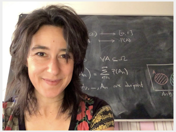
Model uncertainty quantification in presence of missing data

O'Bayes 2022

María Eugenia Castellanos[1], Stefano Cabras[2], Anabel Forte [3], Gonzalo García-Donato [4] & Alicia Quirós[5]

[1] Universidad Rey Juan Carlos
[2] Universidad Carlos III de Madrid
[3] Universidad de Valencia
[4] Universidad Castilla-La Mancha
[5] Universidad de León



Santa Cruz, September, 2022

- Consider a random sample $\boldsymbol{z}_i = (y_i, x_{1,i}, \ldots, x_{p,i})$, $i = 1, \ldots n$.

Model selection in presence of missing data

- Consider a random sample $\boldsymbol{z}_i = (y_i, x_{1,i}, \dots, x_{p,i})$, $i = 1, \dots n$.
- Interested in: $\boldsymbol{y} \mid \boldsymbol{x}_1, \dots, \boldsymbol{x}_p$,

- Consider a random sample $\boldsymbol{z}_i = (y_i, x_{1,i}, \ldots, x_{p,i}), i = 1, \ldots n$.
- Interested in: $\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$,
- Several competing regression models (indexed by $\boldsymbol{\gamma} \in \Gamma$),

$$\mathcal{M}_\gamma : \boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \sim f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma).$$

Model selection in presence of missing data

- Consider a random sample $\boldsymbol{z}_i = (y_i, x_{1,i}, \ldots, x_{p,i})$, $i = 1, \ldots n$.
- Interested in: $\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$,
- Several competing regression models (indexed by $\boldsymbol{\gamma} \in \Gamma$),

$$\mathcal{M}_\gamma : \boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \sim f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma).$$

- Only $\boldsymbol{z}_i$ with $i \in \boldsymbol{i}^{obs} = \{i_1, \ldots, i_m\}$ are fully observed (first m components) due to reasons independent of f_γ (Missing At Random scenario)

Model selection in presence of missing data

- Consider a random sample $\boldsymbol{z}_i = (y_i, x_{1,i}, \ldots, x_{p,i})$, $i = 1, \ldots n$.
- Interested in: $\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$,
- Several competing regression models (indexed by $\boldsymbol{\gamma} \in \Gamma$),

$$\mathcal{M}_\gamma : \boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \sim f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma).$$

- Only $\boldsymbol{z}_i$ with $i \in \boldsymbol{i}^{obs} = \{i_1, \ldots, i_m\}$ are fully observed (first m components) due to reasons independent of f_γ (Missing At Random scenario)
- D^{obs} denotes all observed components.

- Consider a random sample $\boldsymbol{z}_i = (y_i, x_{1,i}, \ldots, x_{p,i})$, $i = 1, \ldots n$.
- Interested in: $\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$,
- Several competing regression models (indexed by $\boldsymbol{\gamma} \in \Gamma$),

$$\mathcal{M}_\gamma : \boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \sim f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma).$$

- Only $\boldsymbol{z}_i$ with $i \in \boldsymbol{i}^{obs} = \{i_1, \ldots, i_m\}$ are fully observed (first m components) due to reasons independent of f_γ (Missing At Random scenario)
- D^{obs} denotes all observed components.

In this context...

- How to make model selection?
- How to measure model uncertainty: different models affected by a different set of missing data?

How practitioners do?

- From a non-Bayesian point of view: case-deletion or complete-case strategies applying step-wise selection with AIC or BIC, or other tests;

How practitioners do?

- From a non-Bayesian point of view: case-deletion or complete-case strategies applying step-wise selection with AIC or BIC, or other tests;
- Bad approaches: they sacrifice useful information producing bias results, except perhaps for particular cases as Missing Completely At Random (MCAR).

How practitioners do?

- From a non-Bayesian point of view: case-deletion or complete-case strategies applying step-wise selection with AIC or BIC, or other tests;
- Bad approaches: they sacrifice useful information producing bias results, except perhaps for particular cases as Missing Completely At Random (MCAR).
- Within the multiple imputation (MI) approach, proposed by Rubin (1987), what we call MI world, traditional non-Bayesian variable selection tools are difficult to be applied.

- MI world $\rightsquigarrow$ apply statistical techniques to MI data-sets, there are two proposals in literature:

[*] Imputation and Variable Selection in Linear Regression Models with Missing Covariates, 2005, *Biometrics*.
[†] Computing Bayes factors from data with missing values, 2019, *Psychol Methods*

- MI world $\rightsquigarrow$ apply statistical techniques to MI data-sets, there are two proposals in literature:
- The *Biometrics* paper by Yang, Belin and Moscardin (2005)*, a first attempt to make imputation and Bayesian variable selection in linear regression models.

*Imputation and Variable Selection in Linear Regression Models with Missing Covariates, 2005, *Biometrics*.
†Computing Bayes factors from data with missing values, 2019, *Psychol Methods*

- MI world $\rightsquigarrow$ apply statistical techniques to MI data-sets, there are two proposals in literature:
- The *Biometrics* paper by Yang, Belin and Moscardin (2005)[*], a first attempt to make imputation and Bayesian variable selection in linear regression models.
  - Intuitively, they propose a way to obtain the posterior probability for each possible model in each imputed data set, and averaging posterior probabilities for each model over the MI data-sets.

[*] Imputation and Variable Selection in Linear Regression Models with Missing Covariates, 2005, *Biometrics*.
[†] Computing Bayes factors from data with missing values, 2019, *Psychol Methods*

Model selection with missing data: Bayesian perspective

- MI world ⇝ apply statistical techniques to MI data-sets, there are two proposals in literature:
- The *Biometrics* paper by Yang, Belin and Moscardin (2005)*, a first attempt to make imputation and Bayesian variable selection in linear regression models.
  - Intuitively, they propose a way to obtain the posterior probability for each possible model in each imputed data set, and averaging posterior probabilities for each model over the MI data-sets.
- A recent paper by Hoijtink, Gu, Mulder and Rosseel, 2019, discusses how to compute Bayes Factors (BFs) doing MI for hypothesis testing in Psychology. †

*Imputation and Variable Selection in Linear Regression Models with Missing Covariates, 2005, *Biometrics*.
†Computing Bayes factors from data with missing values, 2019, *Psychol Methods*

- MI world $\rightsquigarrow$ apply statistical techniques to MI data-sets, there are two proposals in literature:
- The *Biometrics* paper by Yang, Belin and Moscardin (2005)*, a first attempt to make imputation and Bayesian variable selection in linear regression models.
  - Intuitively, they propose a way to obtain the posterior probability for each possible model in each imputed data set, and averaging posterior probabilities for each model over the MI data-sets.
- A recent paper by Hoijtink, Gu, Mulder and Rosseel, 2019, discusses how to compute Bayes Factors (BFs) doing MI for hypothesis testing in Psychology. †
  - To average BFs approximating marginal distributions from Gibbs output over the imputed data sets.

