Marco Ferreira's discussion of "Model uncertainty quantification in presence of missing data," by Castellanos et al.

Marco A.R. Ferreira

Virginia Tech, Department of Statistics

marf@vt.edu

OBayes 2022

Table of Contents

Table of Contents

Linear regression model

Let's consider the particular case of variable selection in linear regression model. Model $\mathcal{M}_\gamma$ is

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

In designed experiments (with no missing data), Bayesians usually consider the marginal density of the dependent variable conditional on the covariates and model $\mathcal{M}_\gamma$:

$$m_\gamma(\mathbf{y} \mid \mathbf{X}_\gamma) = \int f_\gamma(\mathbf{y} \mid \mathbf{X}_\gamma, \beta_0, \boldsymbol{\beta}_\gamma, \sigma^2)\pi_\gamma(\beta_0, \boldsymbol{\beta}_\gamma, \sigma^2 \mid \mathbf{X}_\gamma) d\beta_0 d\boldsymbol{\beta}_\gamma d\sigma^2.$$

Then, assuming centered covariates ($\mathbf{X}_\gamma' \mathbf{1} = \mathbf{0}$), a typical prior distribution over $\boldsymbol{\beta}_\gamma$ would be a Zellner-Siow prior:

$$\pi_\gamma(\boldsymbol{\beta}_\gamma | \mathbf{X}_\gamma, \sigma^2) = \int N(\boldsymbol{\beta}_\gamma \mid 0, g\sigma^2(\mathbf{X}_\gamma' \mathbf{X}_\gamma)^{-1})\pi(g) dg,$$

where different choices of $\pi(g)$ lead to different well-known priors (e.g., $\pi(g) = IG(g \mid 1/2, n/2)$ corresponds to the Zellner-Siow prior).

Problem: How to assign a prior for $\boldsymbol{\beta}_\gamma$ when some observations have missing covariate information?

To solve this problem, Castellanos et al. propose

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

or

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

Questions:

- Would the g-prior, Zellner-Siow prior, and similarly constructed priors arise as empirical Bayes implementations of the above proposals?
- And, in the case of fully observed data (no missingness), how different would model selection decisions based on your proposed priors be from model selection decisions based on the g-prior and the Zellner-Siow prior?

Table of Contents

Model definition and model comparison

Recall:

- Missingness matrix $M = (m_{ij})$ with $m_{ij} = 1$ if x_{ij} is missing and $m_{ij} = 0$ otherwise.
- Tilde used to indicate a realization of a random variable or random vector.
- $\widetilde{\mathbf{x}}_{(0)} = \{\widetilde{x}_{ij} : \widetilde{m}_{ij} = 0\}$ is the set of observed values of the covariates.
- $\mathbf{x}_{(1)} = \{x_{ij} : \widetilde{m}_{ij} = 1\}$ is the set of missing values of the covariates.

Model definition and model comparison

Castellanos et al. propose that model $\mathcal{M}_\gamma$ should include the regression model, the generating model for the covariates, and the model for the missingness:

- $f_\gamma(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_p, \beta_0, \boldsymbol{\beta}_\gamma)$;
- $f(\mathbf{x}_1, \ldots, \mathbf{x}_p|\boldsymbol{\nu})$;
- $f(M|\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_p, \psi)$.

Then, model selection decisions would be based on the joint predictive density $m_\gamma(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_{(0)}, \widetilde{M}) = \int m_\gamma(\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}_{(0)}, \mathbf{x}_{(1)}, \widetilde{M}) d\mathbf{x}_{(1)}$.

Suggestion: I think this is a sensible proposal when the focus is variable selection. And this suggests that if we want to investigate the missingness process (e.g., if we suspect that there may have been fraud in the collection of the data), we may want to consider $f_\gamma(M|\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_p, \psi)$.

Question: What difficulties do you foresee in that more general case?