Supplementary material for: "A Model selection approach for Variable selection with censored data"

María Eugenia Castellanos, Gonzalo García-Donato, Stefano Cabras

1 Proofs of main theoretical results

1.1 Proof of Theorem 1

As expressed in De Santis et al. (2001), for a sample of independent survival data with likelihood function as in equation (1) in the main text, the Fisher information matrix can be written as sum of two expectations, the corresponding with the derivatives of the uncensored part and the censored one. Once the second derivatives are obtained for each part, expectations can be calculated easily because censored times $\boldsymbol{c} = (c_1, \ldots, c_n)$ are known.

Indicating $\boldsymbol{\theta} = (\beta_0, \sigma, \boldsymbol{\beta})$ and let the log-likelihood for model in equation (3) in the main text, the $\log\left(\frac{1}{\sigma}\phi\left(\frac{y_i-\beta_0-\boldsymbol{\beta}^{\mathsf{T}}\widetilde{\boldsymbol{x}}_i}{\sigma}\right)\right)$ and the $\log\left(1-\Phi\left(\frac{c_i-\beta_0-\boldsymbol{\beta}^{\mathsf{T}}\widetilde{\boldsymbol{x}}_i}{\sigma}\right)\right)$ be denoted by $l(\boldsymbol{\theta} \mid \boldsymbol{y})$, $l_i^u(\boldsymbol{\theta})$ and $l_i^c(\boldsymbol{\theta})$, respectively.

The element (j,k) of the block Fisher information matrix corresponding to parameters $\boldsymbol{\beta}$ for the likelihood in equation (2) in the main text is

$$\mathcal{I}(\boldsymbol{\beta})_{j,k} = -E\left(\frac{\partial^2}{\partial\beta_j\partial\beta_k} l(\boldsymbol{\theta})|\boldsymbol{\theta}\right) = -[H_u(\beta_j,\beta_k) + H_c(\beta_j,\beta_k)], \tag{1}$$

with

$$\begin{aligned} H_u(\beta_j,\beta_k) &= \sum_{i=1}^{n} E\left(\frac{\partial^2}{\partial\beta_j\partial\beta_k} l_i^u(\boldsymbol{\theta})|\boldsymbol{\theta}\right) Pr(\delta_i = 1|\boldsymbol{\theta}) \\ H_c(\beta_j,\beta_k) &= \sum_{i=1}^{n} E\left(\frac{\partial^2}{\partial\beta_j\partial\beta_k} l_i^c(\boldsymbol{\theta})|\boldsymbol{\theta}\right) Pr(\delta_i = 0|\boldsymbol{\theta}) \end{aligned}$$

In particular, for likelihood in equation (2) in the text

$$\begin{aligned}\frac{\partial^2}{\partial\beta_j\partial\beta_k}l_i^u(\boldsymbol{\theta}) &= -\frac{1}{\sigma^2}\widetilde{x}_{j,i}\widetilde{x}_{k,i}\\ \frac{\partial^2}{\partial\beta_j\partial\beta_k}l_i^c(\boldsymbol{\theta}) &= -\frac{1}{\sigma^2}\widetilde{x}_{j,i}\widetilde{x}_{k,i}h(z_i)(h(z_i)-z_i)\end{aligned}$$

with $z_i = \frac{c_i-\beta_0-\boldsymbol{\beta}^{\mathrm{T}}\widetilde{\boldsymbol{x}}_i}{\sigma}$.

Assuming censored times c_i, $i = 1,\ldots,n$ are known, the above quantities do not depend on data, so expectations equal themselves, while

$$Pr(\delta_i = 1|\boldsymbol{\theta}) = \Phi(z_i) \text{ and } Pr(\delta_i = 0|\boldsymbol{\theta}) = 1-\Phi(z_i)$$

So the block Fisher information matrix corresponding to parameters $\boldsymbol{\beta}$ is given in equation 13 in the main text.

For the null model, this matrix adopts a simpler expression:

$$\mathcal{I}(\beta_0,\sigma) = \frac{1}{\sigma^2}\begin{pmatrix} i_{11} & i_{12}\\ i_{12} & i_{22}\end{pmatrix} \tag{2}$$

where

$$i_{11} = \sum_{i=1}^{n}\phi(z_{i0})(h(z_{i0})-z_{i0}) + 1^T\Delta_0\mathbf{1},$$

$$i_{12} = \sum_{i=1}^{n}\phi(z_{i0})(1+h(z_{i0})z_{i0}-z_{i0}^2) + 2\sum_{i=1}^{n}\Phi(z_{i0})((1-\Phi(z_{i0}))z_{i0}-\phi(z_{i0})),$$

and

$$\begin{aligned}i_{22} &= \sum_{i=1}^{n}\phi(z_{i0})z_{i0}(2+h(z_{i0})z_{i0}-z_{i0}^2) - 1^T\Delta_0\mathbf{1}\\ &+ 3\sum_{i=1}^{n}\Phi(z_{i0})((1-\Phi(z_{i0}))z_{i0}^2 + \Phi(z_{i0})(1-z_{i0}\frac{\phi(z_{i0})}{\Phi(z_{i0})}),\end{aligned}$$

and $\Delta_0 = Diag\{\Phi(z_{i0})\}$.