*Imputation and Variable Selection in Linear Regression Models with Missing Covariates, 2005, *Biometrics*.
†Computing Bayes factors from data with missing values, 2019, *Psychol Methods*

- Revisit all concepts that appear in model selection/model uncertainty obtaining their formulation in presence of missing data;

- Revisit all concepts that appear in model selection/model uncertainty obtaining their formulation in presence of missing data;
- The basic ingredient for model uncertainty quantification (MUQ) is the predictive density in the observed data,

$$m(D^{obs}) = \int f(D^{obs} \mid \boldsymbol{\theta})\, d\Pi(\boldsymbol{\theta}) = \int f(D^{obs}, D^{na} \mid \boldsymbol{\theta})\, dD^{na}\, \pi(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where D^{na} denotes the missing components in $\boldsymbol{z}_i$, with $i \in m+1, m+2, \ldots, n$. ($D^{obs}$ denotes all the components of $\boldsymbol{z}$ observed).

1 Motivation about model selection with missing data

2 Regression models. Full observed data

3 Regression models. Missing data

4 Computing marginals with missing data
- Prior distributions
- Bayes Factors
- Posterior probabilities

5 Simulated example

6 Comments and remarks

- The entertained regression models may differ in a number of ways:
    1. Set of covariates needed to explain Y or,
    2. Density assumed for the errors in linear models.

- The entertained regression models may differ in a number of ways:
  1. Set of covariates needed to explain Y or,
  2. Density assumed for the errors in linear models.
- **Notation**:
  - $\boldsymbol{\theta}_0$ parameters appearing in all competing models;
  - $\boldsymbol{\theta}_\gamma$ specific parameter in $\mathcal{M}_\gamma$;
  - A more precise labelling for the parameters of M_γ is $\big((\boldsymbol{\theta}_0)_\gamma, \boldsymbol{\theta}_\gamma\big)$, we abuse notation considering $\big(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma\big)$.

Full observed data: Designed experiments

- In designed experiments, the marginal distribution is obtained "conditional" on the values of covariates:

$$m_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = \int f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)\, \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_\gamma$$

- Usual prior distributions over $\boldsymbol{\theta}_\gamma$: Zellner–Siow priors, robust priors, hyper-g-priors, etc, use a conditional variance that depends on $\boldsymbol{x}_i$, $i = 1, \ldots, p$.
- This is legitimate as the $\boldsymbol{x}_i$ are fixed covariates designed for the experiment.

- Covariates are random: compare models in the basis of how they predict all observed values.
- Introducing the idea of competing models as joint statistical models:

$$\mathcal{M}_\gamma : \boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \sim f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma) f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \mid \boldsymbol{\nu}).$$

- The marginal density:

$$m_\gamma(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = \int f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)\, f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \mid \boldsymbol{\nu}) \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}) d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_\gamma\, d\boldsymbol{\nu}.$$

- Assuming prior independence: $\pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}) = \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)\pi(\boldsymbol{\nu})$, then

$$m_\gamma(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = \underbrace{\int f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \mid \boldsymbol{\nu})\pi(\boldsymbol{\nu})d\boldsymbol{\nu}}_{m(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p)} \times \int f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)\pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_\gamma.$$

- First factor is independent of the model and would cancel in the BF.
- The distribution $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \mid \boldsymbol{\nu})$ is negligible $\rightarrow$ identical results that in the fixed covariates case.
- Justification of the no discussion about the fixed or random covariates: it does not affect results...

- Assuming prior independence: $\pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}) = \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)\pi(\boldsymbol{\nu})$, then

$$m_\gamma(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = \underbrace{\int f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \mid \boldsymbol{\nu})\pi(\boldsymbol{\nu})d\boldsymbol{\nu}}_{m(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p)} \times \int f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)\pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_\gamma.$$

- First factor is independent of the model and would cancel in the BF.
- The distribution $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \mid \boldsymbol{\nu})$ is negligible $\rightarrow$ identical results that in the fixed covariates case.
- Justification of the no discussion about the fixed or random covariates: it does not affect results...

First message

But... $\pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)$ cannot depend on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, invalidating the most popular priors: g-prior, Zellner-Ziow prior, hyper-g prior, robust prior, etc.

Missing data. Notation

- Missing values only in covariates. Not missing values in $\boldsymbol{y}$.

- Missing values only in covariates. Not missing values in $\boldsymbol{y}$.
- Tilde symbol used to represent a realization of a random or vector variable. $\tilde{\boldsymbol{y}}$ is a realization of the random vector $\boldsymbol{y}$.

- Missing values only in covariates. Not missing values in $\boldsymbol{y}$.
- Tilde symbol used to represent a realization of a random or vector variable. $\tilde{\boldsymbol{y}}$ is a realization of the random vector $\boldsymbol{y}$.
- Consider a $n \times p$ binary matrix $M = (m_{ij})$,

$$\text{with } m_{ij} = 1 \text{ when } x_{ij} \text{ is missing.}$$

- Missing values only in covariates. Not missing values in $\boldsymbol{y}$.
- Tilde symbol used to represent a realization of a random or vector variable. $\tilde{\boldsymbol{y}}$ is a realization of the random vector $\boldsymbol{y}$.
- Consider a $n \times p$ binary matrix $M = (m_{ij})$,

$$\text{with } m_{ij} = 1 \text{ when } x_{ij} \text{ is missing.}$$

- M is a random matrix entering into the competing models, so each $\mathcal{M}_\gamma$ is:

$$\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, M \sim f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma) f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \mid \boldsymbol{\nu})\, f(M \mid \boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \psi).$$

- Missing values only in covariates. Not missing values in $\boldsymbol{y}$.
- Tilde symbol used to represent a realization of a random or vector variable. $\tilde{\boldsymbol{y}}$ is a realization of the random vector $\boldsymbol{y}$.
- Consider a $n \times p$ binary matrix $M = (m_{ij})$,

$$\text{with } m_{ij} = 1 \text{ when } x_{ij} \text{ is missing.}$$

- M is a random matrix entering into the competing models, so each $\mathcal{M}_\gamma$ is:

$$\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, M \sim f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma) f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \mid \boldsymbol{\nu}) f(M \mid \boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\psi}).$$

- Missingness $\rightarrow$ only certain values of the covariates are observed: $\widetilde{\boldsymbol{x}}_{(0)}$, i.e. $\widetilde{\boldsymbol{x}}_{(0)} = \{\widetilde{x}_{ij} : \widetilde{m}_{ij} = 0\}$.

