{i, \mathbf{x}}^* + \pi_{i, \mathbf{0}}^*]},
 \end{aligned}$$

with  $\pi_{i, \mathbf{x}}^* = \sum_{\gamma_1: |\gamma_1|=i} m(\mathbf{y} | \gamma_1) \pi(M_{\gamma_1, \mathbf{x}})$  and  $\pi_{i, \mathbf{0}}^* = \sum_{\gamma_1: |\gamma_1|=i} m(\mathbf{y} | \gamma_1) \pi(M_{\gamma_1, \mathbf{0}})$ , and where

$$m(\mathbf{y} | \mathbf{x}) = \frac{1}{\sqrt{1+g}} \exp \left\{ \frac{g}{2\sigma^2(1+g)} \times [\mathbf{y}'\mathbf{x}]^2 \right\}$$

and  $m(\mathbf{y} | \gamma_1, \mathbf{x}) = m(\mathbf{y} | \gamma_1) m(\mathbf{y} | \mathbf{x})$  was defined in (2.2). When each model has equal prior probability, this simplifies to

$$\pi(\gamma_2 \neq \mathbf{0} | \mathbf{Y}) = \frac{(2^k - 1)m(\mathbf{y} | \mathbf{x})}{1 + (2^k - 1)m(\mathbf{y} | \mathbf{x})}.$$

With the usual choice  $g = n$ , it follows that  $\pi(\gamma_2 \neq \mathbf{0} | \mathbf{Y}) > 0.5$  iff  $z^2 > \log \left( \frac{\sqrt{1+n}}{2^k - 1} \right) 2\sigma^2(1 + \frac{1}{n})$ .  $\square$

From Lemma 2.4 it follows that the optimal predictive model (characterized in Lemma 2.3) will include  $\mathbf{x}$  if the number of duplicates  $k$  is large enough, even when  $\mathbf{x}$  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 problematical. If one is only developing a model for prediction in such a situation, such forced inclusion of  $\mathbf{x}$  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  $\bar{\beta}$  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  $\mathbf{x}$  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  $\pi_{i,\mathbf{x}}^*$  for some  $0 \leq i \leq p$  is much bigger than the others, so that (2.22) becomes

$$\pi(\gamma_2 \neq \mathbf{0} \mid \mathbf{Y}) \approx \frac{\pi_{i,\mathbf{x}}^* m(\mathbf{y} \mid \mathbf{x})}{\pi_{i,\mathbf{x}}^* m(\mathbf{y} \mid \mathbf{x}) + \pi_{i,\mathbf{0}}^*}.$$

Using (2.20), it is immediate that this is bigger than 0.5 if

$$(2.23) \quad 1 < \frac{\pi_{i,\mathbf{x}}^*}{\pi_{i,\mathbf{0}}^*} m(\mathbf{y} \mid \mathbf{x}) = \left[ \prod_{j=1}^k \left( 1 + \frac{i+a}{p+b-i+j-1} \right) - 1 \right] m(\mathbf{y} \mid \mathbf{x}).$$

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

LEMMA 2.5. *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.  $\square$

Utilization of the first order term in (2.23), and again choosing  $g = n$  and assuming  $\|\mathbf{x}\| = 1$ , 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 (2.21), when  $k$  grows, since  $(1+k/p)^{[i+a]}$  is then much smaller than  $2^k$ ; thus having many duplicate  $\mathbf{x}$ 's does not ensure that  $\mathbf{x}$  will be included in the model, as it was in the equal model probability case.

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 this cluster of redundant models from accumulating too much posterior probability, dilution priors may be considered, George (2010). Such priors 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.

For example, to smear out the prior aggregation on  $M_{\gamma_1, \mathbf{x}}$ , one might like to consider different inclusion probabilities. Let  $[\mathbf{x}_1, \dots, \mathbf{x}_p]$  have a prior inclusion probability  $\theta_1$  and each of the  $\mathbf{x}$  clones have a prior inclusion probability  $\theta_2$ . With

$$(2.24) \quad \theta_2 = 1 - (1 - \theta_1)^{1/k}$$

we have

$$(2.25) \quad \pi(M_{\gamma_1, \mathbf{x}}) = \theta_1^{|\gamma_1|} (1 - \theta_1)^{p - |\gamma_1|} \left[ 1 - (1 - \theta_2)^k \right] = \theta_1^{|\gamma_1|+1} (1 - \theta_1)^{p - |\gamma_1|}.$$

and

$$\pi(M_{\gamma_1, \mathbf{0}}) = \theta_1^{|\gamma_1|} (1 - \theta_1)^{p - |\gamma_1| + 1}.$$

Assuming (2.24), variables with correlated copies have smaller inclusion probabilities (the more copies, the smaller the probability). This may correct the imbalance between  $\pi(M_{\gamma_1, \mathbf{x}})$  and  $\pi(M_{\gamma_1, \mathbf{0}})$  by treating the multiple copies of  $\mathbf{x}$  essentially as one variable. This prior allocation would put  $\mathbf{x}$  on an equal footing with  $\mathbf{x}_1, \dots, \mathbf{x}_p$  in the optimal predictive model rule (based on  $\pi(\gamma_2 \neq \mathbf{0} | \mathbf{Y})$ ), but would disadvantage  $\mathbf{x}$  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  $\mathbf{x}$ '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. Define  $\boldsymbol{\alpha}_\gamma$  as the projection of  $\mathbf{Y}$  onto the space spanned by the columns of  $\mathbf{X}_\gamma$ ,  $\boldsymbol{\alpha}_\gamma = \mathbf{X} \mathbf{H}_\gamma \hat{\boldsymbol{\beta}}_\gamma = \mathbf{X}_\gamma (\mathbf{X}'_\gamma \mathbf{X}_\gamma)^{-1} \mathbf{X}'_\gamma \mathbf{Y}$ , and  $\bar{\boldsymbol{\alpha}} = \mathbf{X} \bar{\boldsymbol{\beta}} = \sum_\gamma p_\gamma \mathbf{X} \mathbf{H}_\gamma \hat{\boldsymbol{\beta}}_\gamma$ , where  $p_\gamma = \pi(\gamma | \mathbf{Y})$ . The expected posterior loss (1.2) to be minimized may be written as

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

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

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

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  $r_{12} = \text{Corr}(x_1, x_2)$ ,  $r_{1y} = \text{Corr}(x_1, Y)$  and  $r_{2y} = \text{Corr}(x_2, Y)$ .

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:

$$\begin{array}{llll} \text{Case 1} & r_{12} = -0.5 & r_{1y} = 0.3 & r_{2y} = 0.4 \\ \text{Case 2} & r_{12} = 0.5 & r_{1y} = 0.3 & r_{2y} = 0.4 \\ \text{Case 3} & r_{12} = 0.5 & r_{1y} = 0.1 & r_{2y} = 0.3 \end{array}$$



$r_{12} = 0$	$r_{12} \frac{r_{1y}}{r_{2y}} < 0$	$r_{12} \frac{r_{1y}}{r_{2y}} > 0$	
orthogonal	case 1	$ r_{12}  < \min\{\frac{r_{1y}}{r_{2y}}, \frac{r_{2y}}{r_{1y}}\}$	$ \frac{r_{1y}}{r_{2y}}  <  r_{12} $
		case 2	case 3

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

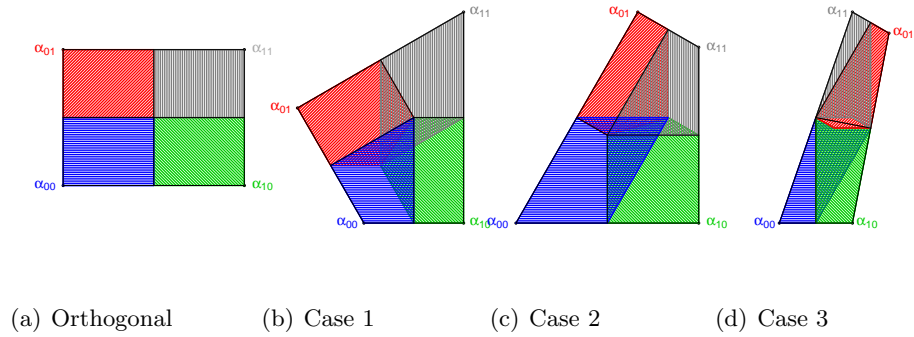


FIG 1. The median posterior probability model

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  $\alpha_{11}$  on the line spanned by  $\alpha_{10}$  is  $\alpha_{10}$  itself and similarly for  $\alpha_{01}$ ].