1.2 Proof of Theorem 2

i) $w(z)$ can be written as,

$$w(z) = \Phi(z) + (1 - \Phi(z)) \underbrace{h(z)(h(z) - z)}_{\zeta(z)},$$

where $\zeta(z) = h'(z)$, verifies $0 < \zeta(z) < 1$, for all $z \in \mathbb{R}$, (Sampford, 1953). So it is clear that $w(z)$ is greater than 0, is sum of positive quantities, and it is smaller than 1, because:

$$w(z) = \Phi(z) + (1 - \Phi(z))\zeta(z) < \Phi(z) + (1 - \Phi(z)) = 1.$$

To show ii), we consider

$$\begin{aligned} w'(z) &= \phi(z) + -\phi(z)\zeta(z) + (1 - \Phi(z))h(z)((2h(z) - z)(h(z) - z) - 1) \\ &> (1 - \Phi(z))h(z)((2h(z) - z)(h(z) - z) - 1) > 0, \end{aligned}$$

this last result is because it consists in a product of positive quantities, $(1-\Phi(z)) > 0$ and $h(z)((2h(z)-z)(h(z)-z)-1) = h''(z) > 0$, because $h(z)$ is convex in the normal model, (Sampford, 1953).

For iii), it suffices to show that the matrices:

$$\widetilde{\boldsymbol{X}}^{\mathsf{T}} \big(\boldsymbol{W}(\beta_0, \sigma) - \boldsymbol{W}(\beta_0, \sigma) \frac{\mathbf{1}\mathbf{1}^{\mathsf{T}}}{N_{(\beta_0,\sigma)}} \boldsymbol{W}(\beta_0, \sigma) \big) \widetilde{\boldsymbol{X}},$$

and:

$$\sum_{i=1}^{n} w_i (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_w)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_w)^{\mathsf{T}},$$

coincide. To check this equivalence, it can be easily seen that the (l, j) element of the first of these matrices is

$$\sum_{i=1}^{n} w_i x_{il} x_{ij} - N_{(\beta_0,\sigma)} \bar{x}_l \bar{x}_j,$$

which clearly coincides with the same element in the second matrix.

1.3 Proof of Theorem 3

Consider first the limit:

$$\lim_{c_c \to -\infty} m_1(\boldsymbol{y}, \boldsymbol{\delta}).$$

Under the conditions here stated, in the proof of Theorem 4 it is proved that the integral in $m_1(\boldsymbol{y}, \boldsymbol{\delta})$ is uniformly bounded (for any $\boldsymbol{c}$) by an integrable function hence allowing the interchange between the limit and the integral by dominated convergence theorem. This observation jointly with the limiting behavior $N \to n_u$ and $\boldsymbol{\Sigma}^M \to \boldsymbol{\Sigma}^U$ and $1 - \Phi(-\infty) = 1$ leads to

$$\lim_{c_c \to -\infty} m_1(\boldsymbol{y}, \boldsymbol{\delta}) = \int N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u \boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \sigma^{-1} \, N_k(\boldsymbol{\beta} \mid \mathbf{0}, g\, \boldsymbol{\Sigma}^U) \pi(g) d\beta_0 d\boldsymbol{\beta} d\sigma dg. \tag{3}$$

To clarify the situation, notice that the product of normal densities in the integral above adopts (because of the definition of $\widetilde{\boldsymbol{X}}_u$ in (4) and $\boldsymbol{\Sigma}^U$ in (17), both in the main text) the form:

$$\begin{aligned} & N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + [\boldsymbol{X}_u - \frac{1}{n}(\mathbf{1}_{n_u}\mathbf{1}_{n_u}^\intercal \boldsymbol{X}_u + \mathbf{1}_{n_u}\mathbf{1}_{n_c}^\intercal \boldsymbol{X}_c)]\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \\ \times \quad & N_k(\boldsymbol{\beta} \mid \mathbf{0}, \sigma^2\, g\, n_u (\boldsymbol{X}_u^\intercal H_{n_u} \boldsymbol{X}_u)^{-1}), \end{aligned}$$

where, in this proof we use the notation $\boldsymbol{H}_m = (\boldsymbol{I} - \mathbf{1}_m \mathbf{1}_m^\intercal / m)$ for this projection matrix.

Similarly with the marginal under the null:

$$\lim_{c_c \to -\infty} m_0(\boldsymbol{y}, \boldsymbol{\delta}) = \int N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0, \sigma^2 \boldsymbol{I}) \sigma^{-1} \, d\beta_0 d\sigma.$$

Equation (10) in the main body of the text, implies that the ratio of the integrals:

$$\frac{\int N_n(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \sigma^{-1} \, N_k(\boldsymbol{\beta} \mid \mathbf{0}, \sigma^2\, g (\boldsymbol{X}^\intercal \boldsymbol{H}_n \boldsymbol{X})^{-1} n) \pi(g) d\beta_0 d\boldsymbol{\beta} d\sigma dg}{\int N_n(\boldsymbol{y} \mid \mathbf{1}\beta_0, \sigma^2 \boldsymbol{I}) \sigma^{-1} d\beta_0 d\sigma}, \tag{4}$$

equals $\mathcal{B}_\pi(\boldsymbol{y}, \widetilde{\boldsymbol{X}}, k, n) = \mathcal{B}_\pi(\boldsymbol{y}, (\mathbf{1}\, \boldsymbol{X}), k, n)$.