- Missing values only in covariates. Not missing values in $\boldsymbol{y}$.
- Tilde symbol used to represent a realization of a random or vector variable. $\tilde{\boldsymbol{y}}$ is a realization of the random vector $\boldsymbol{y}$.
- Consider a $n \times p$ binary matrix $M = (m_{ij})$,

$$\text{with } m_{ij} = 1 \text{ when } x_{ij} \text{ is missing.}$$

- M is a random matrix entering into the competing models, so each $\mathcal{M}_\gamma$ is:

$$\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, M \sim f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma) f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \mid \boldsymbol{\nu})\, f(M \mid \boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \psi).$$

- Missingness $\rightarrow$ only certain values of the covariates are observed: $\widetilde{\boldsymbol{x}}_{(0)}$, i.e. $\widetilde{\boldsymbol{x}}_{(0)} = \{\widetilde{x}_{ij} : \widetilde{m}_{ij} = 0\}$.
- Rest of components of covariates are random, denoted as $\boldsymbol{x}_{(1)}$.

- Different mechanisms assumed to represent the missingness structure (described in Little and Rubin, 2020).
- Consider Missing at Random, (MAR) mechanism, the weakest condition to avoid specifying the probability distribution of M;
- Missing data are MAR for an observed data $(\widetilde{M}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})$ if

$$f(\widetilde{M} \mid \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\psi}) = f(\widetilde{M} \mid \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}^{\star}_{(1)}, \boldsymbol{\psi}), \text{ for any } \boldsymbol{x}_{(1)} \neq \boldsymbol{x}^{\star}_{(1)}$$

Abbreviated: $f(\widetilde{M} \mid \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\psi})$ does not depend on $\boldsymbol{x}_{(1)}$.

- Join prior predictive marginal:

$$
\begin{aligned}
m_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \widetilde{M}) &= \int m_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \widetilde{M}) d\boldsymbol{x}_{(1)} \\
= & \int f_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \widetilde{M} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}, \boldsymbol{\psi})\, \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}, \boldsymbol{\psi}) d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_\gamma\, d\boldsymbol{\nu} d\boldsymbol{\psi} d\boldsymbol{x}_{(1)} \\
= & \int \big[f_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma) f(\widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)} \mid \boldsymbol{\nu}) f(\widetilde{M} \mid \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\psi}) \\
\times & \ \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}, \boldsymbol{\psi}) d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_\gamma\, d\boldsymbol{\nu} d\boldsymbol{\psi} d\boldsymbol{x}_{(1)} \big]
\end{aligned}
$$

- Using the MAR assumption and considering independence between parameters governing the missing mechanism and the rest: $\pi_\gamma(\boldsymbol{\psi} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}) = \pi(\boldsymbol{\psi})$ (we call these MU-ignorable condition).

Prior predictive marginal under MAR

- $$m_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \widetilde{M}) = m_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}) \int f(\widetilde{M} \mid \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{\psi}) \pi(\boldsymbol{\psi}) d\boldsymbol{\psi}$$

  where

  $$m_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}) = \int f_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)\, f(\widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)} \mid \boldsymbol{\nu}) \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}) d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_\gamma\, d\boldsymbol{\nu} d\boldsymbol{x}_{(1)}.$$

- In the above equation (in pink), the second factor does not depend on $\gamma \to$ cancel out in the BFs.

- Under the MU-ignorable condition, and considering conditional prior independence: $\pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}) = \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu})\pi_\gamma(\boldsymbol{\nu})$.
- The marginal prior of interest is:

$$\begin{aligned}
m_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}) &= \int f_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu})\, \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu})\, d\boldsymbol{x}_{(1)}\, d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}) \\
&= \int \Big[\underbrace{\int f_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)\, \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu}) d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma)}_{\mathfrak{m}_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})} \\
&\times \; f(\widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)} \mid \boldsymbol{\nu})\pi_\gamma(\boldsymbol{\nu})\, d\boldsymbol{x}_{(1)} d\boldsymbol{\nu} \Big] \\
&= \int \mathfrak{m}_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})\, f(\widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)} \mid \boldsymbol{\nu})\pi_\gamma(\boldsymbol{\nu})\, d\boldsymbol{x}_{(1)}\, d\boldsymbol{\nu} \\
&= \int \mathfrak{m}_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})\, f(\boldsymbol{x}_{(1)} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{\nu}) f(\widetilde{\boldsymbol{x}}_{(0)} \mid \boldsymbol{\nu})\pi_\gamma(\boldsymbol{\nu})\, d\boldsymbol{x}_{(1)}\, d\boldsymbol{\nu} \\
&= m_\gamma(\widetilde{\boldsymbol{x}}_{(0)}) \int \mathfrak{m}_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})\, f(\boldsymbol{x}_{(1)} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{\nu})\pi_\gamma(\boldsymbol{\nu} \mid \widetilde{\boldsymbol{x}}_{(0)})\, d\boldsymbol{x}_{(1)}\, d\boldsymbol{\nu}
\end{aligned}$$

Marginal approximation via simulation

If $\mathfrak{m}_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})$ can be easily evaluated, then $m_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})$ can be approximated by simulation:

```
For j = 1, ..., N:
  1 :  Draw ν^(j) ~ π_γ(ν | x̃_(0)),
  2 :  draw (x_(1))^(j) ~ f(x_(1) | x̃_(0), ν^(j)),
  3 :  compute m^(j) = m_γ(ỹ | x̃_(0), (x_(1))^(j), ν^(j)),
```

$\texttt{For}\ j = 1, \ldots, N:$

1 : $\texttt{Draw}\ \boldsymbol{\nu}^{(j)} \sim \pi_\gamma(\boldsymbol{\nu} \mid \widetilde{\boldsymbol{x}}_{(0)})$,

2 : $\texttt{draw}\ (\boldsymbol{x}_{(1)})^{(j)} \sim f(\boldsymbol{x}_{(1)} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{\nu}^{(j)})$,

3 : $\texttt{compute}\ \mathfrak{m}^{(j)} = \mathfrak{m}_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, (\boldsymbol{x}_{(1)})^{(j)}, \boldsymbol{\nu}^{(j)})$,

- Approximate

$$m_\gamma \widehat{(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})} = N^{-1} \sum \mathfrak{m}^{(j)}$$

- Steps 1 and 2 are doing with an augmented Gibbs scheme.

Example (Simple linear regression)

- Our data contains observations from three variables $(y, x_1, x_2) \in \mathbb{R}^3$.
- Two competing regression models, explaining Y:

$$\begin{aligned} H_0 &: \quad f_0(y \mid x_1, x_2, \beta_0, \sigma) = N(y \mid \beta_0, \sigma^2), \\ H_1 &: \quad f_1(y \mid x_1, x_2, \beta_0, \beta, \sigma) = N(y \mid \beta_0 + \beta_1 x_1, \sigma^2). \end{aligned}$$

- $\Gamma = \{0, 1\}$, variable x_2 is is not relevant for this model uncertainty problem, but it will be for making imputation.
- Here $\boldsymbol{\theta}_0 = (\beta_0, \sigma)$ are common parameters for both regression models, while $\boldsymbol{\theta}_\gamma = \beta_1$ is specific to $\mathcal{M}_1$.