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

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 pre-

dictive model and the median probability model. These are summarized by the “mini-theorems” below.

**THEOREM 3.1.** (*“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 inclusion probability larger than  $0.5$ , it is optimal.*
6. *In any case, if  $M_{00}$  or  $M_{11}$  has posterior inclusion probability larger than  $0.5$ , the other cannot be optimal.*
7. *In Cases 3, if  $M_{00}$  or  $M_{11}$  has posterior inclusion probability smaller than  $0.5$  it cannot be optimal.*

**PROOF.** Appendix 1.

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 HPM 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. 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 then.)*
- $r_{1y}$  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 Appendix 2.
- Sample sizes  $n = 10, 50$  and  $100$  are considered.

	M=H=O	M=H≠O	M=O≠H	H=O≠M	H>M both ≠ O	M>H both ≠ O	GM $\frac{R(\gamma^M)}{R(\gamma^O)}$	GM $\frac{R(\gamma^H)}{R(\gamma^O)}$
Cases combined: Full model scenario								
n=10	404	93	27	3	4*	3*	1.08	1.10
n=50	505	8	13	0	8*	0	1.02	1.06
n=100	512	8	14	0	0	0	1.02	1.06
Overall	1421	109	54	3	12	3	1.03	1.07
	88.7%	6.8%	3.4%	0.2%	0.7%	0.2%		
Cases combined: $\beta_1 = 0$ and $\beta_2 \neq 0$ scenario								
n=10	424	169	18	2	3*	4*	1.104	1.114
n=50	661	57	3	1	9	0	1.054	1.049
n=100	682	65	1	0	0	1*	1.045	1.046
Overall	1767	291	22	3	12	5	1.065	1.067
	84.1%	13.9%	1.1%	0.1%	0.6%	0.2%		
Cases combined: Null model scenario								
n=10	470	178	50	2	8	7*	1.106	1.120
n=50	682	98	13	3	0	2*	1.039	1.045
n=100	735	60	5	1	0	1	1.023	1.024
Overall	1887	336	68	6	8	10	1.054	1.060
	81.6%	14.5%	2.9%	0.3%	0.3%	0.4%		

TABLE 2

A summary of the numerical study in the two variable case. Legend:  $H = HPM$ ,  $M = MPM$ ,  $O =$  optimal predictive model;  $M>H$  means that  $MPM$  has smaller (1.2) than  $HPM$ ;  $H>M$  means that  $HPD$  has smaller (1.2) than  $MPM$ ; and  $GM$  is the geometric mean of relative risks (to the optimal model) when  $MPM$  or  $HPM$  is not optimal.

\* Curiously, the optimal model,  $O$ , is the lowest probability model in these cases.

- Equal prior probabilities are assumed for the four models.
- The unit information  $g$ -prior is used for the parameters.
- We consider the more realistic scenario where the variance  $\sigma^2$  is unknown and assigned the usual objective prior  $1/\sigma^2$ , risks being computed in this setting.

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  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ ,  $\beta_1$ ,  $\beta_2$ , and the random error vector  $\epsilon$  (or  $\mathbf{Y}$ ). But, because the predictive Bayes risks only depend on  $r_{12}$ ,  $r_{1y}$  and  $r_{2y}$ , we can reduce the study to a three dimensional problem. And, since these are simply correlations, we can 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 Appendix 2.

Tables 4, 5 and 6, in Appendix 2, summarize some features of the simulation study, for the correlation scenarios under the full model, the one-variable model, and the null model, respectively. Those tables present the results separately for the Case 1, Case 2, and Case 3 situations. 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 the cases. 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.

Table 2 summarizes the results, over the three cases, for each of the model correlation scenarios (full, one-variable, and null). The table reports how often the MPM and HPM equal the optimal predictive model ( $O$ ), i.e., the model minimizing (1.2), and presents geometric averages of relative risks of the MPM and HPM to  $O$ .

Here are some observations from Table 2:

- Simpler models are more challenging for the MPM (and HPM); indeed,  $MPM = O$  in 92.1%, 85.2%, and 84.5% 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 HPM do better with larger sample sizes.
- The vast majority of the time, the MPM and HPM are the same model but, when they differ, the MPM is typically better:
  - On average, the MPM does better than the HPM (from the  $M = O \neq H$  and  $M > H$  columns) in 2.7% of the cases; while the HPM does better than the MPM in 0.7% of the cases.
  - When the MPM and HPM are not optimal, the geometric average of the MPM risk (relative to that of  $O$ ) is smaller than the geometric average for the HPM.

Additional insight can be gained by looking at the nature of the ‘failures’ of the MPM and HPM. Figure 2, for the MPM, and Figure 3, for the HPM, show the errors being made, in the numerical study, for each of Case 1, Case 2 and Case 3, under the the full model correlation scenario. Focusing on the MPM for explanation, the color of the dots in Figure 2 indicates which model was the median probability model; thus a blue dot indicates that the median probability model was  $M_{00}$ , because that is the color of  $\alpha_{00}$ . 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  $\alpha_{00}$  as the external vertex, the MPM is the optimal model, while if the blue dot lies within the quadrilateral for which  $\alpha_{10}$  is the external vertex, the MPM is incorrectly saying that  $M_{00}$  is optimal, when actually  $M_{10}$  is optimal.

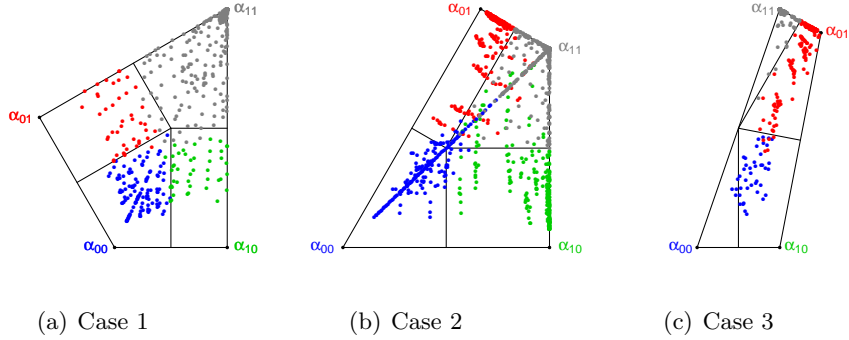


FIG 2. Results from the numerical study under the full model correlation scenario: each dot has the color of the MPM, with the MPM being optimal (or not) if it lies within (or outside) the quadrilateral with external vertex of the same color.

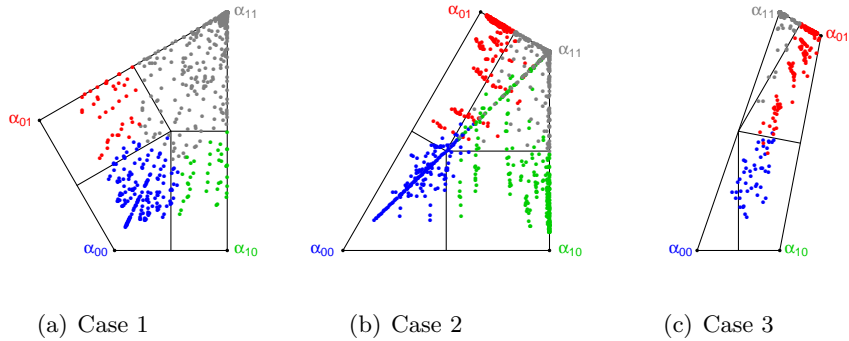


FIG 3. Results from the numerical study under the full model correlation scenario: each dot has the color of the HPM, with the HPM being optimal (or not) if it lies within (or outside) the quadrilateral with external vertex of the same color.

The figures reinforce the earlier messages; Case 1 is nice for the MPM and HPM (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 problematical – 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 HPM, but the improvement is not dramatic.

The interesting feature revealed by the figures is that, essentially always, when the MPM and HPM 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 HPM fail, they may typically do so by choosing too-small models.