Now, in the integral in (3) perform the change of variables:

$$\beta_0^\star = \beta_0 - n^{-1} \left(\mathbf{1}_{n_u}^\intercal \boldsymbol{X}_u + \mathbf{1}_{n_c}^\intercal \boldsymbol{X}_c\right) \boldsymbol{\beta}$$

with the remaining variables unchanged and with unitary Jacobian. Then

$$\begin{aligned}
&\lim_{c_c \to -\infty} m_1(\boldsymbol{y}, \boldsymbol{\delta}) = \\
= &\int N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0^\star + \boldsymbol{X}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I})\sigma^{-1}\, N_k(\boldsymbol{\beta} \mid \mathbf{0}, \sigma^2\, g\, n_u(\boldsymbol{X}_u^\intercal \boldsymbol{H}_{n_u}\boldsymbol{X}_u)^{-1}) \\
\times &\ \pi(g)d\beta_0^\star d\boldsymbol{\beta} d\sigma dg \\
= &\ \mathcal{B}_\pi(\boldsymbol{y}, (\mathbf{1}\,\boldsymbol{X}_u), k, n_u) \times \int N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0, \sigma^2\boldsymbol{I})\sigma^{-1} d\beta_0 d\sigma \\
= &\ \mathcal{B}_\pi(\boldsymbol{y}, (\mathbf{1}\,\boldsymbol{X}_u), k, n_u) \times \lim_{c_c \to -\infty} m_0(\boldsymbol{y}, \boldsymbol{\delta}),
\end{aligned}$$

where the second identity holds true because of (4).

1.4 Proof of Theorem 4

Denote for this proof:

$$\boldsymbol{S} = \sum_{i=1}^{n} w_i(\boldsymbol{x}_i - \boldsymbol{x}_w)(\boldsymbol{x}_i - \boldsymbol{x}_w)^\intercal,$$

so $\boldsymbol{\Sigma}^M = \sigma^2\, \boldsymbol{S}^{-1}\, N$.

For i):

In this case $\boldsymbol{\Sigma}^M$ equals matrix (9) in the text which exists if $n \geq k+1$. Also, the likelihood (2) in the manuscript is bounded by $N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ implying that

$$m(\boldsymbol{y}, \boldsymbol{\delta}) \leq \int m^U(\boldsymbol{y} \mid g)\pi(g)\, dg$$

where

$$\begin{aligned}
m^U(\boldsymbol{y} \mid g) &= \int N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I})\, N_k(\boldsymbol{\beta} \mid \mathbf{0}, g\boldsymbol{\Sigma}^M)\sigma^{-1}\, d\sigma d\beta_0 d\boldsymbol{\beta} \\
&= \int N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0, \sigma^2(\boldsymbol{I} + g\widetilde{\boldsymbol{X}}_u\boldsymbol{S}^{-1}\widetilde{\boldsymbol{X}}_u^\intercal N))\sigma^{-1}\, d\sigma d\beta_0.
\end{aligned}$$

The above marginal defines implicitly a linear model for $\boldsymbol{y}$:

$$\boldsymbol{y} = \beta_0\mathbf{1} + \sigma\epsilon,\ \epsilon \sim N(\mathbf{0}, \boldsymbol{I} + g\widetilde{\boldsymbol{X}}_u\boldsymbol{S}^{-1}\widetilde{\boldsymbol{X}}_u^\intercal N),$$

which is a particular case of the linear models studied in Berger et al. (1998). They show that a *minimal training sample* is formed by two observations impliying that $0 < m^U(\boldsymbol{y} \mid g) < \infty$ if $n_u = 2$. Furthermore, condition in eq.24 in Berger et al. (1998) is satisfied simply because $N(\mathbf{0}, \boldsymbol{I} + g\widetilde{\boldsymbol{X}}_u \boldsymbol{S}^{-1} \widetilde{\boldsymbol{X}}_u^{\mathsf{T}} N)$ is symmetric about the origin with the consequence that if $n_u = 2$ then $m^U(\boldsymbol{y} \mid g)$ does not depend on $\boldsymbol{I} + g\widetilde{\boldsymbol{X}}_u \boldsymbol{S}^{-1} \widetilde{\boldsymbol{X}}_u^{\mathsf{T}} N$ and particularly does not depend on g. Hence, and since $\pi(g)$ is a proper density if $n_u = 2$ then $0 < m(\boldsymbol{y}, \boldsymbol{\delta}) < \infty$ since $m(\boldsymbol{y}, \boldsymbol{\delta}) \leq m^U(\boldsymbol{y} \mid g)$.

For ii):

First we derive a uniform bound for the determinant of $\boldsymbol{S}$. Denote $\boldsymbol{S}_0 = \sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i^{\mathsf{T}}$. It can be easily seen that $\boldsymbol{S}_0 = \boldsymbol{S} + N\, \boldsymbol{x}_\omega^{\mathsf{T}}\, \boldsymbol{x}_\omega$, and hence, by well known properties of determinants

$$|\boldsymbol{S}_0| = |\boldsymbol{S}|(1 + N\, \boldsymbol{x}_\omega \boldsymbol{S}^{-1} \boldsymbol{x}_\omega^{\mathsf{T}}) \geq |\boldsymbol{S}|.$$

