Supplementary Material to "The Median Probability Model and Correlated Variables"

Maria M. Barbieri, James O. Berger, Edward I. George, Veronika Ročková

Appendix 1: Proof of Theorem 1. ("Mini-theorems")

We denote with $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$ the projection of $\boldsymbol{y}$ on the space spanned by the columns of $\mathbf{X}_{\boldsymbol{\gamma}}$. 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

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

where $r_{12} = Corr(x_1, x_2)$, $r_{1y} = Corr(x_1, y)$ and $r_{2y} = Corr(x_2, y)$. 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 $p_{\boldsymbol{\gamma}}$ is the posterior probability of model $M_{\boldsymbol{\gamma}}$.

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.

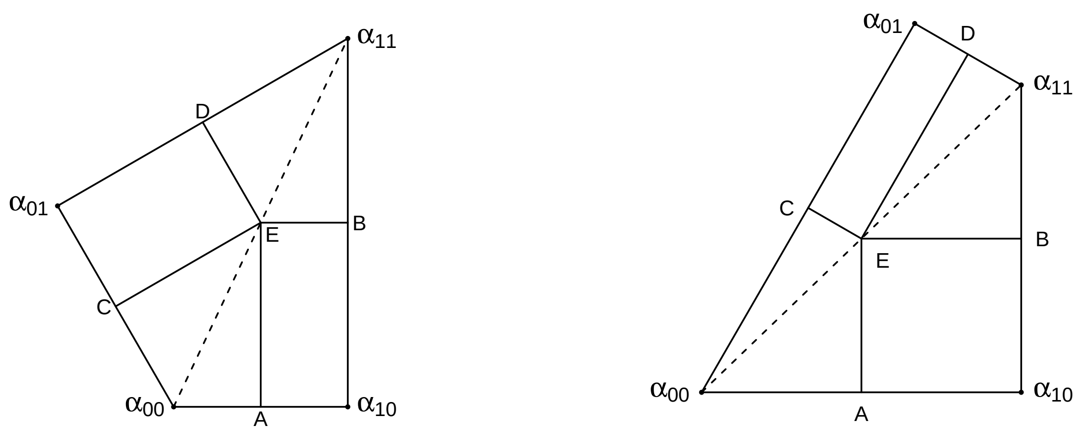
In particular, when we refer to the triangular subregion $S_1 = \{\boldsymbol{\alpha}_{00}, \boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{11}\}$, we write the model average point as

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

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



(a) Case 1 (b) Case 2

Figure S.1: Subregions

that $\bar{\boldsymbol{\alpha}}$ lies on the side of $\boldsymbol{\alpha}_{10}$ with respect to the line through $\boldsymbol{\alpha}_{00}$ and $\boldsymbol{\alpha}_{11}$. If $w_{10}^{(1)} < 0$ then $\bar{\boldsymbol{\alpha}}$ 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 $S_2 = \{\boldsymbol{\alpha}_{00}, \boldsymbol{\alpha}_{01}, \boldsymbol{\alpha}_{11}\}$, we write the model average point as

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

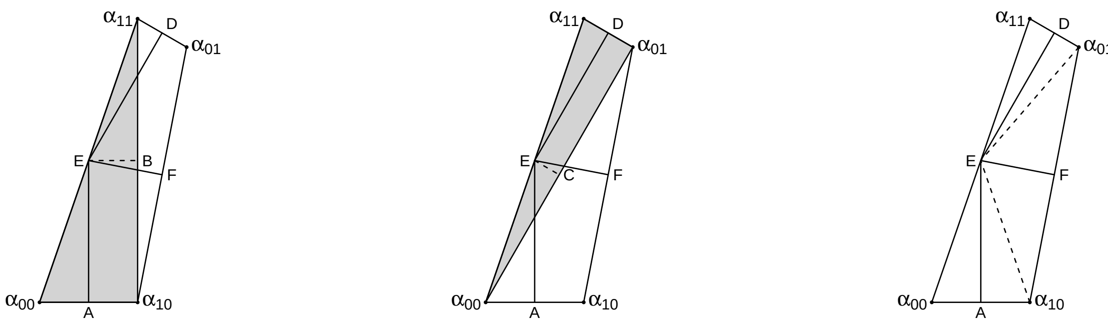
with $w_{00}^{(2)} + w_{01}^{(2)} + w_{11}^{(2)} = 1$ and

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

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

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 $\bar{\boldsymbol{\alpha}}$ lies inside $\{\bar{\boldsymbol{\alpha}}_{00}, A, E\}$, if $w_{11}^{(1)} > 1/2$ inside $\{\bar{\boldsymbol{\alpha}}_{11}, B, E\}$, otherwise inside $\{\bar{\boldsymbol{\alpha}}_{10}, A, E, B\}$.