Prior distributions considered

- Objective or non-informative setting.
- Adaptation of well-known practices in the model uncertainty literature to the missing data problems.
- Prior for $\mathcal{M}_\gamma$:

$$\pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}) = \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu})\pi_\gamma(\boldsymbol{\nu}) = \pi_\gamma(\boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu}, \boldsymbol{\theta}_0)\pi_\gamma(\boldsymbol{\theta}_0 \mid \boldsymbol{\nu})\pi_\gamma(\boldsymbol{\nu}).$$

- Objective or non-informative setting.
- Adaptation of well-known practices in the model uncertainty literature to the missing data problems.
- Prior for $\mathcal{M}_\gamma$:

$$\pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma, \boldsymbol{\nu}) = \pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu})\pi_\gamma(\boldsymbol{\nu}) = \pi_\gamma(\boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu}, \boldsymbol{\theta}_0)\pi_\gamma(\boldsymbol{\theta}_0 \mid \boldsymbol{\nu})\pi_\gamma(\boldsymbol{\nu}).$$

Definition (Prior scheme recommended)

$$\pi_\gamma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu}) = \pi(\boldsymbol{\theta}_0)\pi_\gamma(\boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu}, \boldsymbol{\theta}_0),\ \pi_\gamma(\boldsymbol{\nu}) = \pi(\boldsymbol{\nu}), \tag{1}$$

where $\pi_\gamma(\boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu}, \boldsymbol{\theta}_0)$ *is proper and depends on* $\boldsymbol{\nu}$ *and* $\pi(\boldsymbol{\nu})$ *and/or* $\pi(\boldsymbol{\theta}_0)$ *are potentially improper.*

For models that only have common parameters, (1) *should be understood as:*

$$\pi_\gamma(\boldsymbol{\theta}_0 \mid \boldsymbol{\nu}) = \pi(\boldsymbol{\theta}_0),\ \pi_\gamma(\boldsymbol{\nu}) = \pi(\boldsymbol{\nu}).$$

- For the priors defined above, $m_\gamma(\widetilde{\boldsymbol{x}}_{(0)})$ is independent of γ and

$$m_\gamma(\boldsymbol{y}, \widetilde{\boldsymbol{x}}_{(0)}) = m(\widetilde{\boldsymbol{x}}_{(0)}) \int \mathfrak{m}_\gamma(\boldsymbol{y} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})\, f(\boldsymbol{x}_{(1)} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{\nu}) \pi(\boldsymbol{\nu} \mid \widetilde{\boldsymbol{x}}_{(0)})\, d\boldsymbol{x}_{(1)}\, d\boldsymbol{\nu},$$

$\rightarrow m(\widetilde{\boldsymbol{x}}_{(0)})$ cancels in the BFs. as it appears in all models.

- For the priors defined above, $m_\gamma(\widetilde{\boldsymbol{x}}_{(0)})$ is independent of γ and

$$m_\gamma(\boldsymbol{y}, \widetilde{\boldsymbol{x}}_{(0)}) = m(\widetilde{\boldsymbol{x}}_{(0)}) \int \mathfrak{m}_\gamma(\boldsymbol{y} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})\, f(\boldsymbol{x}_{(1)} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{\nu}) \pi(\boldsymbol{\nu} \mid \widetilde{\boldsymbol{x}}_{(0)})\, d\boldsymbol{x}_{(1)}\, d\boldsymbol{\nu},$$

$\rightarrow m(\widetilde{\boldsymbol{x}}_{(0)})$ cancels in the BFs. as it appears in all models.

About $\pi_\gamma(\boldsymbol{\nu})$

- Contributes to $m_\gamma(\boldsymbol{y}, \widetilde{\boldsymbol{x}}_{(0)})$ through $\pi_\gamma(\boldsymbol{\nu} \mid \widetilde{\boldsymbol{x}}_{(0)})$, under weak conditions this is a proper distribution.
- It can be used an improper prior.
- The meaning of $\boldsymbol{\nu}$ does not change with M_γ, $f(\boldsymbol{x} \mid \boldsymbol{\nu})$ is independent of $\boldsymbol{\gamma} \rightarrow$ same $\pi_\gamma(\boldsymbol{\nu})$ for every model $\mathcal{M}_\gamma$.

Example (Simple linear regression (cont.))

- Consider $X = X_1$ a continuous regressor and assume $X \overset{\text{iid}}{\sim} N(\mu_x, \sigma_x^2)$, so $\boldsymbol{\nu} = (\mu_x, \sigma_x)$.
- The priors to assign can be expressed: $\pi_0(\beta_0, \sigma, \nu) = \pi_0(\beta_0, \sigma \mid \mu_x, \sigma_x)\pi_0(\mu_x, \sigma_x)$, and

$$\pi_1(\beta_1, \beta_0, \sigma, \nu) = \pi_1(\beta_0, \sigma \mid \mu_x, \sigma_x)\,\pi_1(\beta_1 \mid \beta_0, \sigma, \mu_x, \sigma_x)\pi_1(\mu_x, \sigma_x)$$

- The reference prior for X model is $\pi(\mu_x, \sigma_x) = \sigma_x^{-1}$, we consider: $\pi_0(\mu_x, \sigma_x) = \pi_1(\mu_x, \sigma_x) = \sigma_x^{-1}$.

About $\pi_\gamma(\boldsymbol{\theta}_0 \mid \boldsymbol{\nu})$

- Consider a non-informative prior as $\boldsymbol{\theta}_0$ are common parameters.
- Reasonable when $\boldsymbol{\theta}_0$ have a similar interpretation in all models, in this case we should use an objective estimation prior.

Example (Simple linear regression (cont.))

Common parameters are: $\boldsymbol{\theta}_0 = (\beta_0, \sigma)$.

- Under $\mathcal{M}_0$, β_0 represents the mean of *all* y, under M_1 it is the mean of $y \mid x = 0$. Since x has mean μ_x the meaning of both β_0 can be very different.
- To achieve similar meaning, $\rightarrow$ a reparametrization under $\mathcal{M}_1$: $\beta_0^* = \beta_0 + \beta\mu_x$ then $y_i \mid x_i \sim N(\beta_0^* + \beta(x_i - \mu_x), \sigma^2)$.
- Now the prior over common parameters is:

$$\pi_0(\beta_0, \sigma \mid \boldsymbol{\nu}) = \sigma^{-1} \text{ and } \pi_1^*(\beta_0^*, \sigma \mid \boldsymbol{\nu}) = \sigma^{-1}$$

- The most delicate ingredient in the prior assignment.
- Enters into the equation for the $m_\gamma(\cdot)$ in a multiplicative-way. Not possible to use an improper prior, its indeterminate constant will be transferred to the marginal, it will not cancel in the BF calculation (different for each $\mathcal{M}_\gamma$).
- The prior $\pi_\gamma(\boldsymbol{\theta}_\gamma \mid \boldsymbol{\nu})$ has to be proper.
- For full observed data, and within the g-prior approach:

$$\boldsymbol{\theta}_\gamma \sim N_p(\mathbf{0}, \boldsymbol{V}_\gamma),$$

with $\boldsymbol{V}_\gamma$ obtained from the expected Fisher information matrix under $\mathcal{M}_\gamma$

- Many popular proposals in the literature are generalizations of this basic idea: For normal linear models, Benchmark priors (Fernandez et al., 2001), hyper-g priors (Liang et al., 2008); robust prior (Bayarri et al., 2012); etc.