Now,

$$|\boldsymbol{S}| \leq |\boldsymbol{S}_0| \leq \prod_{j=1}^{k} \Big(\sum_{i=1}^{n} w_i x_{ij}^2 \Big) \leq (\max\{x_{ij}^2\})^k\, N^k$$

where the first inequality holds true because the determinant of a matrix is bounded by the product of the diagonal elements. The above observation makes it possible to bound the normal density in the prior in (6) in the manuscript as

$$\begin{aligned} N_k(\boldsymbol{\beta} \mid 0, g\, \boldsymbol{\Sigma}^M) &= N_k(\boldsymbol{\beta} \mid 0, g\, N\sigma^2 \boldsymbol{S}^{-1}) \leq (2\pi N\sigma^2)^{-k/2} |\boldsymbol{S}|^{1/2} \\ &\leq \sigma^{-k} (2\pi g)^{-k/2}\, (\max\{x_{ij}^2\})^{k/2}, \end{aligned} \tag{5}$$

and

$$\begin{aligned} \pi(\boldsymbol{\beta} \mid \beta_0, \sigma) &= \int N_k(\boldsymbol{\beta} \mid 0, g\, \boldsymbol{\Sigma}^M)\, \pi(g)\, dg \\ &\leq \sigma^{-k} (2\pi)^{-k/2} (\max\{x_{ij}^2\})^{k/2} \int g^{-k/2}\, \pi(g)\, dg, \end{aligned}$$

where the last integral is assumed to be finite.

The marginal distribution is then, except for constants, bounded as

$$\begin{aligned} m_1(\boldsymbol{y}, \boldsymbol{\delta}) &\leq \int N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \\ &\times Pr\big(N_{n_c}(\mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_c\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) > \boldsymbol{c}_c\big)\sigma^{-(k+1)}\, d\beta_0\, d\boldsymbol{\beta}\, d\sigma, \end{aligned}$$

and the term on the right above is bounded by above by

$$\begin{aligned} &\int N_{n_u}(\boldsymbol{y} \mid \beta_0\mathbf{1}_{n_u} + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I})\sigma^{-(k+1)}\, d\beta_0\, d\boldsymbol{\beta}\, d\sigma = \\ = &\int (\sqrt{2\pi}\sigma)^{-n_u} \exp\Big\{ -\begin{pmatrix} \beta_0 - \hat{\beta}_0 \\ \beta - \hat{\beta} \end{pmatrix}^{\mathsf{T}} \big((\mathbf{1}_{n_u}\, \widetilde{\boldsymbol{X}}_u)^{\mathsf{T}}(\mathbf{1}_{n_u}\, \widetilde{\boldsymbol{X}}_u)\big) \begin{pmatrix} \beta_0 - \hat{\beta}_0 \\ \beta - \hat{\beta} \end{pmatrix} /2\sigma^2 \Big\} \\ \times &\exp\{-SSE_u/2\sigma^2\}\sigma^{-(k+1)} d\beta_0\, d\boldsymbol{\beta}\, d\sigma \\ \propto &\int \sigma^{-n_u} \exp\{-SSE_u/2\sigma^2\} d\sigma, \end{aligned}$$

where SSE_u is the sum of squared errors

$$SSE_u = \boldsymbol{y}^{\mathsf{T}}\Big(\boldsymbol{I} - (\mathbf{1}_{n_u}\, \widetilde{\boldsymbol{X}}_u)\Big((\mathbf{1}_{n_u}\, \widetilde{\boldsymbol{X}}_u)^{\mathsf{T}}(\mathbf{1}_{n_u}\, \widetilde{\boldsymbol{X}}_u)\Big)^{-1}(\mathbf{1}_{n_u}\, \widetilde{\boldsymbol{X}}_u)^{\mathsf{T}}\Big)\boldsymbol{y},$$

which is strictly positive if $n_u \geq k + 2$, warranting that the last integral above is finite under the conditions in the theorem.

1.5 Proof of Theorem 5

For easiness in the proof, denote $\boldsymbol{\Sigma}^M(\boldsymbol{X})$ the matrix in (14) in the main text obtained from $\boldsymbol{X}$ and similar notation for $\boldsymbol{\Sigma}^M(\boldsymbol{Z})$. Also, following the convention in the paper, let $\widetilde{\boldsymbol{Z}}$ be the matrix $\boldsymbol{Z}$ with columns centered around their means and similarly for $\widetilde{\boldsymbol{Z}}_u$ and $\widetilde{\boldsymbol{Z}}_c$.

Now, the marginal under $\mathcal{M}_1$ with the transformed matrix is

$$\begin{aligned} &\int N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{Z}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I})\, Pr\big(N_{n_c}(\mathbf{1}\beta_0 + \widetilde{\boldsymbol{Z}}_c\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) > \boldsymbol{c}_c\big)\sigma^{-(k+1)} \\ \times &\sigma^{-1}\, N_k(\boldsymbol{\beta} \mid \mathbf{0}, g\, \boldsymbol{\Sigma}^M(\boldsymbol{Z}))\pi(g)\, d\beta_0\, d\boldsymbol{\beta}\, d\sigma dg. \end{aligned}$$

Now it suffices to apply above the identities $\boldsymbol{\Sigma}^M(\boldsymbol{Z}) = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}^M(\boldsymbol{X})\boldsymbol{D}^{-1}$, $\widetilde{\boldsymbol{Z}}_u = \widetilde{\boldsymbol{X}}_u\boldsymbol{D}$ and $\widetilde{\boldsymbol{Z}}_c = \widetilde{\boldsymbol{X}}_c\boldsymbol{D}$ and make the change of variables $\boldsymbol{\gamma} = \boldsymbol{D}\boldsymbol{\beta}$ to get the same marginal obtained from $\boldsymbol{X}$. The marginal under $\mathcal{M}_0$ does not change as it does not depend on the design matrix.