In case 3 the triangular subregions S_1 and S_2 overlap and their union does not cover the entire space $\{\boldsymbol{\alpha}_{00}, \boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{01}, \boldsymbol{\alpha}_{11}\}$ (see Figure 2(a) and 2(b)). However in this case we may refer to $S_3 = \{\boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{01}, E\}$, $S_4 = \{\boldsymbol{\alpha}_{00}, \boldsymbol{\alpha}_{10}, E\}$ and $S_5 = \{\boldsymbol{\alpha}_{01}, \boldsymbol{\alpha}_{11}, E\}$, where $E = \begin{pmatrix} a/2 \\ d/2 \end{pmatrix}$ is the midpoint of the edge linking $\boldsymbol{\alpha}_{00}$ and $\boldsymbol{\alpha}_{11}$ (see Figure 2(c)). To locate the position of the point



(a) (b) (c)

Figure S.2: Subregions: Case 3

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 $S_3 = \{\boldsymbol{\alpha}_{10}, \boldsymbol{\alpha}_{01}, E\}$, from

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

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

$$\begin{aligned} w_{10}^{(3)} &= \frac{(2c-d)\,\bar{\alpha}_1 - (2b-a)\,\bar{\alpha}_2 - ac + bd}{ac + bd - ad} \\ w_{01}^{(3)} &= \frac{d\,\bar{\alpha}_1 + a\,\bar{\alpha}_2 - ad}{ac + bd - ad} \\ w_E^{(3)} &= 2\frac{ac - c\,\bar{\alpha}_1 - (a-b)\,\bar{\alpha}_2}{ac + bd - ad}. \end{aligned}$$

When we refer to the triangular subregion $S_4 = \{\boldsymbol{\alpha}_{00}, \boldsymbol{\alpha}_{10}, E\}$, 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 $S_5 = \{\boldsymbol{\alpha}_{01}, \boldsymbol{\alpha}_{11}, E\}$, 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:

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

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

$$\begin{aligned}
p_1 + p_{01}\,r_{12}\frac{r_{2y}}{r_{1y}} &\leq \frac{1}{2} \\
p_2 + p_{10}\,r_{12}\frac{r_{1y}}{r_{2y}} &\leq \frac{1}{2},
\end{aligned}$$

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:

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

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

$$\begin{aligned}
& p_1 + p_{01}\,r_{12}\frac{r_{2y}}{r_{1y}} \geq \frac{1}{2} \\
& p_2 + p_{01}\,r_{12}\frac{r_{1y}}{r_{2y}}\frac{1 - r_{12}\frac{r_{2y}}{r_{1y}}}{1 - r_{12}\frac{r_{1y}}{r_{2y}}} \leq \frac{1}{2} \\
& \left(\frac{r_{1y}}{r_{2y}}\right)^2 \left[\left(1 - r_{12}\frac{r_{2y}}{r_{1y}}\right) p_1 - \frac{1}{2}\right] \geq \left[\left(1 - r_{12}\frac{r_{1y}}{r_{2y}}\right) p_2 - \frac{1}{2}\right].
\end{aligned}$$

Model M_{01} is optimal if:

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

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

$$\begin{aligned}
& p_1 + p_{10}\,r_{12}\frac{r_{2y}}{r_{1y}}\frac{1 - r_{12}\frac{r_{1y}}{r_{2y}}}{1 - r_{12}\frac{r_{2y}}{r_{1y}}} \leq \frac{1}{2} \\
& p_2 + p_{10}\,r_{12}\frac{r_{1y}}{r_{2y}} \geq \frac{1}{2} \\
& \left(\frac{r_{1y}}{r_{2y}}\right)^2 \left[\left(1 - r_{12}\frac{r_{2y}}{r_{1y}}\right) p_1 - \frac{1}{2}\right] \leq \left[\left(1 - r_{12}\frac{r_{1y}}{r_{2y}}\right) p_2 - \frac{1}{2}\right].
\end{aligned}$$

Finally M_{11} is optimal if:

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

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:

$$\begin{aligned}
R(M_{10}) - R(M_{00}) &= 2\, a^2 \left(w_{00}^{(1)} - \frac{1}{2} \right) \\
R(M_{01}) - R(M_{00}) &= 2\, (b^2 + c^2) \left(w_{00}^{(2)} - \frac{1}{2} \right) \\
R(M_{11}) - R(M_{10}) &= 2\, d^2 \left(\frac{1}{2} - w_{11}^{(1)} \right) \\
R(M_{11}) - R(M_{01}) &= 2\, (a^2 + d^2 - b^2 - c^2) \left(\frac{1}{2} - w_{11}^{(2)} \right) \\
R(M_{01}) - R(M_{10}) &= 2\, (ac + bd - ad) \left(w_{10}^{(3)} - w_{01}^{(3)} \right)
\end{aligned}$$

where all multiplying constants are positive.

After setting

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

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

M_{00} is optimal if

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

M_{10} is optimal if

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

M_{01} is optimal if

$$\begin{aligned}
&p_1 + p_{10}\, A_2 \frac{1 - A_1}{1 - A_2} \leq \frac{1}{2} \\
&p_2 + p_{10}\, A_1 \geq \frac{1}{2} \\
&\left(\frac{r_{1y}}{r_{2y}} \right)^2 \left[(1 - rA_2)\, p_1 - \frac{1}{2} \right] \leq \left[(1 - A_1)\, p_2 - \frac{1}{2} \right],
\end{aligned} \tag{3}$$

M_{11} is optimal if

$$p_2 + p_{01}\, A_1 \frac{1 - A_2}{1 - A_1} \geq \frac{1}{2}$$
$$p_1 + p_{10}\, A_2 \frac{1 - A_1}{1 - A_2} \geq \frac{1}{2}. \tag{4}$$

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 S.1: 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 S.1, where

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

the results 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 β_1 and β_2 in the true model do not change with n. Thus $r_{12} = \boldsymbol{x}_1{}'\boldsymbol{x}_2/[\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|]$. Note that, with $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{I})$, $Z_i = \boldsymbol{x}_i{}'\boldsymbol{\varepsilon} \sim N(0, \|\boldsymbol{x}_i\|^2)$, $Z_i^* = \frac{Z_i}{\|\boldsymbol{x}_i\|} \sim N(0,1)$, and $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \sim \chi^2_n$,

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

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

$$\|\boldsymbol{y}\|^2 = \chi^2_n, \quad r_{1y} = \frac{Z_1^*}{\sqrt{\chi^2_n}}, \quad r_{2y} = \frac{Z_2^*}{\sqrt{\chi^2_n}} .$$

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

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

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

$$\begin{aligned}
\|\boldsymbol{y}\|^2 &= \|\boldsymbol{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi^2_n , \\
r_{1y} &= \frac{r_{12}\|\boldsymbol{x}_2\|\beta_2 + Z_1^*}{\sqrt{|\|\boldsymbol{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi^2_n|}} \approxeq \frac{r_{12}\|\boldsymbol{x}_2\|\beta_2}{\sqrt{|\|\boldsymbol{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi^2_n|}} , \\
r_{2y} &= \frac{\|\boldsymbol{x}_2\|\beta_2 + Z_2^*}{\sqrt{|\|\boldsymbol{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi^2_n|}} \approxeq \frac{\|\boldsymbol{x}_2\|\beta_2}{\sqrt{|\|\boldsymbol{x}_2\|^2\beta_2^2 + 2Z_2\beta_2 + \chi^2_n|}} ,
\end{aligned}$$

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

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

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.

	number of cases	MPM=MAP both=OP (a) %	MPM=MAP both≠OP (b) %	MPM=OP MAP≠OP (c) %	MAP=OP MPM≠OP (d) %	MAP>MPM both≠OP (e) %	MPM>MAP both≠OP (f) %	GM$\frac{R(MPM)}{R(OP)}$	GM$\frac{R(MAP)}{R(OP)}$
Case 1									
n=10	180	87.8	2.8	7.8	1.7	0.0	0.0	1.008	1.029
n=50	180	98.3	0.6	1.1	0.0	0.0	0.0	1.001	1.005
n=100	180	98.9	0.0	1.1	0.0	0.0	0.0	1.000	1.003
Case 2									
n=10	222	70.3	23.4	4.5	0.0	1.8*	0.0	1.110	1.154
n=50	222	90.1	1.4	5.0	0.0	3.6*	0.0	1.036	1.129
n=100	222	92.8	1.8	5.4	0.0	0.0	0.0	1.006	1.130
Case 3									
n=10	132	68.2	27.3	2.3	0.0	0.0	2.3*	1.126	1.152
n=50	132	97.0	3.0	0.0	0.0	0.0	0.0	1.007	1.007
n=100	132	97.0	3.0	0.0	0.0	0.0	0.0	1.017	1.017
Cases combined									
n=10	534	75.7	17.4	5.1	0.6	0.7*	0.6*	1.078	1.110
n=50	534	94.6	1.5	2.4	0.0	1.5*	0.0	1.017	1.056
n=100	534	95.9	1.5	2.6	0.0	0.0	0.0	1.006	1.058
Overall	1602	88.7	6.8	3.4	0.2	0.7*	0.2*	1.033	1.074