Which V_γ with missing data?

- Revisiting the original definition of V_γ in the problem of regression.
- Consider $\boldsymbol{\theta}_0 = (\beta_0, \sigma)$ and $\boldsymbol{\theta}_\gamma \equiv \boldsymbol{\beta}_\gamma$,
- Definition of the prior covariance matrix in the Zellner and Siow, (1980) proposal is

$$V_\gamma = n\big(I(\beta_0)/I(\beta_0, \boldsymbol{\beta}_\gamma)\big)^{-1}$$

- n times the Schur complement of $I(\beta_0)$ in $I(\beta_0, \boldsymbol{\beta}_\gamma)$, the Fisher information matrix for $(\beta_0, \boldsymbol{\beta}_\gamma)$.
- Equal to using the variance matrix of the m.l.e. of $\boldsymbol{\beta}_\gamma$ (Bayarri et al., 2012)
- In normal linear models, with full-observed fixed design matrix it is:

$V_\gamma = n\sigma^2\,(\bar{X}_\gamma^T \bar{X}_\gamma)^{-1}$, with $\bar{X}_\gamma$ made by columns centered around the mean.

Result (Variance matrix)

Suppose $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i) \sim \mathcal{M}_\gamma$ *where*

$$(y_i, \boldsymbol{x}_i) \overset{\text{iid}}{\sim} N(y_i \mid \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_\gamma, \sigma^2)\, f(\boldsymbol{x}_i \mid \boldsymbol{\nu}),$$

then, provided f *has at least the first two moments:*

$$n^{-1}\big(I(\beta_0)/I(\beta_0, \beta_\gamma)\big) = \frac{1}{\sigma^2} V(\boldsymbol{x}_i \mid \boldsymbol{\nu})$$

where $V(\boldsymbol{x}_i \mid \boldsymbol{\nu}) = E[(\boldsymbol{x}_i - E(\boldsymbol{x}_i \mid \boldsymbol{\nu}))^T (\boldsymbol{x}_i - E(\boldsymbol{x}_i \mid \boldsymbol{\nu}))]$ *with expectations respect to* $f(\boldsymbol{x}_i \mid \boldsymbol{\nu})$.

Prior over $\boldsymbol{\beta}_\gamma$

- Our proposal to incorporate the g-priors into the missing context is consider:

$$\boldsymbol{\beta}_\gamma \mid \boldsymbol{\nu}, \beta_0, \sigma \sim N(\mathbf{0},\, \sigma^2\, V(\boldsymbol{x}_i \mid \boldsymbol{\nu})^{-1})$$

- Or with flat-tailed alternatives:

$$\boldsymbol{\beta}_\gamma \mid \boldsymbol{\nu}, \beta_0, \sigma \sim \int N(\mathbf{0},\, g\,\sigma^2\, V(\boldsymbol{x}_i \mid \boldsymbol{\nu})^{-1})\pi(g)\, dg.$$

- Different $\pi(g)$ leads to different well known priors for model selection.
- $V(\boldsymbol{x}_i \mid \boldsymbol{\nu})$ is a completely valid component of the conditional (on $\boldsymbol{\nu}$) prior covariance matrix.
- In fact, this prior is even more Bayesian than the g-prior as it does not depend on n nor on the data $\boldsymbol{x}$.

Result (Priors in linear regression)

To compare $H_1 \equiv \mathcal{M}_\gamma$: $(y_i, \boldsymbol{x}_i) \overset{\text{iid}}{\sim} N(y_i \mid \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_\gamma, \sigma^2)\, f(\boldsymbol{x}_i \mid \boldsymbol{\nu})$ *versus the null model (only intercept), the priors under each hypothesis are:*

$$\begin{aligned} \pi_0(\beta_0, \sigma, \boldsymbol{\nu}) &= \quad \sigma^{-1} \pi^N(\boldsymbol{\nu}), \\ \pi_1(\beta_0, \sigma, \boldsymbol{\beta}_\gamma, \boldsymbol{\nu}) &= \sigma^{-1}\, N(\boldsymbol{\beta}_\gamma \mid 0, g\sigma^2\, V(\boldsymbol{x}_i \mid \boldsymbol{\nu})^{-1})\, \pi^N(\boldsymbol{\nu}) \end{aligned}$$

with $g = 1$, *(or flat-tailed versions,* $g \sim \pi(g)$*) where* $\pi^N(\boldsymbol{\nu})$ *is an appropriate objective prior for* $\boldsymbol{\nu}$ *in relation to* $f(\boldsymbol{x}_i \mid \boldsymbol{\nu})$

Example (Simple linear regression (cont.))

- $X \overset{\text{iid}}{\sim} N(\mu_x, \sigma_x^2)$, so $\boldsymbol{\nu} = (\mu_x, \sigma_x)$, in this case $V(\boldsymbol{x}_i \mid \boldsymbol{\nu}) = \sigma_x^2$, then:

$$\pi_1(\beta \mid \beta_0, \sigma, \sigma_x, \mu_x) = N(\beta \mid 0, \frac{\sigma^2}{\sigma_x^2}).$$

- With the parameterized version of $\mathcal{M}_1$ to define a common prior over $\boldsymbol{\theta}_0$,
- Using the previous result to obtain the prior in this parameterization:

$$n^{-1}\big(I(\beta_0^*)/I(\beta_0, \beta_\gamma)\big) = \frac{1}{\sigma^2} V(x_i - \mu_x \mid \boldsymbol{\nu}) = \frac{1}{\sigma^2} V(x_i \mid \boldsymbol{\nu}),$$

  in the simple regression example: $\pi_1^*(\beta \mid \beta_0^*, \sigma, \sigma_x, \mu_x) = N(\beta \mid 0, \frac{\sigma^2}{\sigma_x^2})$.
- Then, $\pi_0(\beta_0, \sigma, \mu_x, \sigma_x) = (\sigma\sigma_x)^{-1}$ and for $\mathcal{M}_1^*$:
  $\pi_1^*(\beta_0^*, \sigma, \beta, \mu_x, \sigma_x) = (\sigma\sigma_x)^{-1} N(\beta \mid 0, \frac{\sigma^2}{\sigma_x^2})$.
- Same priors in terms of the original problem $\mathcal{M}_0$ vs. $\mathcal{M}_1$ (associated Jacobian is 1).