1.6 Proof of Lemma 1

To prove that $n_u \geq 2$ is sufficient, use the bound $\Phi^c \leq 1$, where $\Phi^c = 1 - \Phi$, that leads to:

$$m_0(\boldsymbol{y}, \boldsymbol{\delta}) \leq \int N_{n_u}(\boldsymbol{y} \mid \beta_0 \mathbf{1}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} \, d\sigma d\beta_0,$$

which is a finite integral if $n_u \geq 2$ as it is proved in Berger et al. (1998).

To prove that it is necessary, denote $c^\star = \max\{c_i \in \boldsymbol{c}_c\}$. Then

$$\begin{aligned} m_0(\boldsymbol{y}, \boldsymbol{\delta}) &\geq \int_{c^\star}^{\infty} \int_0^{\infty} \prod_{i \in \boldsymbol{C}_c} \Phi^c\big(\frac{c_i - \beta_0}{\sigma}\big) \times N_{n_u}(\boldsymbol{y} \mid \beta_0 \mathbf{1}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} \, d\sigma d\beta_0 \\ &\geq \int_{c^\star}^{\infty} \int_0^{\infty} 2^{-n_c} N_{n_u}(\boldsymbol{y} \mid \beta_0 \mathbf{1}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} \, d\sigma d\beta_0, \end{aligned}$$

where the last inequality holds true because $\beta_0 > c_i, \forall i$. Now, if $n_u = 0$ it is obvious that the integral above diverges. In the case where $n_u = 1$, and after integrating out σ we obtain an integral proportional to:

$$\int_{c^\star}^{\infty} \frac{1}{|y_u - \beta_0|} \, d\beta_0 = \infty.$$

1.7 Proof of Lemma 2

From Result 3 in Bayarri et al. (2012) we know that

$$\begin{aligned} &\int N_{k+1}(\boldsymbol{y} \mid \beta_0 \mathbf{1}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} d\beta_0 d\sigma \\ = &\int N_{k+1}(\boldsymbol{y} \mid \beta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} N_k(\boldsymbol{\beta} \mid 0, g\boldsymbol{\Sigma}) \pi(g) \, dg d\beta \, d\beta_0 d\sigma, \end{aligned} \tag{6}$$

if and only if $\boldsymbol{\Sigma} = n\sigma^2 \, (\widetilde{\boldsymbol{X}}^\intercal \widetilde{\boldsymbol{X}})^{-1}$ (or a multiple of this matrix).

In our problem, the marginal under the alternative model can be written as:

$$\begin{aligned} & m(\boldsymbol{y}, \boldsymbol{\delta}) \\ = & \int_{\boldsymbol{c}_c}^{\boldsymbol{\infty}} \int N_{k+1}(\begin{pmatrix} \boldsymbol{t} \\ \boldsymbol{y} \end{pmatrix} \mid \beta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} N_k(\boldsymbol{\beta} \mid 0, g\boldsymbol{\Sigma}) \pi(g)\, dg d\beta\, d\beta_0 d\sigma d\boldsymbol{t} \\ = & \text{ (If } \boldsymbol{\Sigma} = n\sigma^2 \, (\widetilde{\boldsymbol{X}}^{\mathsf{T}} \widetilde{\boldsymbol{X}})^{-1} \text{or a multiple)} \\ = & \int_{\boldsymbol{c}_c}^{\boldsymbol{\infty}} \int N_{k+1}(\begin{pmatrix} \boldsymbol{t} \\ \boldsymbol{y} \end{pmatrix} \mid \beta_0 \mathbf{1}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} d\beta_0 d\sigma d\boldsymbol{t} \\ = & \ m_0(\boldsymbol{y}, \boldsymbol{\delta}) \end{aligned}$$

As a consequence of Result 3 in Bayarri et al. (2012) this equality only holds true for variance matrices of the type $\boldsymbol{\Sigma}^A$ or a multiple.

1.8 Proof of Lemma 3

First express:

$$\begin{aligned} m(\boldsymbol{y}, \boldsymbol{\delta}) \ &= \ \int \prod_{i \in \boldsymbol{C}_c} \Phi^c\big(\frac{c_i - \beta_0}{\sigma}\big) \times N_{k+1}(\boldsymbol{y} \mid \beta_0 \mathbf{1} + \boldsymbol{X}_u \boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \\ &\times \ \frac{1}{\sigma} N_k(\boldsymbol{\beta} \mid 0, g\boldsymbol{\Sigma}) \pi(g)\, dg d\beta\, d\beta_0 d\sigma, \end{aligned}$$

to note that the integral is uniformly bounded, allowing interchanging limits and integrals. Now, clearly

$$\lim_{c_\star \to -\infty} \prod_{i \in \boldsymbol{C}_c} \Phi^c\big(\frac{c_i - \beta_0}{\sigma}\big) = 1,$$

so

$$\lim_{c_\star \to -\infty} m(\boldsymbol{y}, \boldsymbol{\delta}) = \int N_{k+1}(\boldsymbol{y} \mid \beta_0 \mathbf{1} + \boldsymbol{X}_u \boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} N_k(\boldsymbol{\beta} \mid 0, g\boldsymbol{\Sigma}) \pi(g)\, dg d\beta\, d\beta_0 d\sigma. \quad (7)$$

Similarly

$$\lim_{c_\star \to -\infty} m_0(\boldsymbol{y}, \boldsymbol{\delta}) = \int N_{k+1}(\boldsymbol{y} \mid \beta_0 \mathbf{1}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} d\beta_0 d\sigma. \quad (8)$$

Now, as a consequence of (6), both right hand sides of (7) and (8) coincide if and only if $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^U$ (or a multiple).