#### 4. Generalizations of the optimality of the median probability model.

4.1. *More general priors in the orthogonal design case.* In orthogonal designs, the primary condition for optimality of the median probability model is that  $\widehat{\boldsymbol{\beta}}_\gamma$  is obtained by taking the relevant coordinates of the overall posterior mean  $\bar{\boldsymbol{\beta}}$  (condition (17) of [Barbieri and Berger \(2004\)](#)). With  $\mathbf{X}'\mathbf{X} = \mathbf{D} = \text{diag}\{d_i\}_{i=1}^q$ , the likelihood factors into independent likelihoods for each  $\beta_i$  and thereby any independent product prior

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

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

EXAMPLE 4.2. (*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 \(1993\)](#):  $\pi(\boldsymbol{\beta} | \boldsymbol{\gamma}) = \mathcal{N}_q(\mathbf{0}_q, \mathbf{V}_\gamma)$ , where  $\mathbf{V}_\gamma = \text{diag}\{\gamma_i v_1 + (1 - \gamma_i) v_0\}_{i=1}^q$  with  $v_1 \gg v_0$ . 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

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

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

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

The criterion  $R(\gamma)$  in (1.2) can be then written as

$$R(\gamma) = \sum_{i=1}^q \left( \frac{1}{d_i + v_1^{-1}} - \frac{1}{d_i + v_0^{-1}} \right)^2 (\gamma_i - p_i)^2 d_i z_i^2.$$

which easily seen to be minimized by the MPM model.

4.2. *More flexible priors in nested correlated designs.* Barbieri and Berger (2004) show that the MPM is optimal also for correlated regressors, when considering a sequence of nested models. Here, we generalize the class of priors under which such a statement holds. Assume  $q < n$  and denote with  $\mathbf{T}$  the upper Cholesky triangular matrix such that  $\mathbf{X}'\mathbf{X} = \mathbf{T}'\mathbf{T}$ . Then transform the linear model to

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

where  $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ . Note first that, since  $\mathbf{T}^{-1}$  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  $(\mathbf{X}^*)'\mathbf{X}^* = \mathbf{I}_n$ , so the likelihood factors into independent likelihoods for the  $\beta_i^*$ ; and this independence holds within each of the nested models, since the columns of  $\mathbf{X}^*$  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 Section 4.1 that the median probability model is optimal.

EXAMPLE 4.3. (*Rescaled g-priors*) Suppose the prior for  $\boldsymbol{\beta}^*$  is  $\mathcal{N}_p(\mathbf{0}_p, \mathbf{D})$ , where  $\mathbf{D}$  is diagonal. Any such prior results in optimality of the median probability model. If one transforms back to  $\boldsymbol{\beta}$ , the prior is  $\mathcal{N}_p(\mathbf{0}_p, \mathbf{T}^{-1}\mathbf{D}(\mathbf{T}^{-1})')$  which is considerably richer than the g-type priors considered in Barbieri

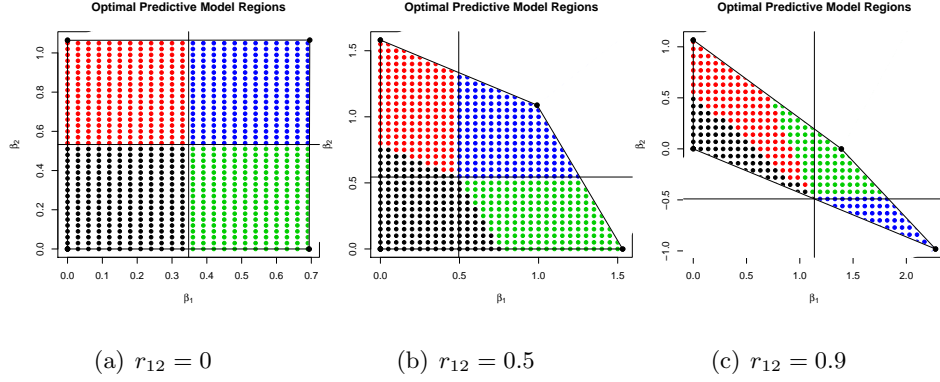


FIG 4. Plots of the optimality regions

and Berger (2004) (which would be these priors with  $\mathbf{D} = g\mathbf{I}$ ). As a specific illustration, suppose

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 + \epsilon & 1 \end{pmatrix},$$

with  $\epsilon > 0$  small. Computation then yields that the prior covariance matrix is

$$\mathbf{T}^{-1}\mathbf{D}(\mathbf{T}^{-1})' = \frac{1}{2\epsilon^2} \begin{pmatrix} 3d_1 & -(3+\epsilon)d_1 \\ -(3+\epsilon)d_1 & [(3+2\epsilon+\frac{1}{3}\epsilon^2)d_1 + \frac{2}{3}\epsilon^2d_2] \end{pmatrix}.$$

The  $g$ -prior choice is  $d_1 = d_2 = g$ , resulting in

$$\mathbf{T}^{-1}\mathbf{D}(\mathbf{T}^{-1})' = \frac{g}{2\epsilon^2} \begin{pmatrix} 3 & -(3+\epsilon) \\ -(3+\epsilon) & [(3+2\epsilon+\frac{1}{3}\epsilon^2) + \frac{2}{3}\epsilon^2] \end{pmatrix}.$$

The alternative choice  $d_1 = g\epsilon^2$ ,  $d_2 = g$  yields

$$\mathbf{T}^{-1}\mathbf{D}(\mathbf{T}^{-1})' = \frac{g}{2} \begin{pmatrix} 3 & -(3+\epsilon) \\ -(3+\epsilon) & [(3+2\epsilon+\frac{1}{3}\epsilon^2) + \frac{2}{3}d_2] \end{pmatrix} \approx \frac{g}{2} \begin{pmatrix} 3 & -3 \\ -3 & \frac{11}{3} \end{pmatrix}.$$

Thus the  $g$ -prior assigns a variance of  $3g/[2\epsilon^2]$  to  $\beta_1$ , while the new prior assigns a variance of  $1.5g$ . This may be much more reasonable in certain contexts.



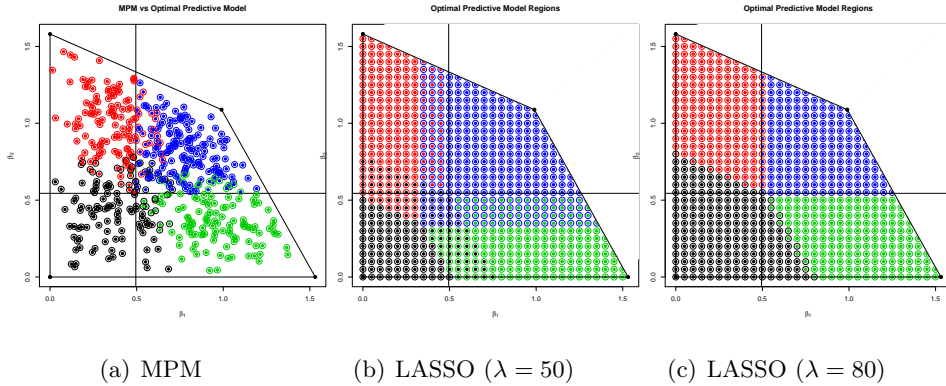


FIG 5. Plots of the actual (full dots) and estimated (round circles) optimal predictive regions

4.3. *The equi-correlated case: optimality can depend on higher order inclusion probabilities.* If we were to visualize the geometry of optimal predictive model selection in the orthogonal case, we would obtain a rectangular partition of a convex hull of posterior means under each model. Figure 4(a) depicts an example of such geometry when  $p = 2$ . The four black dots correspond to the four posterior means  $\hat{\beta}_\gamma$  and comprise a skeleton of a convex hull of all possible locations of the overall posterior mean  $\bar{\beta}$ . Each of these hypothetical means is associated with one optimal predictive model, i.e. the model that is closest to  $\bar{\beta}$  in terms of  $R(\gamma)$ . Denote with  $\hat{\beta}^F = (\hat{\beta}_1^F, \hat{\beta}_2^F)$  the posterior mean under the full model  $\gamma = (1, 1)'$ . When  $r_{12} = 0$ , the optimal predictive model (O) regions are rectangles (marked with 4 colors) where the cuts occur at  $\hat{\beta}_1^F/2$  (vertical line) and  $\hat{\beta}_2^F/2$  (horizontal line). The median probability model is known to be optimal in this case and it can be obtained by element-wise thresholding of  $|\bar{\beta}|$  at  $|\hat{\beta}^F|/2$ . When the predictors are correlated ( $r_{12} = 0.5$  in Figure 4(b)), the regions are no longer rectangular, where simple thresholding of  $\bar{\beta}$  is no longer enough to describe the optimal model. It is worthwhile to note, however, that the full model is the optimal model iff  $|\bar{\beta}| \geq |\hat{\beta}^F|/2$ . Note, also, that Figure 4 (and the later Figure 5) correspond to the Case 2 situation in Section 3.