- The Bayes factor of $\mathcal{M}_\gamma$ to $\mathcal{M}_0$ is obtained as:

$$B_\gamma = \frac{m_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})}{m_0(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})},$$

- In general scenarios, with missing values also in the null model, the previous BF can be approximated as ratio of the approximated marginals:

$$\widehat{B}_\gamma = \frac{m_\gamma\widehat{(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})}}{m_0\widehat{(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})}}$$

- In some cases, i.e. if $\mathcal{M}_0$ does not have missing values, it is possible to average also de BFs.

Result (Ratio of completed-predictive densities)

Under the conditions in the Result about priors for regression models, $\mathfrak{m}_0(\boldsymbol{y} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})$ *does not depend on* $\boldsymbol{\nu}$*, and under the null model this quantity does not depend on* $\widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}$ *neither. So,*

$$\frac{\mathfrak{m}_1(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})}{m_0(\widetilde{\boldsymbol{y}})} = \left[\frac{SSE_0}{SSE_0 - \widetilde{\boldsymbol{y}}^T \bar{X}\left(\bar{X}^T\bar{X} + V(\boldsymbol{x}_i \mid \boldsymbol{\nu})/g\right)^{-1}\bar{X}^T\widetilde{\boldsymbol{y}}}\right]^{(n-1)/2} \times \left|\bar{X}^T\bar{X}\,V(\boldsymbol{x}_i \mid \boldsymbol{\nu})^{-1} + 1/g\,\boldsymbol{I}\right|^{-1/2}, \qquad (2)$$

$\bar{X}$ *is the completed design matrix (filled with* $\widetilde{\boldsymbol{x}}_{(0)}$ *and* $\boldsymbol{x}_{(1)}$*) with columns centered with respect to their mean and* SSE_0 *is the sum of residuals under* $\mathcal{M}_0$*,* $\boldsymbol{I}$ *denotes the* $p \times p$ *identity matrix.*

Result (BF example linear regression)

Under the same conditions of the above results and using the priors proposed above, the expression for the BF for $\mathcal{M}_1$ versus $\mathcal{M}_0$ is:

$$
\begin{aligned}
B_{10}(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}) &= \frac{\int \mathfrak{m}_1(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu}) \prod_{i=m+1}^{n} f(x_i^{na} \mid \boldsymbol{\nu})\, \pi_1(\boldsymbol{\nu} \mid \widetilde{\boldsymbol{x}}_{(0)}) d\boldsymbol{x}_{(1)}\, d\boldsymbol{\nu}}{m_0(\widetilde{\boldsymbol{y}})} \\
&= \int \frac{\mathfrak{m}_1(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})}{m_0(\widetilde{\boldsymbol{y}})} \prod_{i=m+1}^{n} f(x_i^{na} \mid \boldsymbol{\nu})\, \pi_1(\boldsymbol{\nu} \mid \widetilde{\boldsymbol{x}}_{(0)}) d\boldsymbol{x}_{(1)}\, d\boldsymbol{\nu}
\end{aligned}
$$

The previous BF can be approximated as:
For $j = 1, \ldots, N$:

- Step 1: Draw $\boldsymbol{\nu}^{(j)} \sim \pi(\boldsymbol{\nu} \mid \widetilde{\boldsymbol{x}}_{(0)})$,
- Step 2: draw $(\boldsymbol{x}_{(1)})^{(j)} \sim f(\boldsymbol{x}_{(1)} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{\nu}^{(j)})$,
- Step 3: compute the ratio $r_{10}^{j} = \frac{\mathfrak{m}_1(\widetilde{\boldsymbol{y}} | \widetilde{\boldsymbol{x}}_{(0)}, (\boldsymbol{x}_{(1)})^{(j)}, \boldsymbol{\nu}^{(j)})}{m_0(\widetilde{\boldsymbol{y}})}$ given in equation (2).

Approximate

$$B_{10}(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}) \approx N^{-1} \sum_{i=1}^{N} r_{10}^{j}$$

- Finally, for the comparison of the two hypothesis in the example, using $P(H_0) = P(H_1) = 1/2$, the posterior probability for H_1 is:

$$P(H_1 \,|\, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}) = \frac{B_{10}(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})}{1 + B_{10}(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})}$$

- Finally, for the comparison of the two hypothesis in the example, using $P(H_0) = P(H_1) = 1/2$, the posterior probability for H_1 is:

$$P(H_1 \,|\, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}) = \frac{B_{10}(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})}{1 + B_{10}(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})}$$

- However, in the *Biometrics* paper (2005), in any imputed data, $P(H_1 \,|\, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, (\boldsymbol{x}_{(1)})^{(j)})$ is calculated, and the final posterior probability for H_1 is obtained as a mean:

$$\frac{1}{N}\sum_{j=1}^{N} P(H_1 \mid \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)}, (\boldsymbol{x}_{(1)})^{(j)}) = \frac{1}{N}\sum_{j=1}^{N} \frac{BF^j}{1 + BF^j}$$

- Not admissible because of Jensen's inequality.

Example (Simple linear regression)

- We have simulated from:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & \rho^* \\ \rho^* & 1 \end{pmatrix}\right), \quad Y \mid X_1 = x_1, X_2 = x_2 \sim N_1(1+\beta_1^* x_1+\beta_2^* x_2, 1),$$

  with ρ^*, β_1^* and β_2^* are prefixed values used to reproduce several real scenarios.
- The complete simulated data D consist on $n = 100$ draws from $\boldsymbol{Z}$.
- Interest: whether X_1 is a explanatory variable for Y:

$$\begin{aligned} H_0 &: f_0(y \mid x_1, x_2, \beta_0, \sigma) = N(y \mid \beta_0, \sigma^2), \\ H_1 &: f_1(y \mid x_1, x_2, \beta_0, \beta, \sigma) = N(y \mid \beta_0 + \beta_1 x_1, \sigma^2). \end{aligned}$$

- If D would be completely observed, the Bayesian answer is the posterior probability $p(\mathcal{M}_1 \mid D)$. This is the oracle response.
- We simulate a MAR mechanism for a proportion π in X_1, obtaining $(\boldsymbol{y}, \widetilde{\boldsymbol{x}}_{(0)})$.

Example (Simple linear regression)

Consider 3 scenarios:

E1: $\beta_1^* = 0.3$, $\beta_2^* = 0$,

E2: $\beta_1^* = \beta_2^* = 0$,

E3: $\beta_1^* = 0$, $\beta_2^* = 0.3$.