2 An illustrative application and model averaged estimations

For illustrative purposes we analyze the heart transplant survival dataset, considered previously by several authors (see Crowley and Hu, 1977, and references therein). It contains data from the Stanford Heart Transplantation Program. Data is available in the library `survival` from `R` software (Therneau, 2015). The original data contain information about each patient related to the day he/she agreed to enter the program and the last day he/she has been seen, so the survival time is calculated as the difference between these two times. We study the group of patients that have received a transplant. Each survival time is censored (uncensored) depending on whether the last time seen is the date of death or the closing date of the study: April 1, 1974. To confine the study to the comparison of only two models, we restrict our interest to investigating whether survival time depends on the patient age (for a more in-depth study on this dataset see Crowley and Hu, 1977; Brown et al., 1973).

For this dataset, $n = 69$, 24 of which survived to the end of study, so $n_u = 45$ and there is approximately 35% censoring. Following previous notation the model with age is called $\mathcal{M}_1$ while the model with just the intercept is $\mathcal{M}_0$.

The Bayes factor we obtain with our proposal slightly favors the null model leading to $B_1 = 0.63$ ($B_1 = 0.69$ for $\pi(g)$ the inverse gamma, equation (9) in the manuscript and $B_1 = 0.89$ for fixed $g = 1$). Assuming equal prior probabilities, these results would translate in $p(\mathcal{M}_0 \mid \boldsymbol{y}, \boldsymbol{\delta}) = 0.61$ (0.59 and 0.53, respectively, for the other choices of the mixing function). These results agree with the conclusions obtained in Crowley and Hu (1977).

An interesting by-product of our methodology is the effective sample size defined in equation number (15) in the manuscript. As we have argued, and broadly speaking, this unknown quantity measures the value of the missing observations ($N_{(\beta_0,\sigma)}$ close to n_u would mean barely any extra information, while at the opposite extreme $N_{(\beta_0,\sigma)} \approx n$ would mean highly informative censored observations). Of course, the estimation of $N_{(\beta_0,\sigma)}$ varies on the model used, but given that β_0 and σ have a similar

meaning across models, it seems valid and natural to report a model averaging estimation of $N_{(\beta_0,\sigma)}$. Model averaged (MA) estimations are obtained by summarizing the posterior distribution of $N_{(\beta_0,\sigma)}$:

$$\pi(N_{(\beta_0,\sigma)} \mid \boldsymbol{y}, \boldsymbol{\delta}) = p(\mathcal{M}_1 \mid \boldsymbol{y}, \boldsymbol{\delta})\pi_1(N_{(\beta_0,\sigma)} \mid \boldsymbol{y}, \boldsymbol{\delta}) + p(\mathcal{M}_0 \mid \boldsymbol{y}, \boldsymbol{\delta})\pi_0(N_{(\beta_0,\sigma)} \mid \boldsymbol{y}, \boldsymbol{\delta}), \tag{9}$$

from which drawing simulations, from those for the parameters under each model is straightforward. In this example, the posterior distribution of $N_{(\beta_0,\sigma)}$ is represented in Figure 1 under M_0 (histogram in the upper left corner); under M_1 (upper right) and the Model Averaged distribution (lower left). Under the Model Averaged distribution, a credible interval for $N_{(\beta_0,\sigma)}$ at 90% probability is [59, 64] with the posterior median at 61.9. Recall that in this case $n_u = 45$ and $n = 69$, so the censored observations contain relevant information equivalent to around 17 uncensored points.

We can also obtain MA estimations of the coefficient β associated to the effect of age. In this case, the corresponding posterior distribution (similar to that in eq 9) is a mixture of a degenerate distribution at zero (coming from $\mathcal{M}_0$) and a continuous distribution (coming from $\mathcal{M}_1$). In Figure 1 (lower right corner) we have provided a possible graphical representation of this distribution, in which the dark gray bar represents the probability under the null while the light gray area is the distribution of the probability under $\mathcal{M}_1$ (both areas sum one). From this distribution we conclude that, in the case of an effect of age, it is of a negative sign of about -0.06. Furthermore, the probability that this parameter is *strictly* negative is approximately 0.38.

3 Details about computation of BFs

In this section we provide details about the computation of BFs using the important sampling algorithm specified in Section 4.5 in the manuscript. These are the practical details for steps 1, 2 and 4 in the algorithm:

1. In this first step simulations of the posterior distribution are obtained, for each

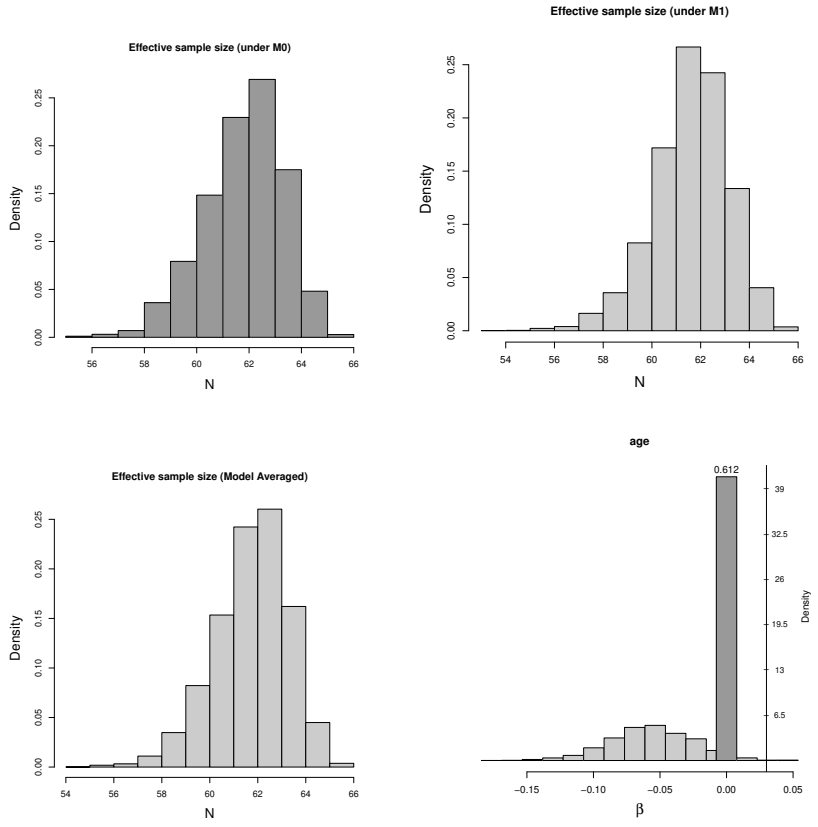

Figure 1: Heart transplant example. First row: posterior distributions of the effective sample size $N_{(\beta_0,\sigma)}$ under M_0 (left) and under M_1 (right). Second row: Model averaged posterior distribution of $N_{(\beta_0,\sigma)}$ (left) and of the regression coefficient for the covariate age (right) –dark gray area represents the probability under the null model (no effect) and the light gray area the distribution of probability under the alternative (effect)–.

type of prior for g, the inverse gamma prior or the robust prior given in equations (9) and (10) in the manuscript, or fixing $g = 1$, the target distribution is:

$$\pi(\boldsymbol{\beta}, \beta_0, \log(\sigma), g \mid \boldsymbol{y}, \boldsymbol{\delta}) \propto f_1(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \log(\sigma), \boldsymbol{\beta}) \int N_k(\boldsymbol{\beta} \mid 0, g\, \boldsymbol{\Sigma}^M)\, \pi(g) dg \tag{10}$$

where f_1 is given in equation (3) in the manuscript.

Denoting by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \beta_0, \log(\sigma))$, the proposal distribution at step (t) is the following product:

$$q^{MH}(\boldsymbol{\theta}, g) = N_k(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t-1)}, \boldsymbol{\Sigma}^q)\pi(g),$$

where $\boldsymbol{\Sigma}^q$ is the inverse of the hessian calculated at the mode of $\boldsymbol{\theta}$ previously calculated and multiplied by a fixed scale factor, chosen in order to have an acceptance rate of 30-40%, and $\pi(g)$ is the inverse gamma prior or the robust prior given in equations (9) and (10) in the manuscript, as it is mention before. When $g = 1$, the target distribution is $N_k(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t-1)}, \boldsymbol{\Sigma}^q)$.

2. It is constructed a proposal distribution to do important sampling in order to approximate the predictive distribution, the proposal distribution we use is:

$$q^{IS}(\boldsymbol{\theta}, g) = t_k(\boldsymbol{\theta} \mid \nu = 3, \Delta^{post}, \Sigma^{post})\pi(g),$$

where t_k denotes a non-central k-variate distribution with 3 degrees of freedom, Δ^{post} and Σ^{post} are the mean and covariance matrix calculated over the posterior sample obtained in step 1. $\pi(g)$ is the inverse gamma prior or the robust prior.

4. Using a sample of size N, in the application $N = 10000$ simulations, $((\boldsymbol{\theta}_1, g_1), \ldots, (\boldsymbol{\theta}_N, g_N))$ obtained from $q^{IS}(\boldsymbol{\theta}, g)$ the approximated marginal predictive distribution for model with likelihood given by f_1, equation (3) in the manuscript, is:

$$\widehat{m_1(\boldsymbol{y},\boldsymbol{\delta})} = \frac{1}{N}\sum_{i=1}^{N}\frac{\pi(\boldsymbol{\theta}_i, g_i \mid \boldsymbol{y},\boldsymbol{\delta})}{q^{IS}(\boldsymbol{\theta}_i, g_i)}.$$

where $\pi(\boldsymbol{\theta}_i, g_i \mid \boldsymbol{y},\boldsymbol{\delta})$ is given in equation (10). This approximation is used in the numerator of the BF.

4 Simulation study for comparing different versions of BF

Here we report the results of the simulated experiment obtained from the heart transplant data set. We simulated 50 data sets to which we performed variable selection based on BF_{robust}, TBF_{EB} and TBF_{ZS}. Inclusion probabilities in the form of cloud of points are represented in Figure 2 and further summarized in Figure 3.

References

Bayarri, M., Berger, J., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian Model Choice with Application to Variable Selection." *The Annals of Statistics*, 40: 1550–1577.

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998). "Bayes Factors and Marginal Distributions in Invariant Situations." *Sankhya: The Indian Journal of Statistics, Series A*, 60(3): 307–321.

Brown, B. W., Jr., Hollander, M., and Korwar, R. M. (1973). "Nonparametric tests of independence for censored data with application to heart transplant studies." Technical report, Florida State Univ Tallahassee Dept of Statistics.

Crowley, J. and Hu, M. (1977). "Covariance analysis of heart transplant survival data." *Journal of the American Statistical Association*, 72(357): 27–36.

De Santis, F., Mortera, J., and Nardi, A. (2001). "Jeffreys priors for survival models with censored data." *Journal of statistical planning and inference*, 99(2): 193–209.

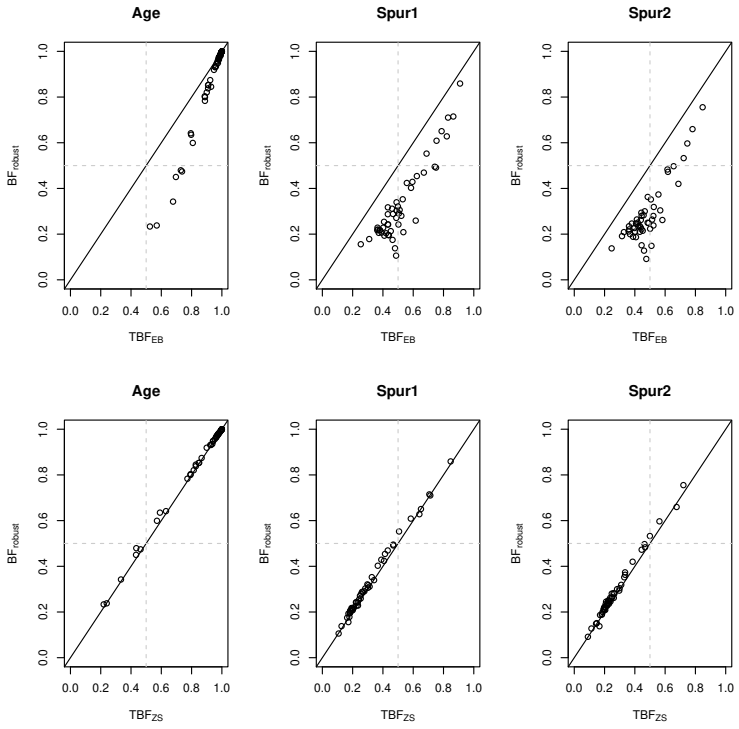

Figure 2: Inclusion probabilities for the 50 simulated data sets related, based on BF_{robust}, TBF_{EB}, and TBF_{ZS}. Variable Age is a true explanatory variable while Spur1 and Spur2 do not affect the response.

Sampford, M. R. (1953). "Some inequalities on Mill's ratio and related functions." *The Annals of Mathematical Statistics*, 24(1): 130–132.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. Version 2.38. URL `http://CRAN.R-project.org/package=survival`

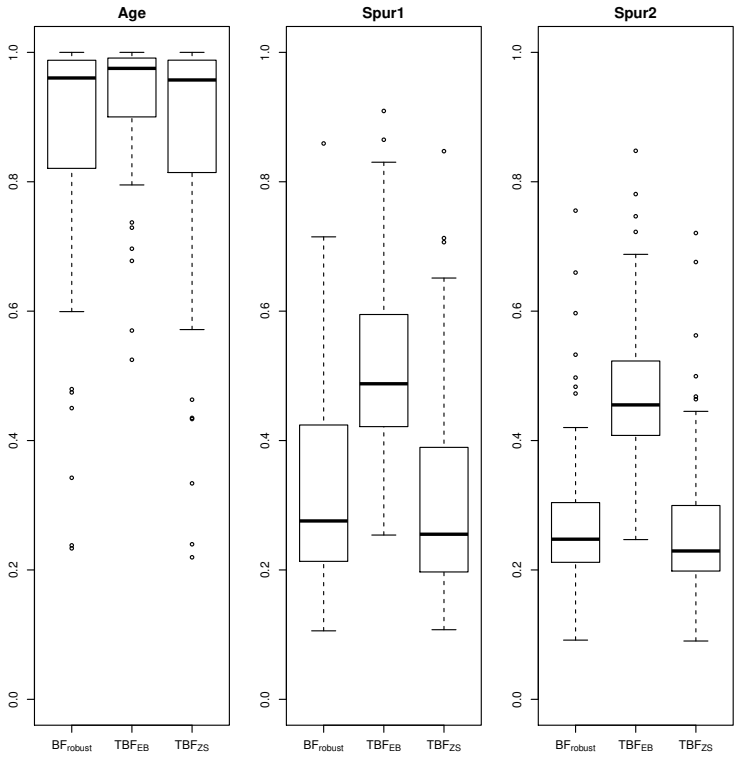

Figure 3: Inclusion probabilities for the 50 simulated data sets related, based on BF_{robust}, TBF_{EB}, and TBF_{ZS}. Variable Age is a true explanatory variable while Spur1 and Spur2 do not affect the response.