Table S.2: The case of two covariates: performance of MPM and MAP under the full model.

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

* denotes cases when OP is the *lowest* probability model.

	number of cases	MPM=MAP both=OP (a)	MPM=MAP both≠OP (b)	MPM=OP MAP≠OP (c)	MAP=OP MPM≠OP (d)	MAP>MPM both≠OP (e)	MPM>MAP both≠OP (f)	GM$\frac{R(MPM)}{R(OP)}$	GM$\frac{R(MAP)}{R(OP)}$
Case 1									
n=10	130	91.5	1.5	5.4	1.5	0.0	0.0	1.007	1.012
n=50	75	100.0	0.0	0.0	0.0	0.0	0.0	1.000	1.000
n=100	45	100.0	0.0	0.0	0.0	0.0	0.0	1.000	1.000
Case 2									
n=10	247	65.6	32.0	1.2	0.0	1.2*	0.0	1.136	1.136
n=50	328	87.8	8.8	0.6	0.3	2.4*	0.0	1.095	1.085
n=100	350	92.9	7.1	0.0	0.0	0.0	0.0	1.058	1.058
Case 3									
n=10	243	58.8	36.2	3.3	0.0	0.0	1.6*	1.126	1.151
n=50	328	90.9	8.5	0.3	0.0	0.3	0.0	1.028	1.026
n=100	354	88.1	11.3	0.3	0.0	0.0	0.3*	1.037	1.039
Cases combined									
n=10	620	68.4	27.3	2.9	0.3	0.5*	0.6*	1.030	1.032
n=50	731	90.4	7.8	0.4	0.1	1.2	0.0	1.019	1.017
n=100	749	91.1	8.7	0.1	0.0	0.0	0.1*	1.016	1.016
Overall	2100	84.1	13.9	1.0	0.1	0.6	0.2*	1.021	1.022

Table S.3: The case of two covariates: performance of MPM and MAP under the one-variable ($\beta_1 = 0$ and $\beta_2 \neq 0$) model.

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

* denotes cases when OP is the *lowest* probability model.

	number of cases	MPM=MAP both=OP (a)	MPM=MAP both≠OP (b)	MPM=OP MAP≠OP (c)	MAP=OP MPM≠OP (d)	MAP>MPM both≠OP (e)	MPM>MAP both≠OP (f)	GM$\frac{R(MPM)}{R(OP)}$	GM$\frac{R(MAP)}{R(OP)}$
Case 1									
n=10	321	83.5	5.0	10.9	0.6	0.0	0.0	1.011	1.038
n=50	401	95.3	1.2	2.7	0.7	0.0	0.0	1.002	1.008
n=100	405	98.0	0.5	1.0	0.2	0.0	0.2	1.001	1.004
Case 2									
n=10	239	66.5	29.3	1.3	0.0	2.9⋆	0.0	1.124	1.090
n=50	239	97.5	2.5	0.0	0.0	0.0	0.0	1.006	1.006
n=100	239	100.0	0.0	0.0	0.0	0.0	0.0	1.000	1.000
Case 3									
n=10	159	29.6	57.9	7.5	0.0	0.6	4.4⋆	1.293	1.356
n=50	159	42.8	54.7	1.3	0.0	0.0	1.3⋆	1.198	1.209
n=100	161	63.4	36.0	0.6	0.0	0.0	0.0	1.116	1.118
Cases combined									
n=10	719	65.9	24.8	7.0	0.3	1.1	1.0⋆	1.032	1.036
n=50	799	85.5	12.3	1.6	0.4	0.0	0.3⋆	1.013	1.015
n=100	805	91.7	7.5	0.6	0.1	0.0	0.1	1.008	1.008
Overall	2323	81.6	14.5	2.9	0.3	0.3	0.4	1.018	1.020

Table S.4: The case of two covariates: performance of MPM and MAP models under the null model.

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

⋆ denotes cases when OP is the *lowest* probability model.

Appendix 3: A Simulation Study

To glean more insights into the predictive optimality of the MPM model, we conduct a simulation study with $q = 5$ covariates. We consider three setups: (1) the full model with $\boldsymbol{b} = (1,1,1,1,1)'$, (2) a sparse model with $\boldsymbol{b} = (1,1,1,0,0)'$ and (3) the null model with $\boldsymbol{b} = (0,0,0,0,0)'$. We assume $\boldsymbol{x}_i \overset{ind}{\sim} \mathcal{N}_5(\boldsymbol{0}_5, \Sigma)$, where $\Sigma = (\sigma_{ij})_{i,j=1}^{5,5}$ is an equi-correlated matrix with $\sigma_{ij} = \rho \times \mathbb{I}(i \neq j) + \mathbb{I}(i = j)$. We also consider various degrees of correlation $\rho \in \{0, 0.5, 0.9, 0.99\}$. For each degree of correlation and a model setting, we generate 1 000 datasets $(\boldsymbol{Y}, \boldsymbol{X})$ assuming $\sigma^2 = 1$. For each dataset we record whether MAP (MPM) was optimal etc. The predictors are recentered and rescaled to have mean 0 and an $\|\cdot\|$ norm $\sqrt{n}$. We assign the unit-information g-prior with $g = n$ and the inverse gamma prior (2.9) with $\eta = \lambda = 1$. We consider two model priors (1) the uniform prior assigning a probability 1/32 on each model (results reported in Table S.6) and (2) the beta-binomial prior with $a = b = 1$ (results reported in Table S.5).

Table S.6 summarizes findings obtained with equal prior model probabilities. We reiterate some of the conclusions obtained earlier in Section 3.3. Again, simpler models are more challenging and both MPM and MAP perform (a) better with larger sample sizes and (b) worse with larger correlations. Note that, unlike when the predictors are orthogonal, MPM is not guaranteed to be optimal when $\rho = 0$. MPM and MAP are seen to agree very often and, again, when they do not agree MPM is better more often. It is interesting to compare Table S.6 with Table S.5 which summarizes results for the beta-binomial prior with $a = b = 1$. We have seen in Section 2.4 that the beta-binomial prior can cope better with variable redundancy. We can see a robust performance for spare and null settings. Interestingly, in all simulated datasets for our setups, the MAP model was the same as the MPM model.

	MPM=MAP both=OP	MPM=MAP both≠OP	MPM=OP MAP≠OP	MAP=OP MPM≠OP	MAP>MPM both≠OP	MPM>MAP both≠OP	MPM=MAP both=OP	MPM=MAP both≠OP	MPM=OP MAP≠OP	MAP=OP MPM≠OP	MAP>MPM both≠OP	MPM>MAP both≠OP
Full model scenario: $\boldsymbol{b} = (1,1,1,1,1)'$												
	$\rho = 0$						$\rho = 0.5$					
n=10	96.4	3.6	0	0	0	0	86.3	13.7	0	0	0	0
n=50	100	0	0	0	0	0	100	0	0	0	0	0
n=100	100	0	0	0	0	0	100	0	0	0	0	0
	$\rho = 0.9$						$\rho = 0.99$					
n=10	36.6	63.4	0	0	0	0	13.8	86.2	0	0	0	0
n=50	98.4	1.6	0	0	0	0	56.2	43.8	0	0	0	0
n=100	100	0	0	0	0	0	78.3	21.7	0	0	0	0
Sparse scenario: $\boldsymbol{b} = (1,1,1,0,0)'$												
	$\rho = 0$						$\rho = 0.5$					
n=10	88.3	11.7	0	0	0	0	75.6	24.4	0	0	0	0
n=50	89.4	10.6	0	0	0	0	60.6	39.4	0	0	0	0
n=100	98.7	1.3	0	0	0	0	96.9	3.1	0	0	0	0
	$\rho = 0.9$						$\rho = 0.99$					
n=10	44.7	55.3	0	0	0	0	29.4	70.6	0	0	0	0
n=50	89.4	10.6	0	0	0	0	60.6	39.4	0	0	0	0
n=100	94.9	5.1	0	0	0	0	66.9	33.1	0	0	0	0
Null model scenario: $\boldsymbol{b} = (0,0,0,0,0)'$												
	$\rho = 0$						$\rho = 0.5$					
n=10	57.8	42.2	0	0	0	0	51.4	48.6	0	0	0	0
n=50	74.6	25.4	0	0	0	0	60.3	39.7	0	0	0	0
n=100	75.5	24.5	0	0	0	0	65	35	0	0	0	0
	$\rho = 0.9$						$\rho = 0.99$					
n=10	52.6	47.4	0	0	0	0	52	48	0	0	0	0
n=50	60.9	39.1	0	0	0	0	62.7	37.3	0	0	0	0
n=100	60.5	39.5	0	0	0	0	67.6	32.4	0	0	0	0

Table S.5: The case of $q = 5$. Performance of MPM and MAP models under the full, one-variable and null models using the beta-binomial prior on the model space with $a = b = 1$ (percentage of cases, out of 1 000 simulated datasets).

Legend: OP = optimal predictive model; MPM>MAP (resp. MAP>MPM) means that MPM (resp. MAP) has a smaller value of risk defined in (1.2) than MAP (resp. MPM).

	MPM=MAP both=OP	MPM=MAP both≠OP	MPM=OP MAP≠OP	MAP=OP MPM≠OP	MAP>MPM both≠OP	MPM>MAP both≠OP	MPM=MAP both=OP	MPM=MAP both≠OP	MPM=OP MAP≠OP	MAP=OP MPM≠OP	MAP>MPM both≠OP	MPM>MAP both≠OP
Full model scenario: $\boldsymbol{b} = (1,1,1,1,1)'$												
	$\rho = 0$						$\rho = 0.5$					
n=10	43.4	30.3	15.2	1.9	1.8	7.4	10	34.8	27.7	1	2.7	23.8
n=50	100	0	0	0	0	0	100	0	0	0	0	0
n=100	100	0	0	0	0	0	100	0	0	0	0	0
	$\rho = 0.9$						$\rho = 0.99$					
n=10	1.5	45.1	8.5	.4	13.9	30.6	8.6	35.9	3.2	1.4	49.3	1.6
n=50	40.6	33.8	23	.1	.1	2.4	4.3	51.8	12	.1	6	25.8
n=100	95.4	2.9	1.7	0	0	0	5.3	52.7	17.5	.1	1.8	22.6
Sparse scenario: $\boldsymbol{b} = (1,1,1,0,0)'$												
	$\rho = 0$						$\rho = 0.5$					
n=10	30.1	45.2	13.1	1.1	3.3	7.2	15.3	45.7	15.2	.8	4.5	18.5
n=50	86.3	11.2	.8	1.6	.1	0	72.4	23.5	2	1.3	.7	.1
n=100	90.7	7.3	.7	.9	.4	0	75.6	20.4	1.9	1.1	.8	.2
	$\rho = 0.9$						$\rho = 0.99$					
n=10	6.4	49.5	12.7	.3	8.6	22.5	14.3	38.1	6.4	2.1	32.3	6.8
n=50	29.3	49.4	12.8	0	1.3	7.2	7.4	57.4	14.4	.4	4	16.4
n=100	53.3	36.9	6.5	.4	1	1.9	10.2	54.8	15.8	.2	4.4	14.6
Null model scenario: $\boldsymbol{b} = (0,0,0,0,0)'$												
	$\rho = 0$						$\rho = 0.5$					
n=10	12.5	56.1	6.8	.7	7.9	16	10.1	57.1	6.4	.7	10	15.7
n=50	36.4	50.7	3.9	1.8	3.6	3.6	16.5	58	7.5	1	8.4	8.6
n=100	52	39.7	3.6	1.6	1.1	2	15.9	61.2	6.5	1.3	6.6	8.5
	$\rho = 0.9$						$\rho = 0.99$					
n=10	9.1	53.7	7.1	1.9	13.3	14.9	8	51.2	8.8	3.7	12.5	15.8
n=50	10.7	52.3	13.8	2.1	9.9	11.2	7.4	55.5	10.3	4.8	8.9	13.1
n=100	13.4	53.2	12.9	1.7	10	8.8	8.4	57.5	9.1	4.6	8.4	12

Table S.6: The case of $q = 5$. Performance of MPM and MAP models under the full, sparse and null settings using the uniform prior on the model space (percentage of cases, out of 1 000 simulated datasets).

Legend: OP = optimal predictive model; MPM>MAP (resp. MAP>MPM) means that MPM (resp. MAP) has a smaller value of risk defined in (1.2) than MAP (resp. MPM).