For each scenario $N = 100$ datasets are generated for the combinations:

- $\rho^* \in \{0, 0.4, 0.7, 0.9\}$
- $\pi \in \{0.05, 0.15, 0.40, 0.60, 0.75\}$

Example (Simple linear regression)

Integrating out X_2 in the data generative process,

$$Y \mid X_1 = x_1 \sim N_1(1 + \beta_2^*(2 - \rho^*) + (\beta_1^* + \rho^* \beta_2^*)x_1, 1 + (\beta_2^*)^2(1 - (\rho^*)^2)).$$

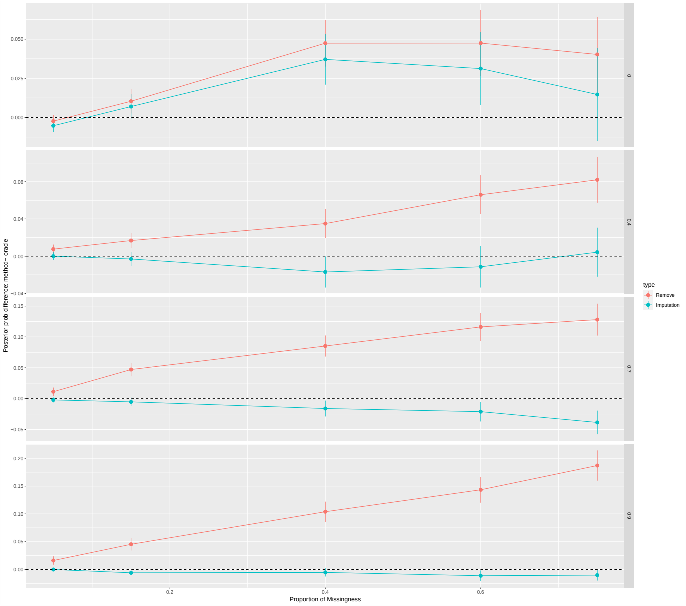
The ratio of the signal to the standard deviation is;

$$S^2 = \frac{(\beta_1^* + \rho^* \beta_2^*)^2}{1 + (\beta_2^*)^2(1 - (\rho^*)^2)}.$$

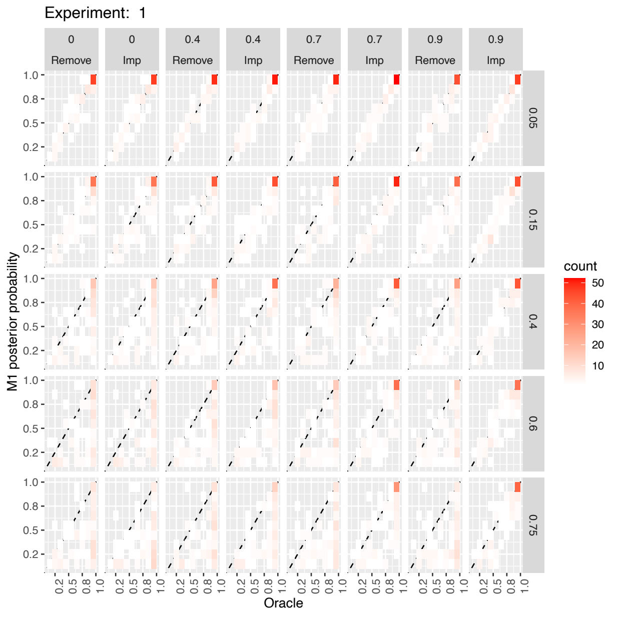
Summarized in the considered scenarios:

	$\rho^* = 0$	$\rho^* = 0.4$	$\rho^* = 0.7$	$\rho^* = 0.9$
E1	0.3^2	0.3^2	0.3^2	0.3^2
E2	0	0	0	0
E3	0	0.12^2	0.21^2	0.27^2

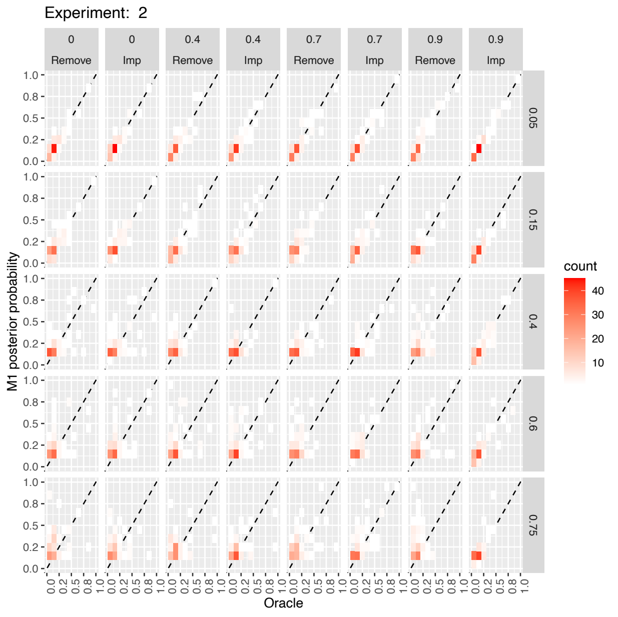
Results comparing Imputation with Remove