We focus on the example with  $r_{12} = 0.5$  a bit more closely in Figure 5. On the left, Figure 5(a), we have a comparison with the MPM. The dots correspond to locations of the posterior mean  $\bar{\beta}$ , where the posterior model probabilities were sampled from  $Dir(1, 1, 1, 1)$ . The color of the solid dot designates the optimal predictive model. The color of the round circle sur-

rounding each dot designates the median probability model. We can see an agreement between the MPM and O when the posterior model probabilities put a lot of weight onto one model (corners of the hull). When there is model selection uncertainty (the centre of the hull), the MPM does not have to be optimal. It is interesting to note that the regions of the MPM selection are overlapping suggesting that using only first posterior moments  $\mathbb{E}[\boldsymbol{\gamma} | \mathbf{Y}]$  may not be enough to characterize the optimal model. The following lemma provides a full characterization of the optimal model in terms of both the first and the second moments.

LEMMA 4.1. *Assume  $\mathbf{Q} = \mathbf{X}'\mathbf{X} = (1-r)\mathbf{I}_p + r\mathbf{1}\mathbf{1}'$  for some  $-1 < r < 1$ . Denote  $\boldsymbol{\pi} = \mathbb{E}[\boldsymbol{\gamma} | \mathbf{Y}]$  the vector of posterior inclusion probabilities and  $\boldsymbol{\Pi} = \mathbb{E}\left[\frac{\boldsymbol{\gamma}\boldsymbol{\gamma}'}{1-r+r\|\boldsymbol{\gamma}\|} \mid \mathbf{Y}\right]$ . Under (2.6), the optimal predictive model minimizes*

$$(4.2) \quad R(\boldsymbol{\gamma}) = (1-r) \sum_{i=1}^p \left( \sum_{j=1}^p b_{ij}(\boldsymbol{\gamma}) z_j \right)^2 + r \sum_{i=1}^p \left[ \left( \sum_{j=1}^p b_{ij}(\boldsymbol{\gamma}) \right) z_i \right]^2,$$

where  $\mathbf{Z} = \mathbf{X}'\mathbf{Y} = (z_1, \dots, z_p)'$  and  $\mathbf{B}(\boldsymbol{\gamma}) = \left[ \text{diag}\{\boldsymbol{\gamma} - \boldsymbol{\pi}\} - r \left( \frac{\boldsymbol{\gamma}\boldsymbol{\gamma}'}{1-r+r\|\boldsymbol{\gamma}\|} - \boldsymbol{\Pi} \right) \right]$ .

PROOF. We begin by noting that

$$\mathbf{H}_\gamma \widehat{\boldsymbol{\beta}}_\gamma = \frac{g}{1+g} \frac{1}{1-r} \left( \text{diag}\{\boldsymbol{\gamma}\} - \frac{r\boldsymbol{\gamma}\boldsymbol{\gamma}'}{1-r+r\|\boldsymbol{\gamma}\|} \right) \mathbf{Z},$$

where we used the fact

$$(\mathbf{G} + \mathbf{H})^{-1} = \mathbf{G}^{-1} - \frac{1}{1+a} \mathbf{G}^{-1} \mathbf{H} \mathbf{G}^{-1},$$

for  $\mathbf{G} = (1-r)\mathbf{I}$  and  $\mathbf{H} = r\mathbf{1}\mathbf{1}'$ , where  $a = \text{trace}(\mathbf{H}\mathbf{G}^{-1})$ . Denote with  $\boldsymbol{\pi} = \mathbb{E}[\boldsymbol{\gamma} | \mathbf{Y}]$ . Then  $\mathbf{H}_\gamma \widehat{\boldsymbol{\beta}}_\gamma - \bar{\boldsymbol{\beta}} =$

$$\frac{g}{1+g} \frac{1}{1-r} \left\{ \text{diag}\{\boldsymbol{\gamma} - \boldsymbol{\pi}\} - r \left[ \frac{\boldsymbol{\gamma}\boldsymbol{\gamma}'}{1-r+r\|\boldsymbol{\gamma}\|} - \mathbb{E} \left( \frac{\boldsymbol{\gamma}\boldsymbol{\gamma}'}{1-r+r\|\boldsymbol{\gamma}\|} \mid \mathbf{Y} \right) \right] \right\} \mathbf{Z}.$$

The rest follows from matrix algebra.  $\square$

When  $r = 0$ , we obtain the usual criterion  $R(\boldsymbol{\gamma})$  minimized by the MPM model. The larger the correlation  $r$ , the more weight is put on the joint inclusion probabilities, where the optimal model is the one whose matrix  $\boldsymbol{\gamma}\boldsymbol{\gamma}'$  is closest to the posterior mean of  $\boldsymbol{\gamma}\boldsymbol{\gamma}'$  (normalized by the model size).

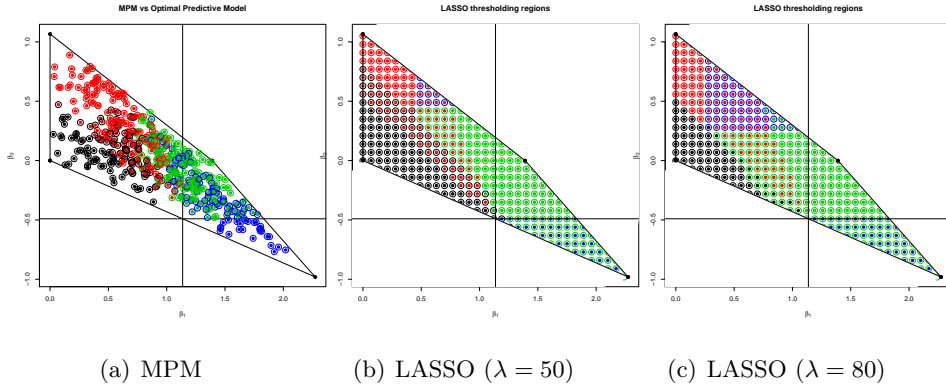


FIG 6. Plots of the actual (full dots) and estimated (round circles) optimal predictive regions

It is also useful to point out that the MPM no longer corresponds to simple thresholding of  $\bar{\beta}$  when  $r \neq 0$ . Because the predictors are correlated, it would seem natural to threshold some functional of  $\bar{\beta}$  which takes into account the correlation. An example of one possible approach is given in [Hahn and Carvalho \(2015\)](#), who suggest running a lasso regression of  $\mathbf{X}\bar{\beta}$  onto  $\mathbf{X}$ . Such a LASSO post-processing step yields a model which summarizes  $\bar{\beta}$  while taking into account the correlation pattern between  $\mathbf{x}$ 's. The regions of such LASSO selected model are depicted in Figure 5(b) and 5(c) (again the solid dots are O and the round circle around are the LASSO selected models). We can see that, indeed, the LASSO selection takes into account the correlation and, interestingly, it can almost exactly match the optimal predictive regions for a suitably chosen hyper-parameter  $\lambda$ . This connection between the LASSO post-processing of  $\bar{\beta}$  and optimal predictive model is curious. However, when the predictors are highly correlated ( $r = 0.9$  in Figure 6), the LASSO regions do not yield the O regions, not even remotely (Figure 6).

**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 and second, by doing an extensive numerical study to see how often the MPM (and HPM) fail to be optimal. The conclusions from the numerical study are given in Section 3, and won’t be repeated here, except to say that the performance of the MPM was overall excellent, even in highly correlated situations, and was measurably better than the performance of the HPM.

The MPM can fail, however, and fail badly, so we finish with a discussion of when this happens, focusing (for simplicity) on the case where there are many replicates of the covariate vector  $\boldsymbol{x}$  in the model; then the median probability model will not include that covariate. Consider four cases.

*Case 1.  $\boldsymbol{x}$  is not 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  $\boldsymbol{x}$ , while the model averaged answer insists on including it.

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

*Case 3.  $\boldsymbol{x}$  is helpful but not crucial for good prediction:* This is like the situation in Section 1.1. The harm in the median probability model ignoring  $\boldsymbol{x}$  may be 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.)

**Appendix 1: Proof of Mini-Theorems.** We denote with  $\boldsymbol{\alpha}_\gamma$  the projection of  $\mathbf{y}$  on the space spanned by the columns of  $\mathbf{X}_\gamma$ . Assume that all variables have been standardized, so that

$$\boldsymbol{\alpha}_{00} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\alpha}_{10} = \begin{pmatrix} a \\ 0 \end{pmatrix}, \quad \boldsymbol{\alpha}_{01} = \begin{pmatrix} b \\ c \end{pmatrix}, \quad \boldsymbol{\alpha}_{11} = \begin{pmatrix} a \\ d \end{pmatrix},$$

with

$$a = r_{1y}, \quad b = r_{12} r_{2y}, \quad c = (1 - r_{12}^2)^{1/2} r_{2y}, \quad d = \frac{r_{2y} - r_{12} r_{1y}}{(1 - r_{12}^2)^{1/2}},$$

where  $r_{12} = \text{Corr}(x_1, x_2)$ ,  $r_{1y} = \text{Corr}(x_1, y)$  and  $r_{2y} = \text{Corr}(x_2, y)$ . Actually the original expression of each coordinate has an irrelevant common factor equal to  $\sqrt{n}$ , which has been ignored. The model average point  $\bar{\boldsymbol{\alpha}}$  has coordinates  $\bar{\alpha}_1$  and  $\bar{\alpha}_2$  given by

$$\begin{pmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \end{pmatrix} = p_{10} \begin{pmatrix} a \\ 0 \end{pmatrix} + p_{01} \begin{pmatrix} b \\ c \end{pmatrix} + p_{11} \begin{pmatrix} a \\ d \end{pmatrix}$$

where  $p_\gamma$  is the posterior probability of model  $M_\gamma$ .

Suppose that we would like to check if the model average point  $\bar{\boldsymbol{\alpha}}$  lies inside a particular triangular subregion of the space  $\{\boldsymbol{\alpha}_{00}, \boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{01}, \boldsymbol{\alpha}_{11}\}$ . To this aim, we express the coordinates of  $\bar{\boldsymbol{\alpha}}$  as a linear combination of the coordinates of the vertexes of the triangular subregion. The model average point is inside the triangular subregion if the weights of the vertexes result to be all positive.

In particular, when we refer to the triangular subregion  $S_1 = \{\boldsymbol{\alpha}_{00}, \boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{11}\}$ , we write the model average point as

$$\begin{pmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \end{pmatrix} = w_{10}^{(1)} \begin{pmatrix} a \\ 0 \end{pmatrix} + w_{11}^{(1)} \begin{pmatrix} a \\ d \end{pmatrix},$$

with  $w_{00}^{(1)} + w_{10}^{(1)} + w_{11}^{(1)} = 1$ , and we may find that:

$$\begin{aligned} w_{00}^{(1)} &= 1 - \frac{\bar{\alpha}_1}{a} \\ w_{10}^{(1)} &= \frac{\bar{\alpha}_1}{a} - \frac{\bar{\alpha}_2}{d} \\ w_{11}^{(1)} &= \frac{\bar{\alpha}_2}{d}. \end{aligned}$$

Note that the sign of each weight gives us information on the position of  $\bar{\boldsymbol{\alpha}}$  with respect to the segment joining the other two vertexes. In fact if one

of the weight is positive, say  $w_{10}^{(1)}$ , this means that  $\bar{\alpha}$  lies on the side of  $\alpha_{10}$  with respect to the line through  $\alpha_{00}$  and  $\alpha_{11}$ . If  $w_{10}^{(1)} < 0$  then  $\bar{\alpha}$  lies on the other side, while if  $w_{10}^{(1)} = 0$  it lies on the segment.

In the same way, when we consider the triangular subregion  $S_2 = \{\alpha_{00}, \alpha_{01}, \alpha_{11}\}$ , we write the model average point as

$$\begin{pmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \end{pmatrix} = w_{01}^{(2)} \begin{pmatrix} b \\ c \end{pmatrix} + w_{11}^{(2)} \begin{pmatrix} a \\ d \end{pmatrix}$$

with  $w_{00}^{(2)} + w_{01}^{(2)} + w_{11}^{(2)} = 1$  and

$$\begin{aligned} w_{00}^{(2)} &= 1 + \frac{(d-c)\bar{\alpha}_1 + (b-a)\bar{\alpha}_2}{ac-bd} \\ w_{01}^{(2)} &= \frac{a\bar{\alpha}_2 - d\bar{\alpha}_1}{ac-bd} \\ w_{11}^{(2)} &= \frac{c\bar{\alpha}_1 - b\bar{\alpha}_2}{ac-bd}. \end{aligned}$$

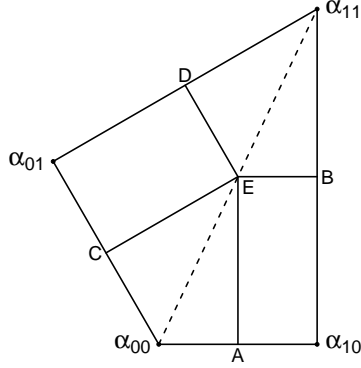
In case 1 and 2 the triangular subregions  $S_1$  and  $S_2$  are disjoint and their union covers the entire space  $\{\alpha_{00}, \alpha_{10}, \alpha_{01}, \alpha_{11}\}$  (see Figure 7).

Note also that to locate the position of the point inside  $S_1$  or  $S_2$  we just need to check the values of the weights  $w^{(1)}$  or  $w^{(2)}$ . In fact in the nested models case the optimal model is the median. Thus, taking into account  $S_1$ , we know that if  $w_{00}^{(1)} > 1/2$  then  $\bar{\alpha}$  lies inside  $\{\bar{\alpha}_{00}, A, E\}$ , if  $w_{11}^{(1)} > 1/2$  inside  $\{\bar{\alpha}_{11}, B, E\}$ , otherwise inside  $\{\bar{\alpha}_{10}, A, E, B\}$ .

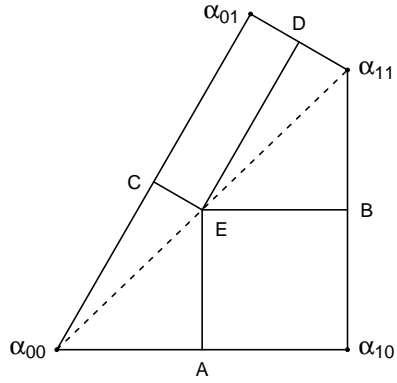
In case 3 the triangular subregions  $S_1$  and  $S_2$  overlap and their union does not cover the entire space  $\{\alpha_{00}, \alpha_{10}, \alpha_{01}, \alpha_{11}\}$  (see Figure 8(a) and 8(b)). However in this case we may refer to  $S_3 = \{\alpha_{10}, \alpha_{01}, E\}$ ,  $S_4 = \{\alpha_{00}, \alpha_{10}, E\}$  and  $S_5 = \{\alpha_{01}, \alpha_{11}, E\}$ , where  $E = \begin{pmatrix} a/2 \\ d/2 \end{pmatrix}$  is the midpoint of the edge linking  $\alpha_{00}$  and  $\alpha_{11}$  (see Figure 8(c)). To locate the position of the point inside  $S_3$ ,  $S_4$  or  $S_5$  we just need to check the value of which of the weights of the two vertexes different from  $E$  is the largest.

In the rest of the section, the weights for these new subregions are reported. In particular, when we refer to the triangular subregion  $S_3 = \{\alpha_{10}, \alpha_{01}, E\}$ , from

$$\begin{pmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \end{pmatrix} = w_{10}^{(3)} \begin{pmatrix} a \\ 0 \end{pmatrix} + w_{01}^{(3)} \begin{pmatrix} b \\ c \end{pmatrix} + w_E^{(3)} \begin{pmatrix} a/2 \\ d/2 \end{pmatrix}$$



(a) Case 1



(b) Case 2

FIG 7. *Subregions*

and  $w_E^{(3)} + w_{10}^{(3)} + w_{01}^{(3)} = 1$ , we obtain

$$w_{10}^{(3)} = \frac{(2c - d) \bar{\alpha}_1 - (2b - a) \bar{\alpha}_2 - ac + bd}{ac + bd - ad}$$

$$w_{01}^{(3)} = \frac{d \bar{\alpha}_1 + a \bar{\alpha}_2 - ad}{ac + bd - ad}$$

$$w_E^{(3)} = 2 \frac{ac - c \bar{\alpha}_1 - (a - b) \bar{\alpha}_2}{ac + bd - ad}.$$

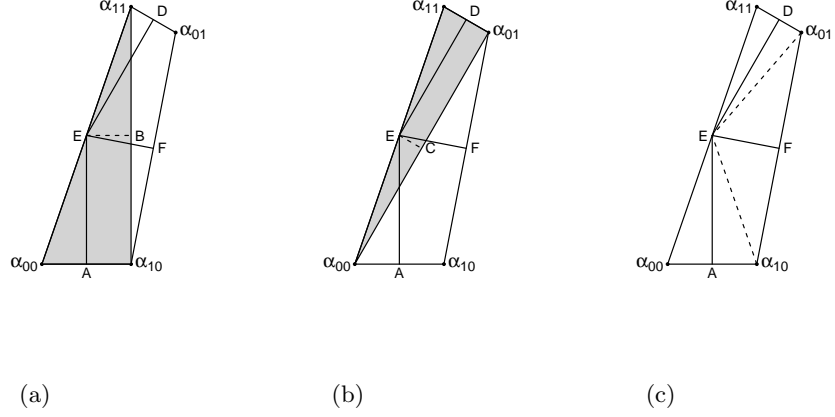


FIG 8. Subregions: Case 3

When we refer to the triangular subregion  $S_4 = \{\alpha_{00}, \alpha_{10}, E\}$ , from

$$\begin{pmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \end{pmatrix} = w_{10}^{(4)} \begin{pmatrix} a \\ 0 \end{pmatrix} + w_E^{(4)} \begin{pmatrix} a/2 \\ d/2 \end{pmatrix}$$

and  $w_E^{(4)} + w_{00}^{(4)} + w_{10}^{(3)} = 1$ , we obtain

$$\begin{aligned} w_{00}^{(4)} &= 1 - \frac{\bar{\alpha}_1}{a} - \frac{\bar{\alpha}_2}{d} \\ w_{10}^{(4)} &= \frac{\bar{\alpha}_1}{a} - \frac{\bar{\alpha}_2}{d} \\ w_E^{(4)} &= 2\frac{\bar{\alpha}_2}{d}. \end{aligned}$$

When we refer to the triangular subregion  $S_5 = \{\alpha_{01}, \alpha_{11}, E\}$ , from

$$\begin{pmatrix} \bar{\alpha}_1 \\ \bar{\alpha}_2 \end{pmatrix} = w_{01}^{(5)} \begin{pmatrix} b \\ c \end{pmatrix} + w_{11}^{(5)} \begin{pmatrix} a \\ d \end{pmatrix} + w_E^{(5)} \begin{pmatrix} a/2 \\ d/2 \end{pmatrix}$$

and  $w_E^{(5)} + w_{01}^{(5)} + w_{11}^{(5)} = 1$ , we obtain

$$\begin{aligned} w_{01}^{(5)} &= \frac{a\bar{\alpha}_2 - d\bar{\alpha}_1}{ac - bd} \\ w_{11}^{(5)} &= \frac{(2c - d)\bar{\alpha}_1 - (2b - a)\bar{\alpha}_2}{ac - bd} - 1 \\ w_E^{(5)} &= 2\frac{(d - c)\bar{\alpha}_1 + (b - a)\bar{\alpha}_2}{ac - bd} + 2. \end{aligned}$$



Conditions under which each model is optimal may be derived using the sets of  $w$ 's weights. In particular,  $M_{00}$  is optimal if:

$$w_{00}^{(1)} \geq \frac{1}{2} \quad w_{00}^{(2)} \geq \frac{1}{2} \quad w_{00}^{(4)} \geq w_{10}^{(4)}.$$

However, since  $w_{00}^{(4)} = w_{10}^{(4)} + 2w_{00}^{(1)} - 1$ , the third condition is equivalent to the first and the first two give:

$$p_1 + p_{01} r_{12} \frac{r_{2y}}{r_{1y}} \leq \frac{1}{2}$$

$$p_2 + p_{10} r_{12} \frac{r_{1y}}{r_{2y}} \leq \frac{1}{2},$$

where  $p_1 = p_{10} + p_{11}$  and  $p_2 = p_{01} + p_{11}$  are the posterior inclusion probabilities of the two covariates.

Model  $M_{10}$  is optimal if:

$$w_{00}^{(1)} \leq \frac{1}{2} \quad w_{00}^{(1)} + w_{10}^{(1)} = 1 - w_{11}^{(1)} \geq \frac{1}{2} \quad w_{10}^{(3)} \geq w_{01}^{(3)} \quad w_{10}^{(4)} \geq w_{00}^{(4)}.$$

Where, as before, the last condition is equivalent to the first and the other three may be restated as:

$$p_1 + p_{01} r_{12} \frac{r_{2y}}{r_{1y}} \geq \frac{1}{2}$$

$$p_2 + p_{01} r_{12} \frac{r_{1y}}{r_{2y}} \frac{1 - r_{12} \frac{r_{2y}}{r_{1y}}}{1 - r_{12} \frac{r_{1y}}{r_{2y}}} \leq \frac{1}{2}$$

$$\left( \frac{r_{1y}}{r_{2y}} \right)^2 \left[ \left( 1 - r_{12} \frac{r_{2y}}{r_{1y}} \right) p_1 - \frac{1}{2} \right] \geq \left[ \left( 1 - r_{12} \frac{r_{1y}}{r_{2y}} \right) p_2 - \frac{1}{2} \right].$$

Model  $M_{01}$  is optimal if:

$$w_{00}^{(2)} \leq \frac{1}{2} \quad w_{00}^{(2)} + w_{01}^{(2)} = 1 - w_{11}^{(2)} \geq \frac{1}{2} \quad w_{10}^{(3)} \leq w_{01}^{(3)} \quad w_{01}^{(5)} \geq w_{11}^{(5)}.$$

Since  $w_{11}^{(5)} = 2w_{11}^{(2)} + w_{01}^{(5)} - 1$ , the last condition is equivalent to the second and the first three give:

$$p_1 + p_{10} r_{12} \frac{r_{2y}}{r_{1y}} \frac{1 - r_{12} \frac{r_{1y}}{r_{2y}}}{1 - r_{12} \frac{r_{2y}}{r_{1y}}} \leq \frac{1}{2}$$

$$p_2 + p_{10} r_{12} \frac{r_{1y}}{r_{2y}} \geq \frac{1}{2}$$

$$\left( \frac{r_{1y}}{r_{2y}} \right)^2 \left[ \left( 1 - r_{12} \frac{r_{2y}}{r_{1y}} \right) p_1 - \frac{1}{2} \right] \leq \left[ \left( 1 - r_{12} \frac{r_{1y}}{r_{2y}} \right) p_2 - \frac{1}{2} \right].$$

Finally  $M_{11}$  is optimal if:

$$w_{11}^{(1)} \geq \frac{1}{2} \quad w_{11}^{(2)} \geq \frac{1}{2} \quad w_{01}^{(5)} \leq w_{11}^{(5)}.$$

Where, as before, the third is equivalent to the second and the first two may be restated as:

$$p_2 + p_{01} r_{12} \frac{r_{1y}}{r_{2y}} \frac{1 - r_{12} \frac{r_{2y}}{r_{1y}}}{1 - r_{12} \frac{r_{1y}}{r_{2y}}} \geq \frac{1}{2}$$

$$p_1 + p_{10} r_{12} \frac{r_{2y}}{r_{1y}} \frac{1 - r_{12} \frac{r_{1y}}{r_{2y}}}{1 - r_{12} \frac{r_{2y}}{r_{1y}}} \geq \frac{1}{2}.$$

The same conclusions may be obtained using the risks. In fact:

$$R(M_{10}) - R(M_{00}) = 2a^2 \left( w_{00}^{(1)} - \frac{1}{2} \right)$$

$$R(M_{01}) - R(M_{00}) = 2(b^2 + c^2) \left( w_{00}^{(2)} - \frac{1}{2} \right)$$

$$R(M_{11}) - R(M_{10}) = 2d^2 \left( \frac{1}{2} - w_{11}^{(1)} \right)$$

$$R(M_{11}) - R(M_{01}) = 2(a^2 + d^2 - b^2 - c^2) \left( \frac{1}{2} - w_{11}^{(2)} \right)$$

$$R(M_{01}) - R(M_{10}) = 2(ac + bd - ad) \left( w_{10}^{(3)} - w_{01}^{(3)} \right)$$

where all multiplying constants are positive.

After setting

$$A_1 = r_{12} \frac{r_{1y}}{r_{2y}} \quad \text{and} \quad A_2 = r_{12} \frac{r_{2y}}{r_{1y}},$$

we may restate the optimality conditions of each model as follows.

$M_{00}$  is optimal if

$$(5.1) \quad \begin{aligned} p_1 + p_{01} A_2 &\leq \frac{1}{2} \\ p_2 + p_{10} A_1 &\leq \frac{1}{2}, \end{aligned}$$

$M_{10}$  is optimal if

$$(5.2) \quad \begin{aligned} p_1 + p_{01} A_2 &\geq \frac{1}{2} \\ p_2 + p_{01} A_1 \frac{1 - A_2}{1 - A_1} &\leq \frac{1}{2} \\ \left( \frac{r_{1y}}{r_{2y}} \right)^2 \left[ (1 - A_2) p_1 - \frac{1}{2} \right] &\geq \left[ (1 - A_1) p_2 - \frac{1}{2} \right], \end{aligned}$$

$M_{01}$  is optimal if

$$(5.3) \quad \begin{aligned} p_1 + p_{10} A_2 \frac{1 - A_1}{1 - A_2} &\leq \frac{1}{2} \\ p_2 + p_{10} A_1 &\geq \frac{1}{2} \\ \left(\frac{r_{1y}}{r_{2y}}\right)^2 \left[ (1 - r A_2) p_1 - \frac{1}{2} \right] &\leq \left[ (1 - A_1) p_2 - \frac{1}{2} \right], \end{aligned}$$

$M_{11}$  is optimal if

$$(5.4) \quad \begin{aligned} p_2 + p_{01} A_1 \frac{1 - A_2}{1 - A_1} &\geq \frac{1}{2} \\ p_1 + p_{10} A_2 \frac{1 - A_1}{1 - A_2} &\geq \frac{1}{2}. \end{aligned}$$

Case 1	Case 2	Case 3
$A_1 < 0$	$0 < A_1 < 1$	$0 < A_1 < 1$
$A_2 < 0$	$0 < A_2 < 1$	$1 < A_1$
$B_1 < 0$	$0 < B_1$	$B_1 < 0$
$B_2 < 0$	$0 < B_2$	$B_2 < 0$

TABLE 3  
*Characterization of possible scenarios in term of  $A_1, A_2, B_1$  and  $B_2$ .*

From the optimality conditions and the results in Table 3, where

$$B_1 = A_1 \frac{1 - A_2}{1 - A_1} \quad \text{and} \quad B_2 = A_2 \frac{1 - A_1}{1 - A_2},$$

the Mini-Theorems 1-7 in Theorem 3.1 from Section 3.2 follow.

**Appendix 2: Details from the Numerical Study.** We first discuss the choice of the correlation ranges adopted in the numerical studies. The idea is to find, for each possible true model – null, one-variable and full – the natural ranges of  $r_{1y}$  and  $r_{2y}$ , in the sense of spanning the high probability region of data arising from the true model.

We do the computations in this appendix without standardizing variables, so that  $\beta_1$  and  $\beta_2$  in the true model do not change with  $n$ . Thus  $r_{12} = \mathbf{x}_1' \mathbf{x}_2 / [\|\mathbf{x}_1\| \|\mathbf{x}_2\|]$ . Note that, with  $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{I})$ ,  $Z_i = \mathbf{x}_i' \boldsymbol{\varepsilon} \sim N(0, \|\mathbf{x}_i\|^2)$ ,

$$Z_i^* = \frac{Z_i}{\|\mathbf{x}_i\|} \sim N(0, 1), \text{ and } \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \sim \chi_n^2,$$

$$\begin{aligned} \|\mathbf{y}\|^2 &= \|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\|^2 = \|\mathbf{x}_1\|^2\beta_1^2 + \|\mathbf{x}_2\|^2\beta_2^2 + 2r_{12}\|\mathbf{x}_1\|\|\mathbf{x}_2\|\beta_1\beta_2 + 2Z_1\beta_1 + 2Z_2\beta_2 + \chi_n^2, \\ r_{1y} &= \frac{\mathbf{x}_1'\mathbf{y}}{\|\mathbf{x}_1\|\|\mathbf{y}\|} = \frac{\mathbf{x}_1'[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}]}{\|\mathbf{x}_1\|\|\mathbf{y}\|} = \frac{\|\mathbf{x}_1\|^2\beta_1 + r_{12}\|\mathbf{x}_1\|\|\mathbf{x}_2\|\beta_2 + Z_1}{\|\mathbf{x}_1\|\|\mathbf{y}\|} = \\ &= \frac{\|\mathbf{x}_1\|\beta_1 + r_{12}\|\mathbf{x}_2\|\beta_2 + Z_1^*}{\|\mathbf{y}\|}, \\ r_{2y} &= \frac{\mathbf{x}_2'\mathbf{y}}{\|\mathbf{x}_2\|\|\mathbf{y}\|} = \frac{\mathbf{x}_2'[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}]}{\|\mathbf{x}_2\|\|\mathbf{y}\|} = \frac{\|\mathbf{x}_2\|^2\beta_2 + r_{12}\|\mathbf{x}_1\|\|\mathbf{x}_2\|\beta_1 + Z_2}{\|\mathbf{x}_2\|\|\mathbf{y}\|} = \\ &= \frac{\|\mathbf{x}_2\|\beta_2 + r_{12}\|\mathbf{x}_1\|\beta_1 + Z_2^*}{\|\mathbf{y}\|}. \end{aligned}$$

**When the full model is true:** There is nothing unusual about the behavior of  $r_{1y}$  and  $r_{2y}$ , so they are allowed to vary independently over the grid  $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ , but with  $r_{1y} \leq r_{2y}$  to eliminate duplicates. Also, only correlations for which the resulting correlation matrix is positive definite are considered.

**When the null model is true:** Now the expressions above become

$$\|\mathbf{y}\|^2 = \chi_n^2, \quad r_{1y} = \frac{Z_1^*}{\sqrt{\chi_n^2}}, \quad r_{2y} = \frac{Z_2^*}{\sqrt{\chi_n^2}}.$$

So, if we want to cover, say, 90% of the probability range of the  $r_{iy}$ , we should use a grid such as

$$\left\{ \frac{0.2}{\sqrt{n}}, \frac{0.4}{\sqrt{n}}, \frac{0.6}{\sqrt{n}}, \frac{0.8}{\sqrt{n}}, \frac{1.0}{\sqrt{n}}, \frac{1.2}{\sqrt{n}}, \frac{1.4}{\sqrt{n}}, \frac{1.6}{\sqrt{n}}, \frac{1.8}{\sqrt{n}} \right\},$$

again with  $r_{1y} \leq r_{2y}$  and keeping only those for which the resulting correlation matrix is positive definite. (For small  $n$ , one would want to use a grid from the  $t$ -distribution with  $n$  degrees of freedom, since that is the distribution of the  $r_{iy}$  but, for the numerical study, this is not necessary.)

**When  $\beta_1 = 0$  and  $\beta_2 \neq 0$ :** Now the expressions above become

$$\begin{aligned} \|\mathbf{y}\|^2 &= \|\mathbf{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi_n^2, \\ r_{1y} &= \frac{r_{12}\|\mathbf{x}_2\|\beta_2 + Z_1^*}{\sqrt{\|\mathbf{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi_n^2}} \cong \frac{r_{12}\|\mathbf{x}_2\|\beta_2}{\sqrt{\|\mathbf{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi_n^2}}, \\ r_{2y} &= \frac{\|\mathbf{x}_2\|\beta_2 + Z_2^*}{\sqrt{\|\mathbf{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi_n^2}} \cong \frac{\|\mathbf{x}_2\|\beta_2}{\sqrt{\|\mathbf{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi_n^2}}, \end{aligned}$$

	M=H=O	M=H≠O	M=O≠H	H=O≠M	H>M both ≠ O	M>H both ≠ O	GM $\frac{R(\gamma^M)}{R(\gamma^O)}$	GM $\frac{R(\gamma^H)}{R(\gamma^O)}$
Case 1: Full model scenario								
n=10	158	5	14	3	0	0	1.01	1.03
n=50	177	1	2	0	0	0	1.001	1.006
n=100	178	0	2	0	0	0	1	1.003
Case 2: Full model scenario								
n=10	156	52	10	0	4*	0	1.11	1.15
n=50	200	3	11	0	8*	0	1.04	1.13
n=100	206	4	12	0	0	0	1.01	1.13
Case 3: Full model scenario								
n=10	90	36	3	0	0	3*	1.13	1.15
n=50	128	4	0	0	0	0	1.01	1.01
n=100	128	4	0	0	0	0	1.02	1.02
Cases combined: Full model scenario								
n=10	404	93	27	3	4*	3*	1.08	1.10
n=50	505	8	13	0	8*	0	1.02	1.06
n=100	512	8	14	0	0	0	1.02	1.06
Overall	1421	109	54	3	12*	3*	1.03	1.07
	88.7%	6.8%	3.4%	0.2%	0.7%	0.2%		

TABLE 4

A summary of the numerical study in the two variable case. Legend:  $H = HPM$ ,  $M = MPM$ ,  $O =$  optimal predictive model;  $M>H$  means that  $MPM$  has smaller (1.2) than  $HPM$ ;  $H>M$  means that  $HPD$  has smaller (1.2) than  $MPM$ ; and  $GM$  is the geometric mean of relative risks (to the optimal model) when  $MPM$  or  $HPM$  is not optimal.

\* Curiously, the optimal model,  $O$ , is the lowest probability model in these cases.

the last approximations following because the  $Z_i^*$  are  $O(1)$  and the other terms are  $O(\sqrt{n})$ . As in the full model case, both correlations are  $O(1)$ , so nothing has to go to zero. But note that

$$r_{1y} \cong r_{12}r_{2y}.$$

Since the error in the approximation is  $O(1/\sqrt{n})$  (and looks to be smaller than  $1/\sqrt{n}$ ), this suggests gridding  $r_{2y}$  in the usual way (from 0.1 to 0.9) and then using a grid for  $r_{1y}$  such as

$$\left\{ \left( r_{12}r_{2y} + \frac{h}{\sqrt{n}} \right), \quad h \in \{-0.9, -0.7, -0.5, -0.3, -0.1, 0.1, 0.3, 0.5, 0.7, 0.9\} \right\},$$

again with  $r_{1y} \leq r_{2y}$  and keeping only those for which the resulting correlation matrix is positive definite.

	M=H=O	M=H≠O	M=O≠H	H=O≠M	H>M both ≠ O	M>H both ≠ O	GM $\frac{R(\gamma^M)}{R(\gamma^O)}$	GM $\frac{R(\gamma^H)}{R(\gamma^O)}$
Case 1: $\beta_1 = 0$ and $\beta_2 \neq 0$ scenario								
n=10	119	2	7	2	0	0	1.007	1.012
n=50	75	0	0	0	0	0	1	1
n=100	45	0	0	0	0	0	1	1
Case 2: $\beta_1 = 0$ and $\beta_2 \neq 0$ scenario								
n=10	162	79	3	0	3*	0	1.136	1.036
n=50	288	29	2	1	8*	0	1.095	1.085
n=100	325	25	0	0	0	0	1.058	1.058
Case 3: $\beta_1 = 0$ and $\beta_2 \neq 0$ scenario								
n=10	143	88	8	0	0	4*	1.126	1.151
n=50	298	28	1	0	1	0	1.028	1.026
n=100	312	40	1	0	0	1*	1.037	1.039
Cases combined: $\beta_1 = 0$ and $\beta_2 \neq 0$ scenario								
n=10	424	169	18	2	3*	4*	1.104	1.114
n=50	661	57	3	1	9	0	1.054	1.049
n=100	682	65	1	0	0	1*	1.045	1.046
Overall	1767	291	22	3	12	5*	1.065	1.067
	84.1%	13.9%	1.1%	0.1%	0.6%	0.2%		

TABLE 5

A summary of the numerical study in the two variable case. Legend:  $H = \text{HPM}$ ,  $M = \text{MPM}$ ,  $O = \text{optimal predictive model}$ ;  $M>H$  means that MPM has smaller (1.2) than HPM;  $H>M$  means that HPD has smaller (1.2) than MPM; and GM is the geometric mean of relative risks (to the optimal model) when MPM or HPM is not optimal.

\* Curiously, the optimal model,  $O$ , is the lowest probability model in these cases.

	M=H=O	M=H≠O	M=O≠H	H=O≠M	H>M both ≠ O	M>H both ≠ O	GM $\frac{R(\gamma^M)}{R(\gamma^O)}$	GM $\frac{R(\gamma^H)}{R(\gamma^O)}$
Case 1: Null model scenario								
n=10	268	16	35	2	0	0	1.01	1.04
n=50	382	5	11	3	0	0	1.002	1.008
n=100	397	2	4	1	0	1	1.0009	1.0036
Case 2: Null model scenario								
n=10	159	70	3	0	7*	0	1.12	1.09
n=50	233	6	0	0	0	0	1.006	1.006
n=100	239	0	0	0	0	0	1	1
Case 3: Null model scenario								
n=10	43	92	12	0	1	7*	1.38	1.37
n=50	67	87	2	0	0	2*	1.20	1.21
n=100	99	58	1	0	0	0	1.03	1.12
Cases combined: Null model scenario								
n=10	470	178	50	2	8	7*	1.106	1.120
n=50	682	98	13	3	0	2*	1.039	1.045
n=100	735	60	5	1	0	1	1.023	1.024
Overall	1887	336	68	6	8	10	1.054	1.060
	81.6%	14.5%	2.9%	0.3%	0.3%	0.4%		

TABLE 6

A summary of the numerical study in the two variable case. Legend:  $H = HPM$ ,  $M = MPM$ ,  $O =$  optimal predictive model;  $M>H$  means that MPM has smaller (1.2) than HPM;  $H>M$  means that HPD has smaller (1.2) than MPM; and GM is the geometric mean of relative risks (to the optimal model) when MPM or HPM is not optimal.

\* Curiously, the optimal model,  $O$ , is the lowest probability model in these cases.

## References.

- BARBIERI, M. M. and BERGER, J. O. (2004). Optimal Predictive Model Selection. *The Annals of Statistics* **32** 870–897.
- BAYARRI, S., BERGER, J. O., FORTE, A. and GARCIA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics* **40** 1550–1577.
- BERGER, J. O. and PERICCHI, R. L. (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison (with discussion). In *Model Selection* (P. Lahiri, ed.) 135–207. Institute of Mathematical Statistics Lecture Notes- Monograph Series, volume 38.
- 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** 80–101.
- CUI, W. and GEORGE, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference* **138** 888–900.
- DRACHAL, K. (2018). Comparison between Bayesian and information-theoretic model averaging: Fossil fuels prices example. *Energy Economics*.
- FELDKIRCHER, M. (2012). Forecast combination and Bayesian model averaging: A prior sensitivity analysis. *Journal of Forecasting* **31** 361–376.
- FOUSKAKIS, D., NTZOUFRAS, I. and PERRAKIS, K. (2018). Power-expected-posterior priors for generalized linear models. *Bayesian Analysis* **13** 721–748.
- GARCIA-DONATO, G. and MARTINEZ-BENEITO, M. A. (2013). On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association* **108** 340–352.
- GEORGE, E. I. (2010). Dilution Priors: Compensating for Model Space Redundancy. *IMS Collections: Borrowing Strength: Theory Powering Applications - A. Festschrift for Lawrence D. Brown* **6** 158–165.
- GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association* **88** 881–889.
- GHOSH, J. (2015). Bayesian model selection using the median probability model. *Wiley Interdisciplinary Reviews: Computational Statistics* **7** 185–193.
- HAHN, R. and CARVALHO, C. (2015). Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective. *Journal of the American Statistical Association* **110** 435–448.
- 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.
- 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** 410–423.
- PIIRONEN, J. and VEHTARI, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing* **27** 711–735.
- ROČKOVÁ, V. (2018). Bayesian Estimation of Sparse Signals with a Continuous Spike-and-Slab Prior. *The Annals of Statistics* **46** 401–437.
- ROČKOVÁ, V. and GEORGE, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association* **109** 828–846.
- 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.
- SCOTT, J. G. and CARVALHO, C. (2008). Feature-inclusion stochastic search for Gaussian



- graphical models. *Journal of Computational and Graphical Statistics* **17** 790–808.
- 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.