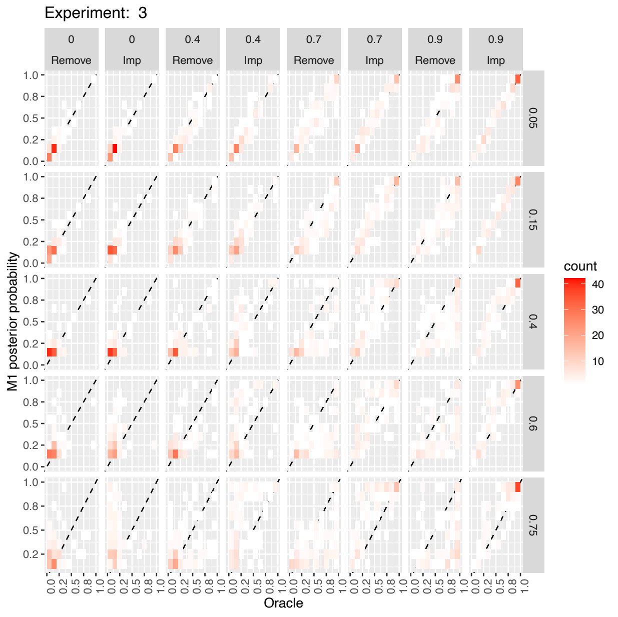


Results Imputation vs Remove. Probabilities Exp 1

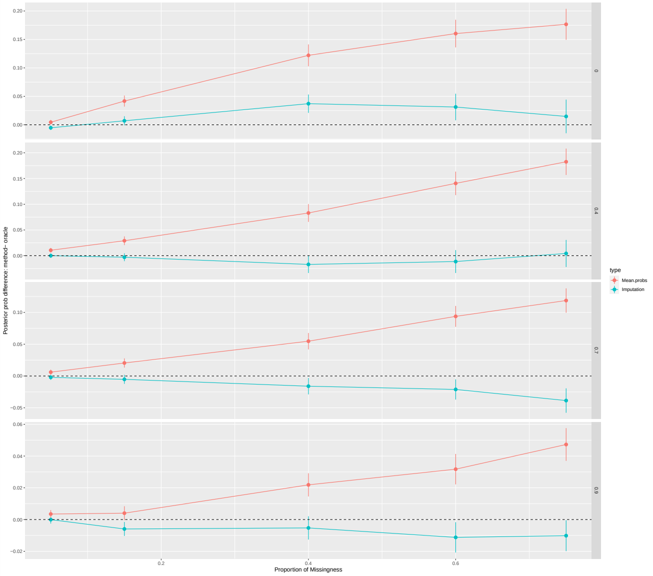


Results Imputation vs Remove. Probabilities Exp 2

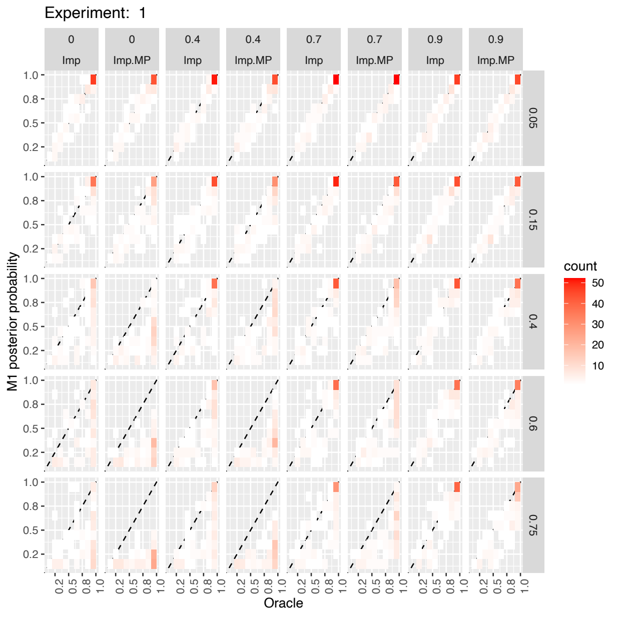


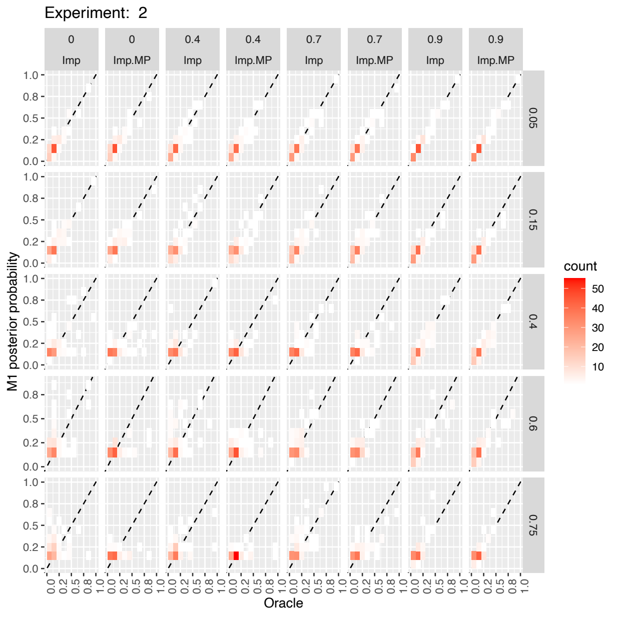

Experiment: 3
0 Remove
0 Imp
0.4 Remove
0.4 Imp
0.7 Remove
0.7 Imp
0.9 Remove
0.9 Imp
0.05
0.15
0.4
0.6
0.75
M1 posterior probability
Oracle
count
40
30
20
10
0.0
0.2
0.5
0.8
1.0

Results comparing Imputation with Mean probabilities

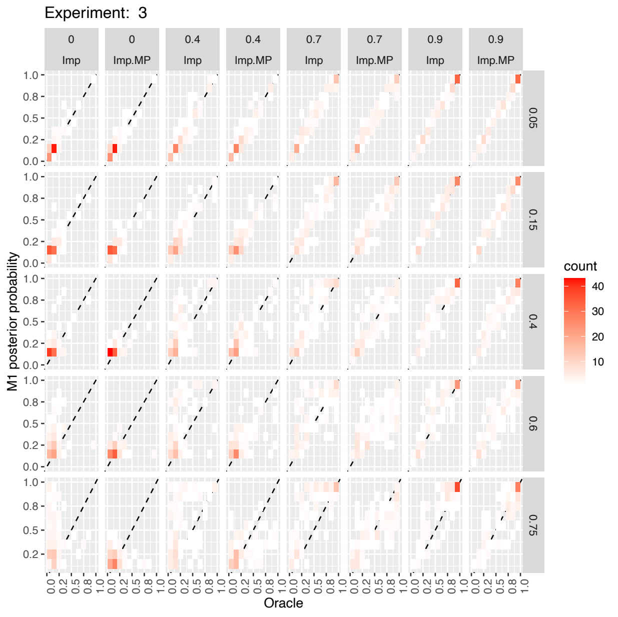


Results Imputation vs Mean Prob. Exp 1




Experiment: 2
0
Imp
0
Imp.MP
0.4
Imp
0.4
Imp.MP
0.7
Imp
0.7
Imp.MP
0.9
Imp
0.9
Imp.MP
0.05
0.15
0.4
0.6
0.75
M1 posterior probability
Oracle
count
50
40
30
20
10

Results Imputation vs Mean Prob. Exp 3



- $m_\gamma(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{x}}_{(0)})$: always an average of marginals obtained on imputed values: $m_\gamma(\widetilde{\boldsymbol{y}} \mid \widetilde{\boldsymbol{x}}_{(0)}, \boldsymbol{x}_{(1)})$, over the posterior predictive $\boldsymbol{x}_{(1)} \mid \widetilde{\boldsymbol{x}}_{(0)}$.
- Bayes factors: sometimes an average of BFs calculated in imputed data, in the sense of the MI word. Not longer true when $\mathcal{M}_0$ has also missing values.
- Posterior probabilities: never an average of posterior probabilities calculated in imputed data.

Present/Future work

- Study characteristics as predictive matching criteria and others in the predictive marginals obtained with the proposal priors.
- Analyze and develop the variable selection problem in the context of comparing a set of possible models.
- Analyze how to search in the space of all models when p (number of covariates is large) with missing data and exhaustive calculations are not possible.

Thanks for the attention! ‡

‡Grant PID2019-104790GB-I00 funded by

MINISTERIO
DE CIENCIA
E INNOVACIÓN

AGENCIA
ESTATAL DE
INVESTIGACIÓN

Some References

- Bayarri, M. J., J. O. Berger, A. Forte, and G. García-Donato (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics.*
- Fernández, C., E. Ley, and M. F. Steel (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics.*
- Hoijtink H., Gu, X., Mulder, J. and Rosseel, Y. (2019). Computing Bayes Factors from Data with Missing Values, *Psychological Methods.*
- Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association.*
- Little, R. and D. Rubin (2020). *Statistical Analysis with Missing Data* (3rd ed.). Wiley
- Yang, X., Belin, T. R. and Boscardin, W. J. (2005). Imputation and Variable Selection in Linear Regression Models with Missing Covariates. *Biometrics.*
- Zellner, A. and A. Siow (1980). Posterior odds for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics*