Supplementary material for "Bayesian Survival Modelling of University Outcomes"

C.A. Vallejos and M.F.J. Steel

This document extends the descriptive analysis of the PUC dataset and provides further empirical results on the effects of covariates. We also describe the marginal likelihood estimator used in the implementation. In addition, documentation for the freely available R code is provided. Throughout, Sections refer to the paper.

A. Descriptive analysis of PUC dataset

Table 1 breaks down the percentage of students satisfying the inclusion criteria (see Section 2) by program. This inclusion percentage is above 78% for the programmes analyzed in Section 5.

Table 1: PUC dataset. Amount of students satisfying the inclusion criteria using in this study by program.

Program	No. students	% students
Acting	362	80.1
Agronomy and Forestry Engineering	2,466	85.2
Architecture	841	69.9
Art	688	76.3
Astronomy	295	88.3
Biochemistry	331	85.5
Biology	791	83.9
Business Administration and Economics	2,027	72.7
Chemistry	379	82.0
Chemistry and Pharmacy	687	85.6
Civil Construction	1,930	86.0
Design	651	65.2
Education, elementary school	1,277	81.4
Education, elementary school (Villarrica campus)	301	80.5
Education, preschool	949	83.2
Engineering	3,522	69.3
Geography	534	84.5
History	552	76.6
Journalism and Media Studies	876	76.2
Law	2,303	84.2
Literature (Spanish and English)	911	80.8
Mathematics and Statistics	598	78.0
Medicine	972	89.8
Music	161	74.5
Nursing	886	78.6
Physics	237	85.9
Psychology	801	75.9
Social Work	440	87.5
Sociology	421	74.0
Total	27,189	78.7

Figures 1 to 8 summarize a more complete descriptive analysis of the PUC dataset. These Figures confirm strong levels of heterogeneity between different programmes of the PUC. As described in Section 2, this suggests modelling each programme independently.

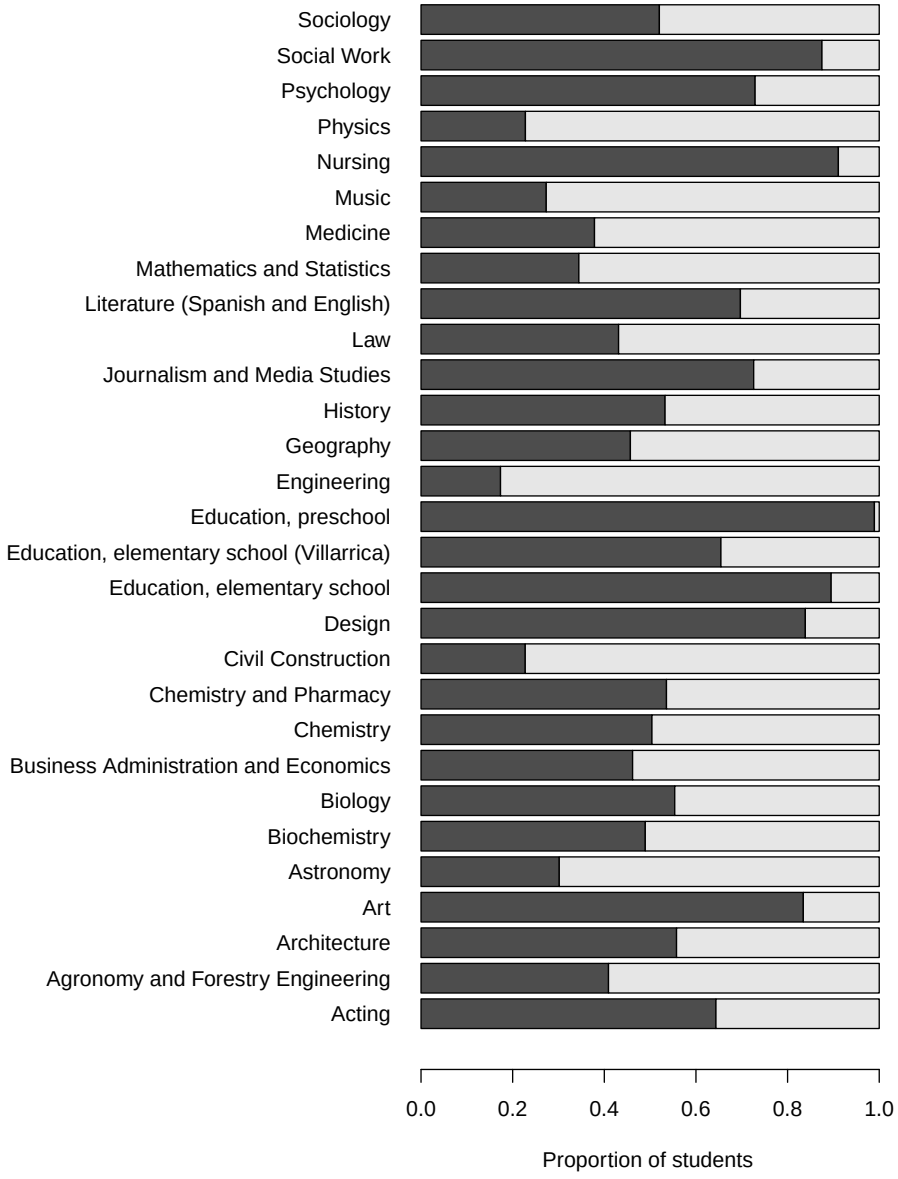

Figure 1: Distribution of students according to sex (lighter area: males).

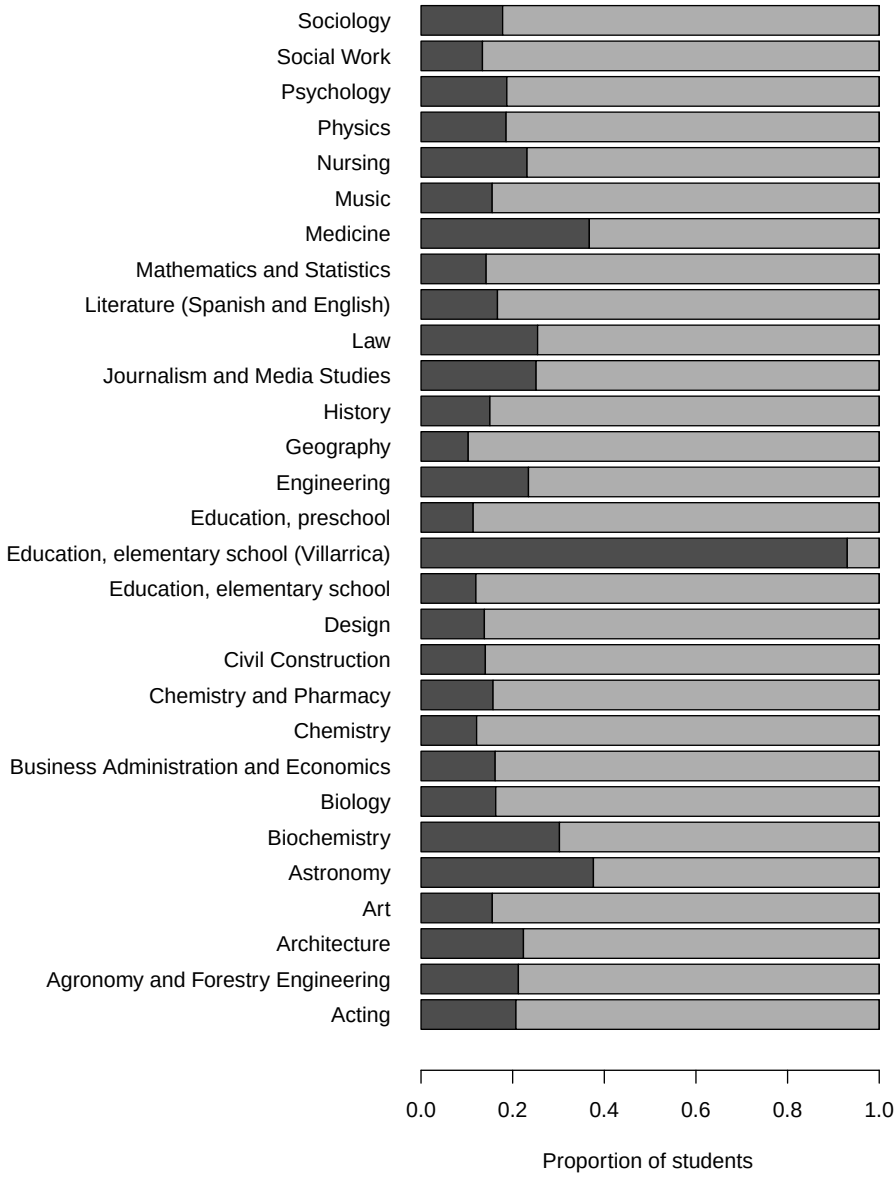

Figure 2: Distribution of students according to region of residence (lighter area: Metropolitan area).

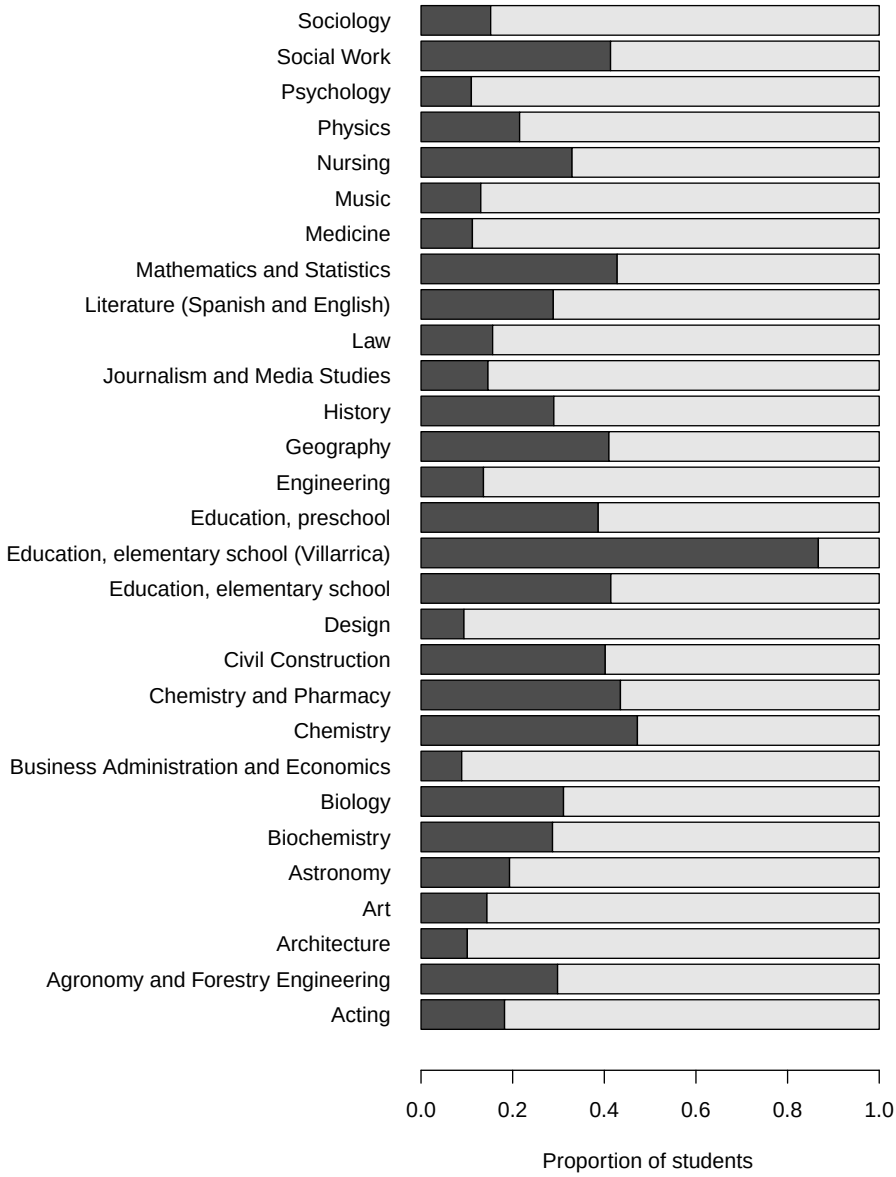

Figure 3: Distribution of students according to educational level of the parents (lighter area: students for which at least one of the parents has a university or technical degree).

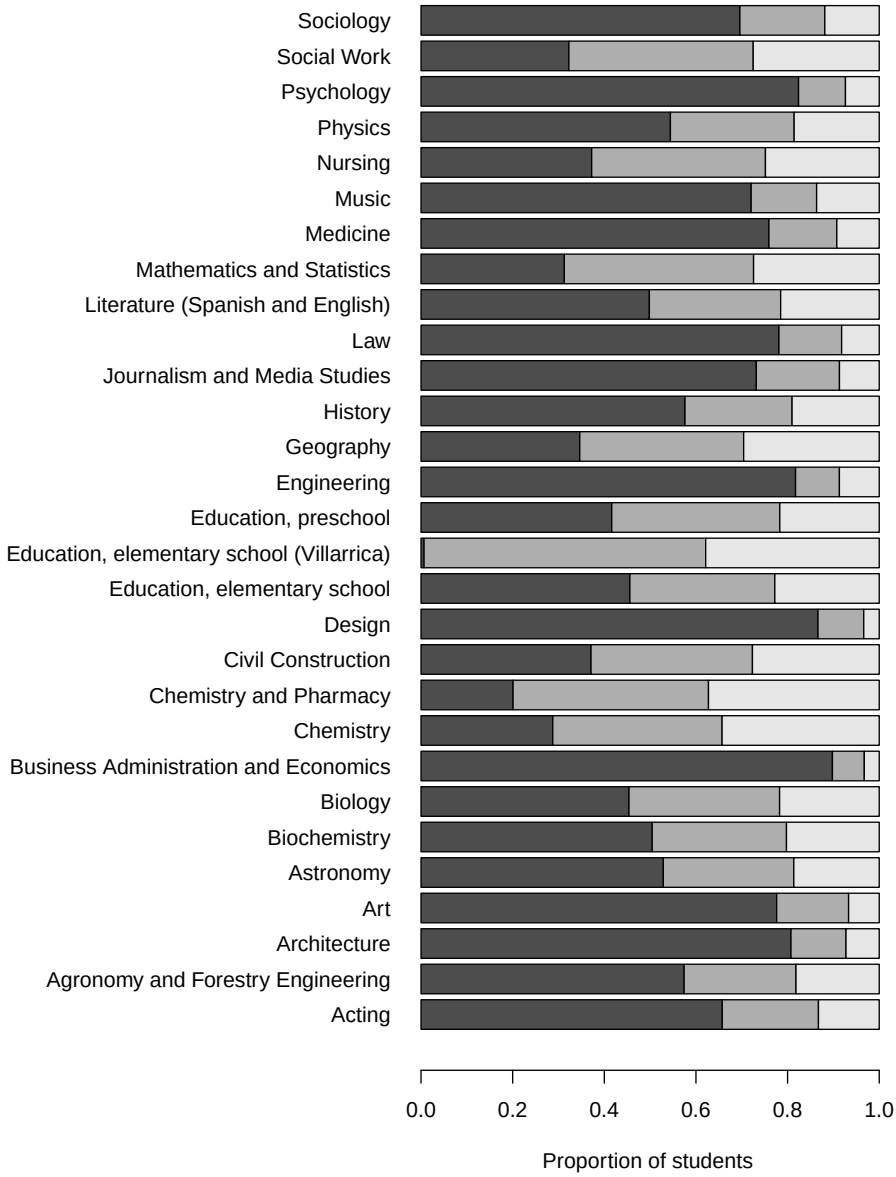

Figure 4: Distribution of students according to type of high school (from darkest to lightest, colored areas represent the proportion of students whose high school was: private, subsidized private and public, respectively).

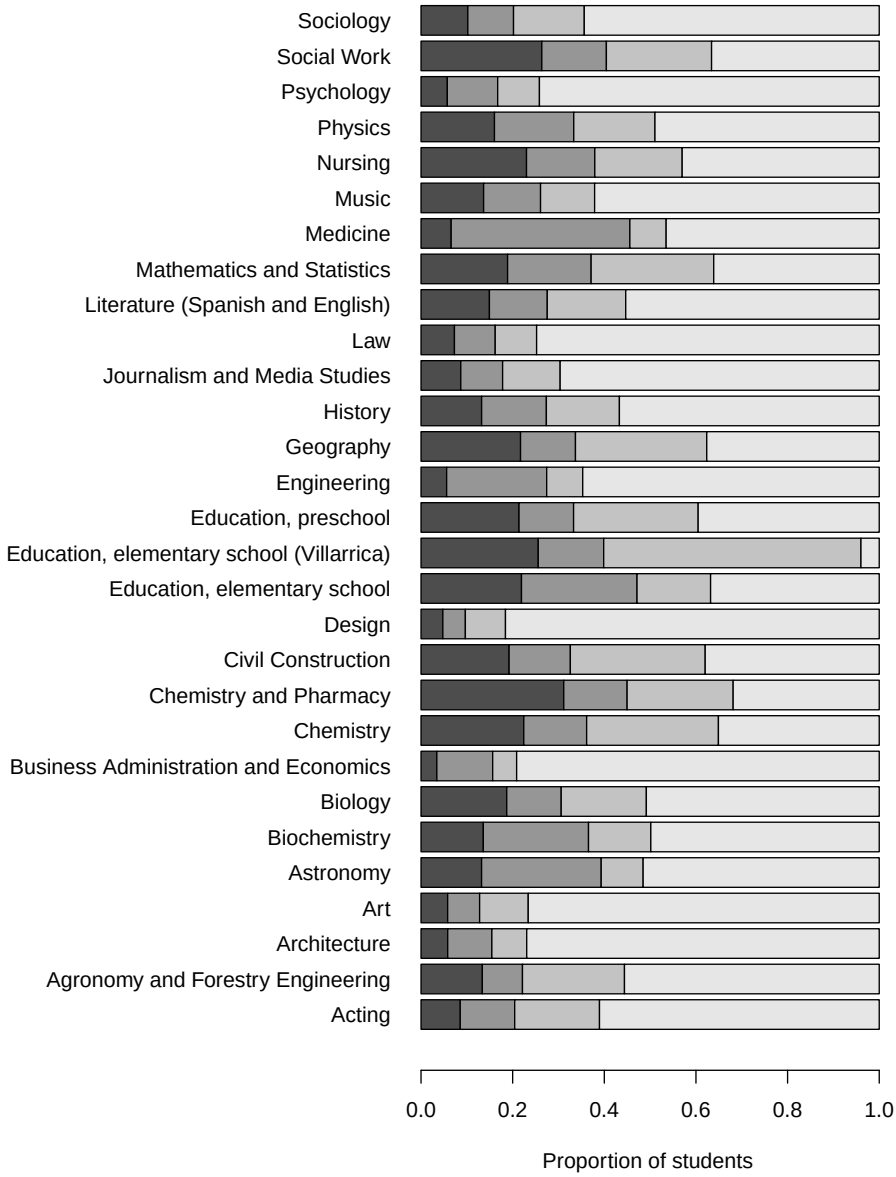

Figure 5: Distribution of students according to funding (from darkest to lightest, colored areas represent the proportion of students who have: scholarship and loan, scholarship only, loan only and no aid, respectively).

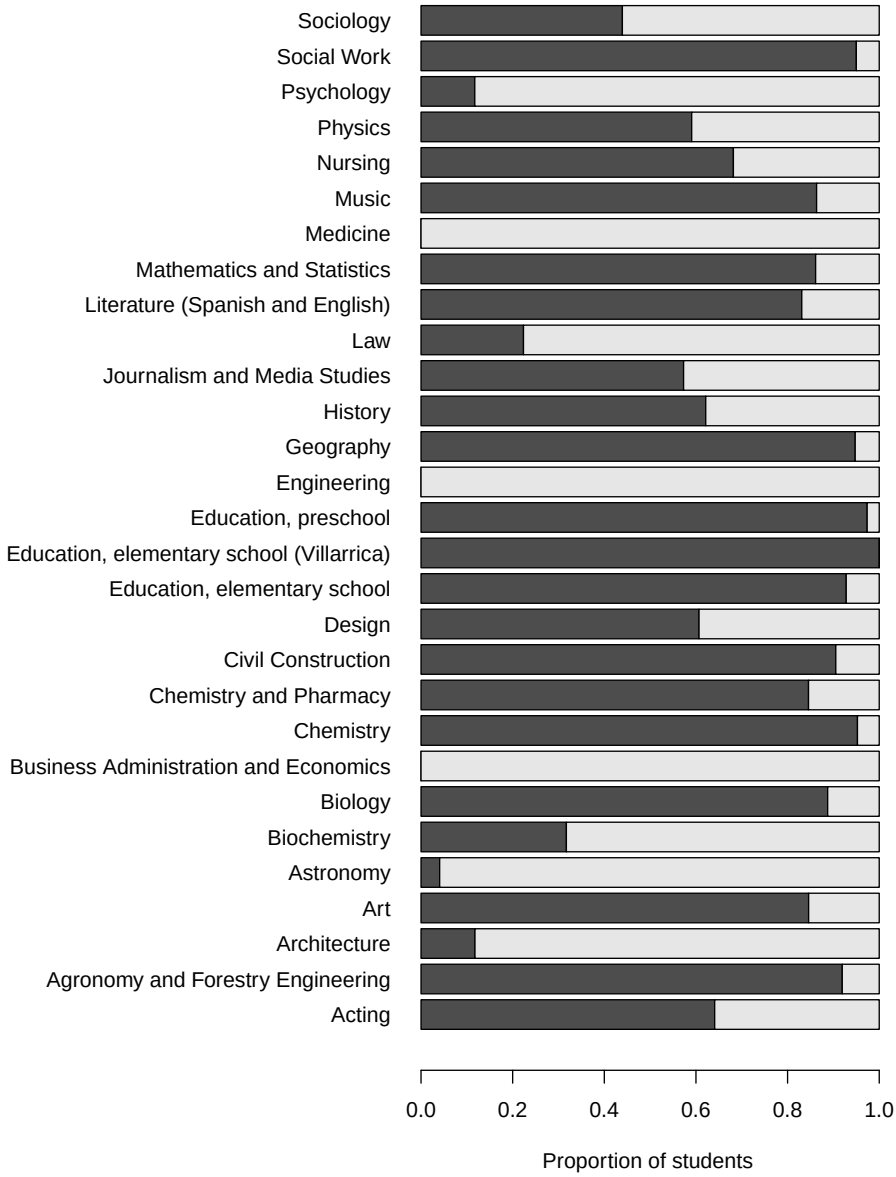

Figure 6: Distribution of students according to their selection score (lighter area: students with a selection score of 700 or more, which is typically considered a high value - the maximum possible score is 850). The minimum score required when applying to the PUC is 600 but exceptions apply for some education-related programmes.

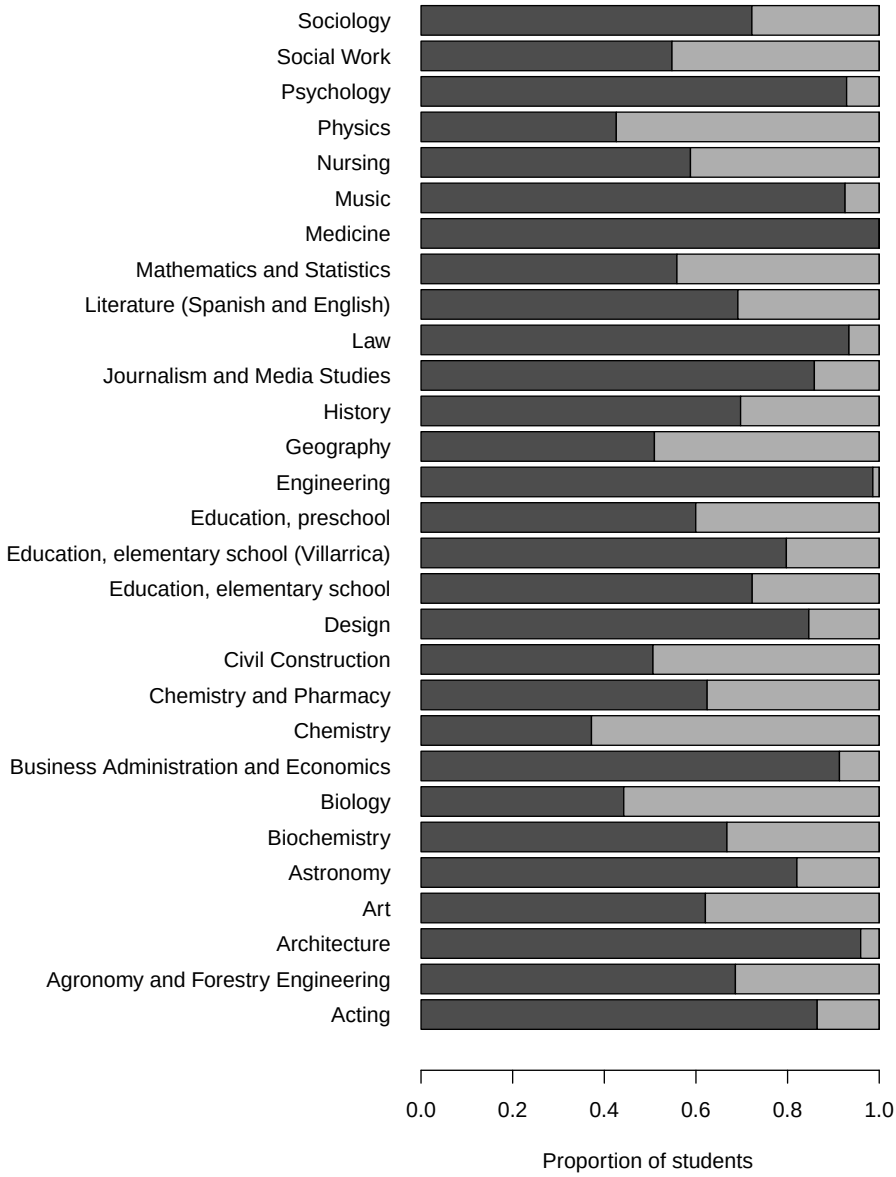

Figure 7: Distribution of students according to their application preference (lighter area: students who applied with second or lower preference to their current degree).

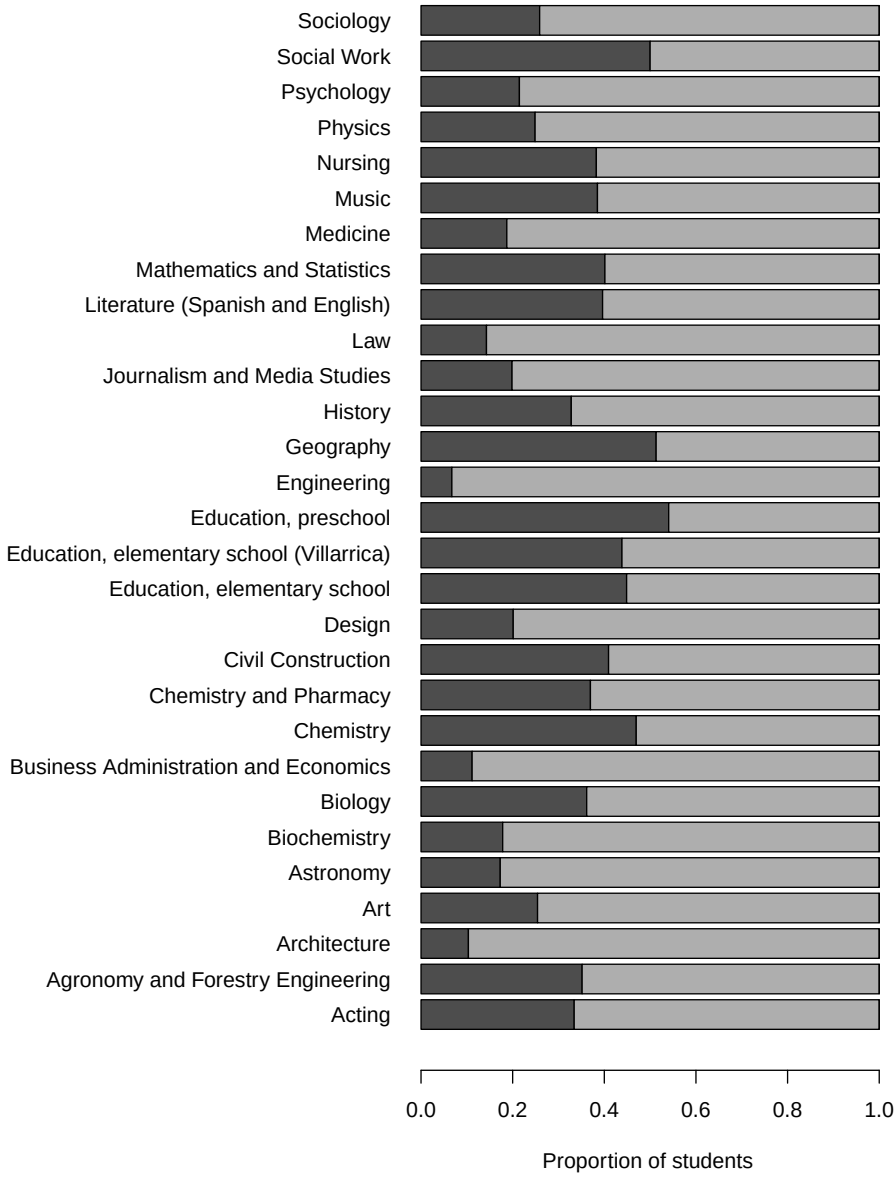

Figure 8: Distribution of students according to the gap between High School graduation and admission to PUC (lighter area: students with no gap).

B. Further empirical results

Figures 9 to 16 display the continuous component of the posterior distribution of the regression coefficients for the analyzed degree programmes. The effects that were already shown in the paper are excluded.

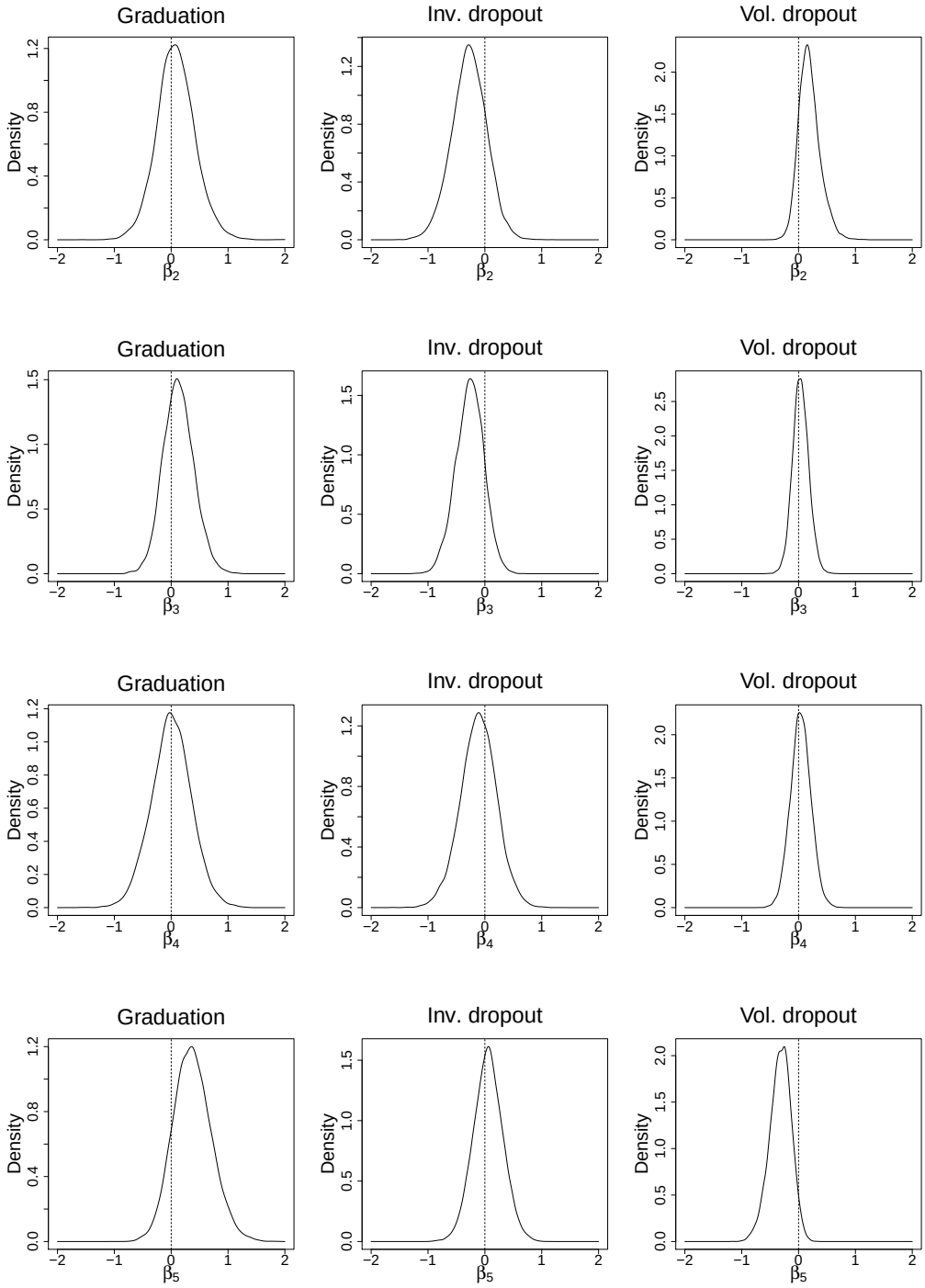

Figure 9: Chemistry students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: region - metro area (β_2), parents' education - with degree (β_3), high school - private (β_4) and high school - subsidized private (β_5). A vertical dashed line was drawn at zero for reference. The prior in equation (19) of the paper was adopted for the model space.

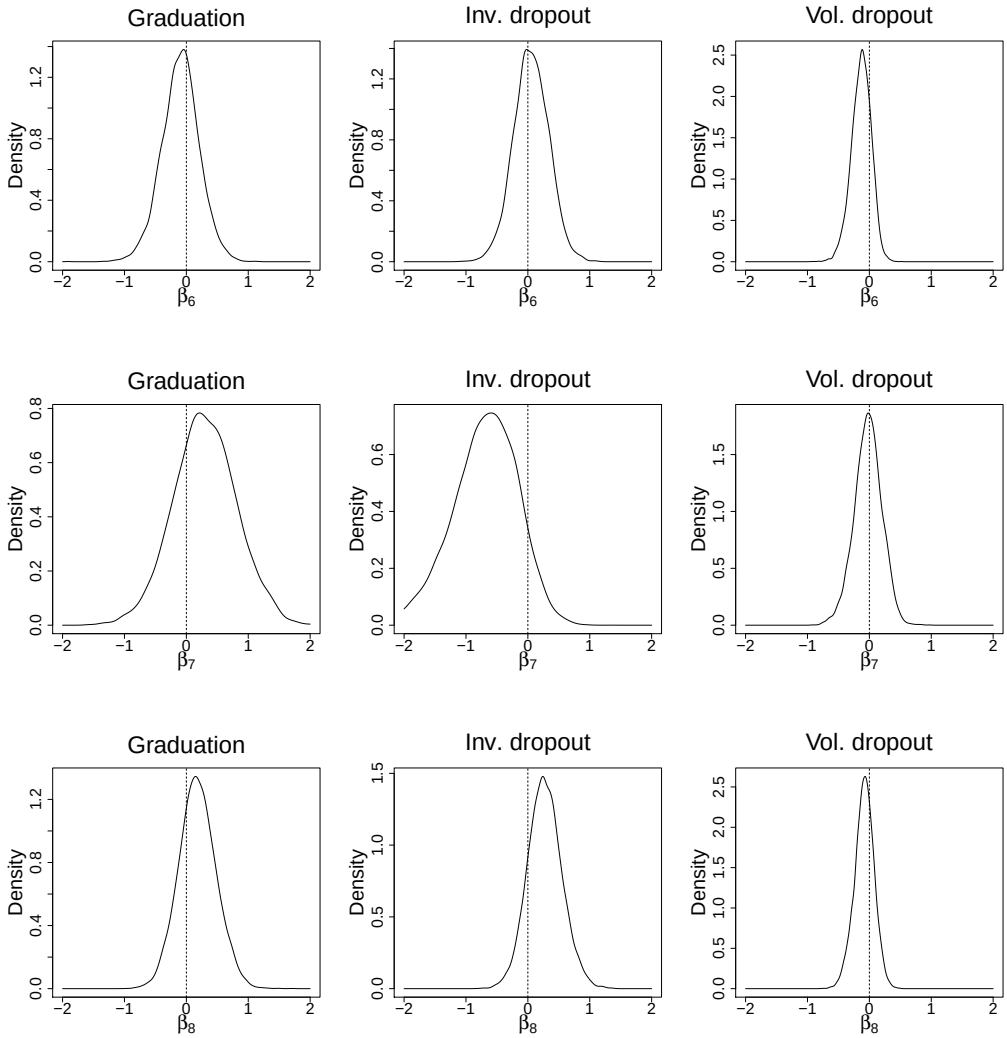

Figure 10: Chemistry students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: funding - scholarship only (β_6), funding - scholarship and loan (β_7) and funding - loan only (β_8). A vertical dashed line was drawn at zero for reference. The prior in equation (19) of the paper was adopted for the model space.

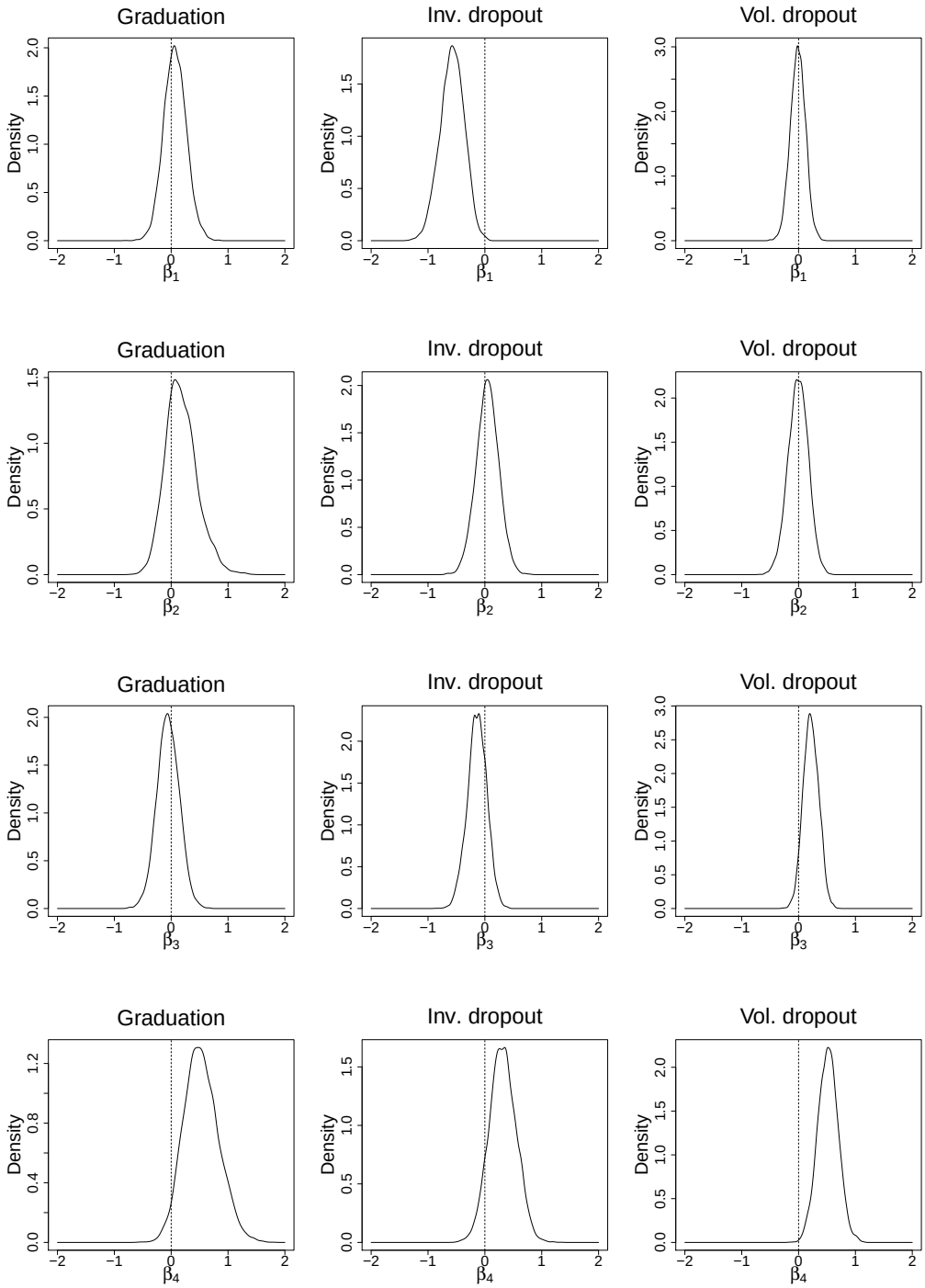

Figure 11: Mathematics and Statistics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: sex - female (β_1), region - metro area (β_2), parents' education - with degree (β_3) and high school - private (β_4). A vertical dashed line was drawn at zero for reference. The prior in equation (19) of the paper was adopted for the model space.

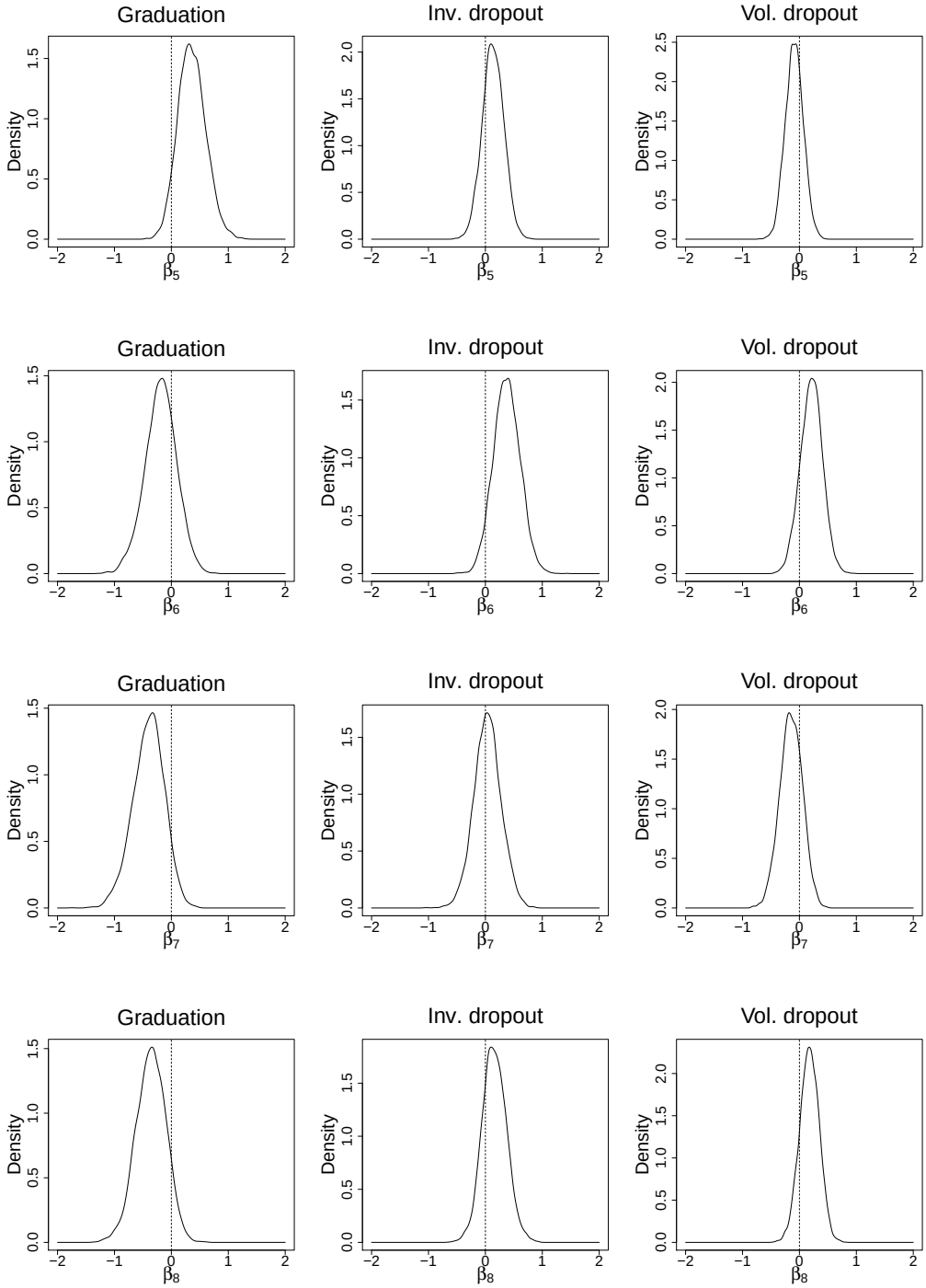

Figure 12: Mathematics and Statistics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: high school - subsidized private (β_5), funding - scholarship only (β_6), funding - scholarship and loan (β_7) and funding - loan only (β_8). A vertical dashed line was drawn at zero for reference. The prior in equation (19) of the paper was adopted for the model space.

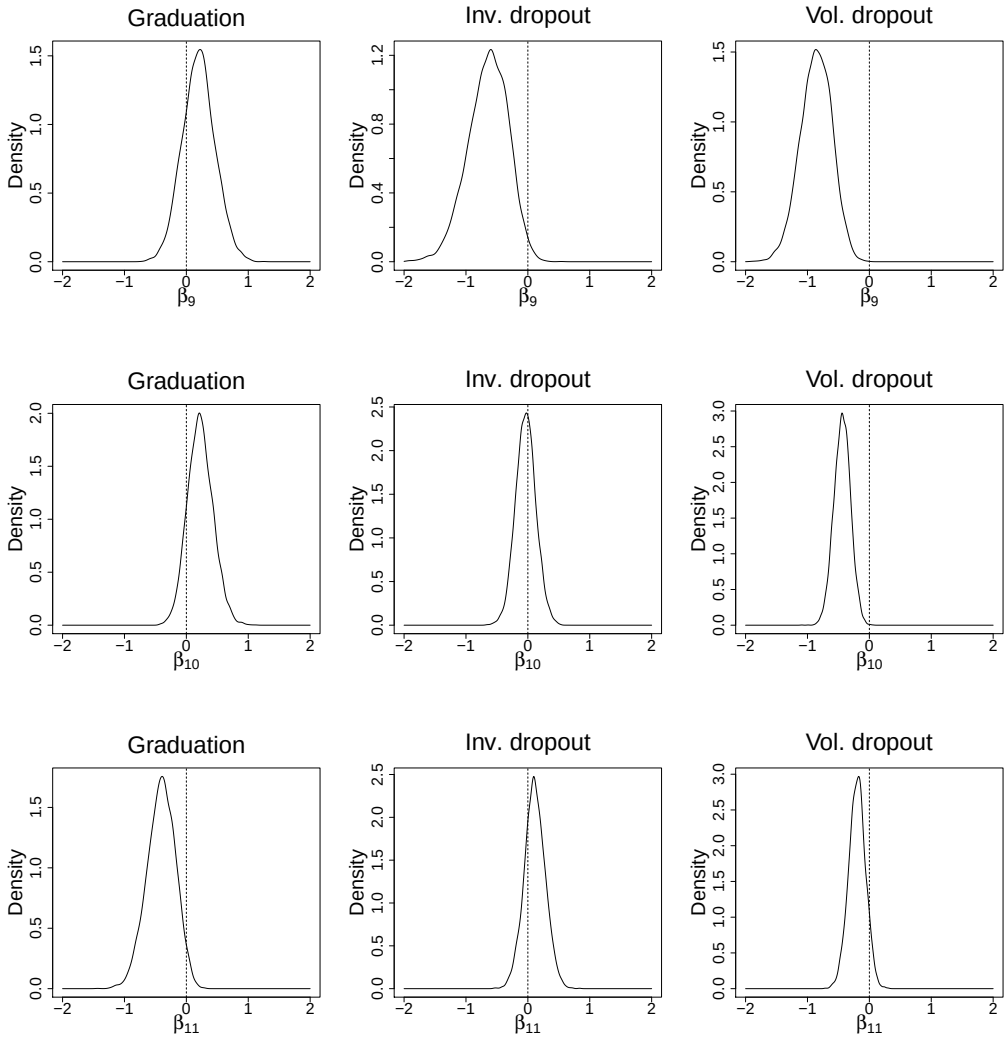

Figure 13: Mathematics and Statistics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: selection score - top 10% (β_9), preference - first (β_{10}) and gap - yes (β_{11}). A vertical dashed line was drawn at zero for reference. The prior in equation (19) of the paper was adopted for the model space.

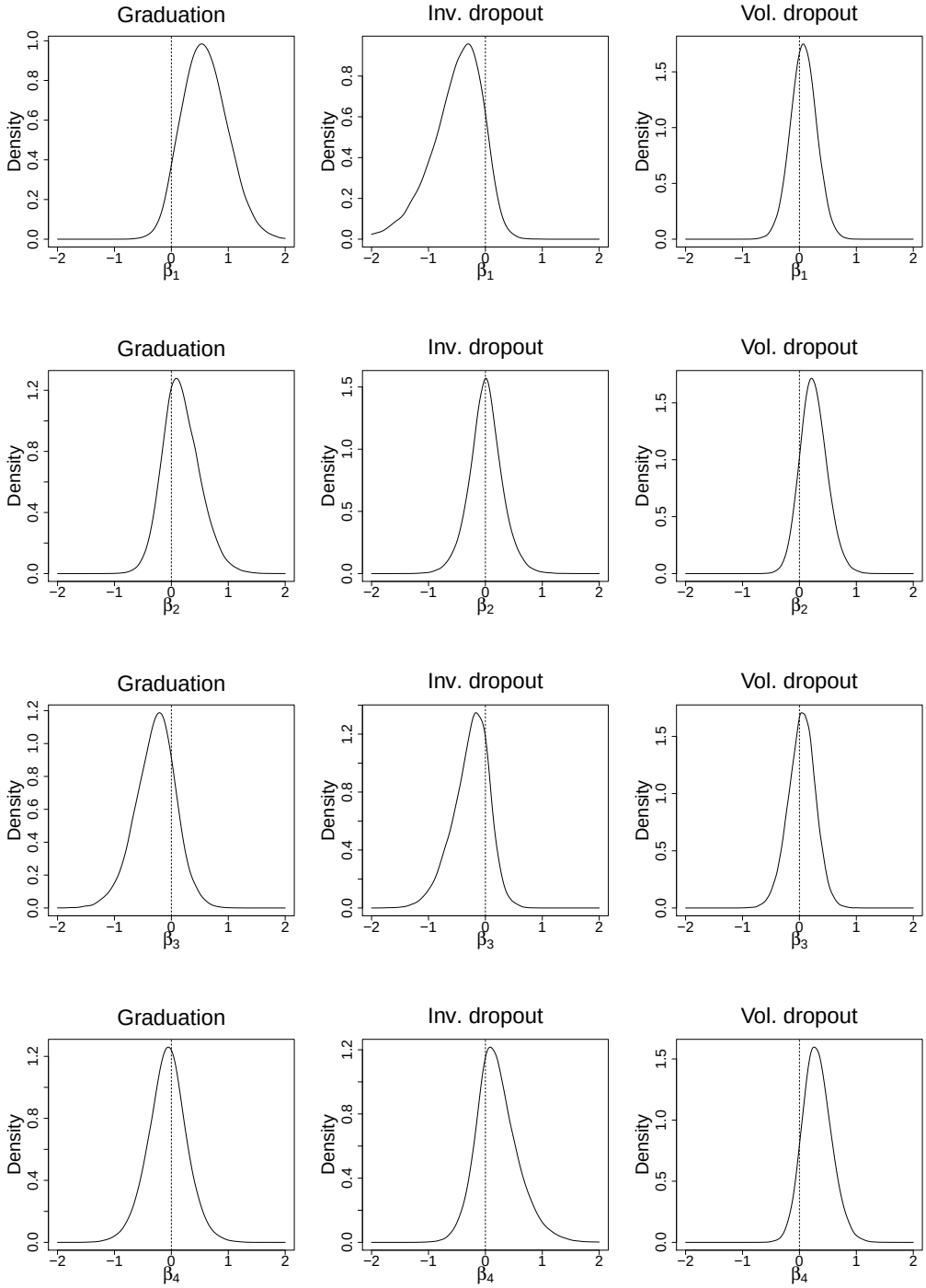

Figure 14: Physics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: sex - female (β_1), region - metro area (β_2), parents' education - with degree (β_3) and high school - private (β_4). A vertical dashed line was drawn at zero for reference. The prior in equation (19) of the paper was adopted for the model space.

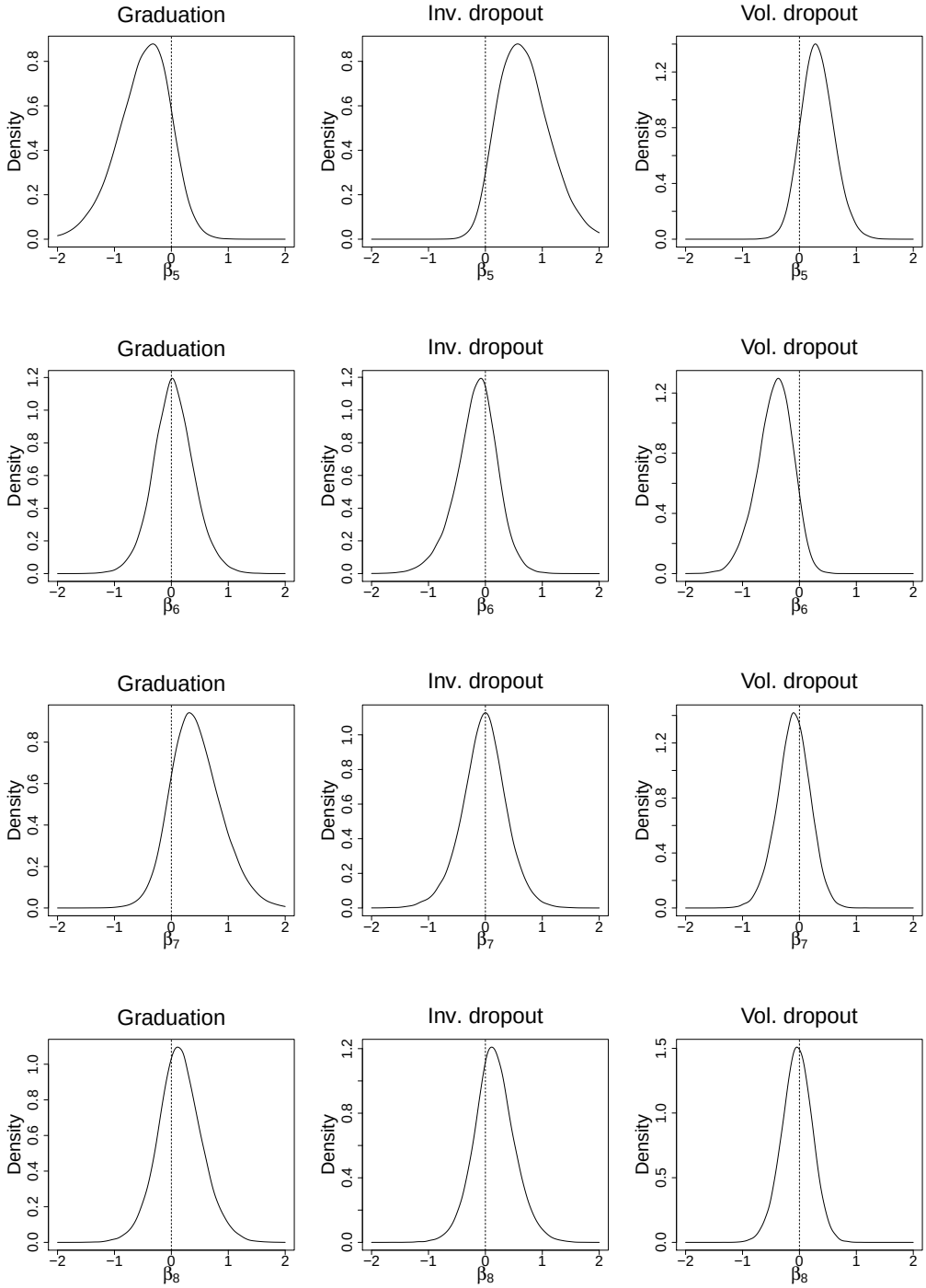

Figure 15: Physics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: high school - subsidized private (β_5), funding - scholarship only (β_6), funding - scholarship and loan (β_7) and funding - loan only (β_8). A vertical dashed line was drawn at zero for reference. The prior in equation (19) of the paper was adopted for the model space.

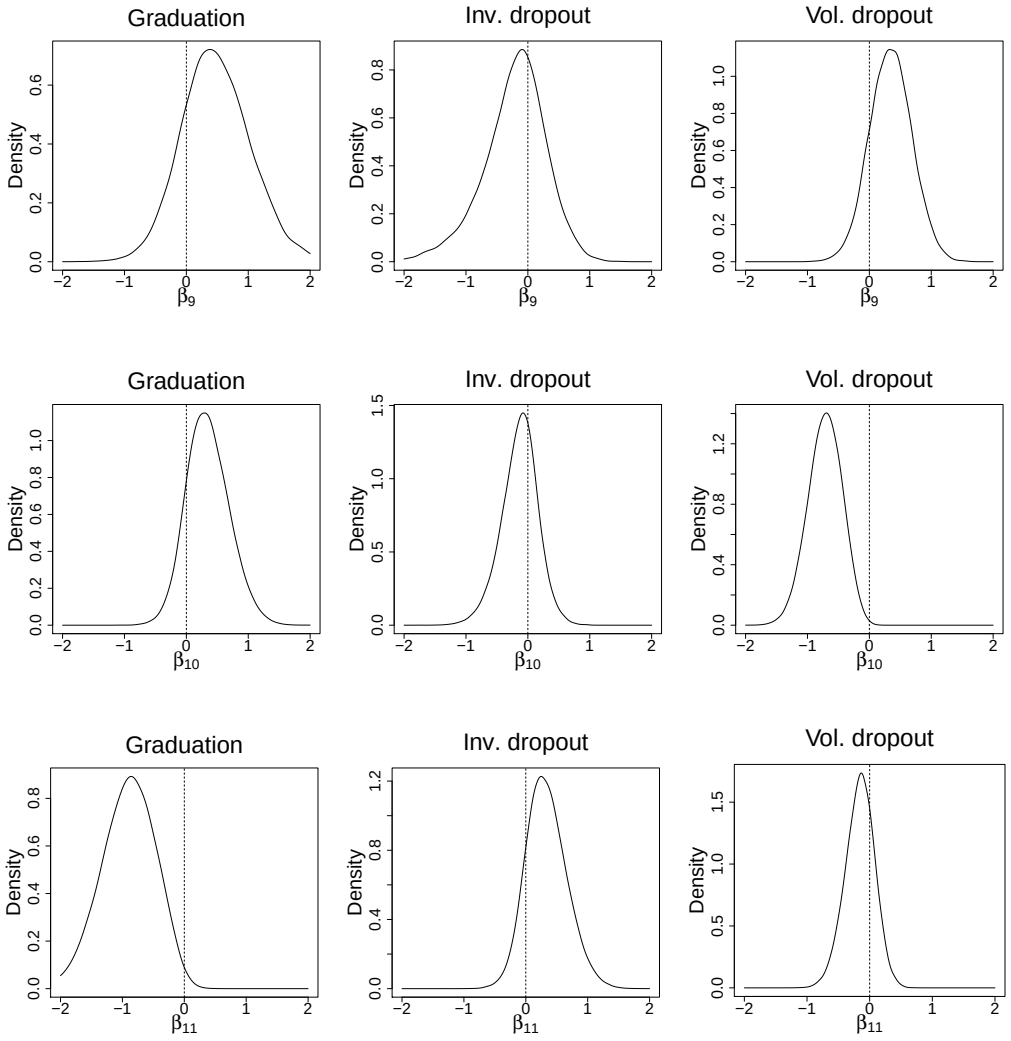

Figure 16: Physics students: posterior density (given that the corresponding covariate is included in the model) of some selected regression coefficients: selection score - top 10% (β_9), preference - first (β_{10}) and gap - yes (β_{11}). A vertical dashed line was drawn at zero for reference. The prior in equation (19) of the paper was adopted for the model space.

C. Marginal likelihood estimator

Computing marginal likelihoods is a very challenging endeavour. A survey of several methods is provided in Section 7.3 of Robert [2007]. In this paper we employ the Bridge sampling method proposed in Meng and Wong [1996]. Let $g_0(\cdot)$ and $g_1(\cdot)$ be two densities for a random variable ψ sharing the same support and that are known only up to proportionality constants c_0 and c_1, respectively. Inspired by the physics literature, Meng and Wong [1996] show that for any arbitrary *bridge* function $\alpha(\cdot)$ (such that the required expectations exist), it follows that

$$r = \frac{c_1}{c_0} = \frac{\mathrm{E}_{g_0}\left(\tilde{g}_1(\psi)\alpha(\psi)\right)}{\mathrm{E}_{g_1}\left(\tilde{g}_0(\psi)\alpha(\psi)\right)}, \tag{1}$$

where $\tilde{g}_0(\cdot)$ and $\tilde{g}_1(\cdot)$ are the known un-normalized versions of $g_0(\cdot)$ and $g_1(\cdot)$, respectively. The expectations in (1) are with respect to $g_0(\cdot)$ and $g_1(\cdot)$, respectively. Using (1), the *bridge sampling* estimator of c_1/c_0 is defined as

$$\hat{r}_\alpha = \frac{1/n_0 \sum_{i=1}^{n_0} \tilde{g}_1(\psi_{0i})\alpha(\psi_{0i})}{1/n_1 \sum_{i=1}^{n_1} \tilde{g}_0(\psi_{1i})\alpha(\psi_{1i})}, \tag{2}$$

where $\psi_{01}, \ldots, \psi_{0n_0}$ and $\psi_{11}, \ldots, \psi_{1n_1}$ are random samples from $g_0(\cdot)$ and $g_1(\cdot)$, respectively. If draws within each of these samples are independent, Meng and Wong [1996] deduced that the variance of $\log(\hat{r_\alpha})$ is minimized when

$$\alpha^*(\psi) \propto \frac{1}{s_1\tilde{g}_1(\psi) + r s_0 \tilde{g}_0(\psi)}, \tag{3}$$

where $s_j = n_j/(n_0 + n_1)$, $j = 0, 1$. As discussed in Meng and Schilling [2002], dependencies between the draws are not to critical for this optimization as long as they are weak. Since r is unknown, $\alpha^*(\psi)$ cannot be directly used. Nevertheless, given an initial guess $\hat{r}_\alpha^{(0)}$, an optimal bridge estimator can be defined iteratively as

$$\hat{r}_\alpha^{(m+1)} = \frac{1/n_0 \sum_{i=1}^{n_0} l_{0i}/(s_1 l_{0i} + s_0 \hat{r}_\alpha^{(m)})}{1/n_1 \sum_{i=1}^{n_1} 1/(s_1 l_{1i} + s_0 \hat{r}_\alpha^{(m)})}, \quad m = 1, 2, \ldots \tag{4}$$

where $l_{ji} = \tilde{g}_1(\psi_{ji})/\tilde{g}_0(\psi_{ji})$, $i = 1, \ldots, n_j$, $j = 0, 1$. The latter defines a consistent estimator of r. Nonetheless, the method in Meng and Wong [1996] is restrictive in the sense that it requires the same support for $g_0(\cdot)$ and $g_1(\cdot)$. In particular, this condition does not hold when the aim is to estimate the BF between two models M_0 and M_1 which have different number of parameters (*e.g.* in variable selection). As a solution, Chen and Shao [1997] proposed to augment the smaller support, introducing a correction factor in (2). Alternatively, Meng and Schilling [2002] suggested a different solution that computes c_0 and c_1 independently. They pointed out that (2) defines an estimator of c_1 when $\tilde{g}_0(\psi)$ is replaced by an arbitrary auxiliary normalized density $g(\psi)$ which has the same support as $g_1(\psi)$. The latter approach is adopted throughout the paper, where the marginal likelihood of each possible model is computed by defining $g(\psi)$ as a multivariate normal density (specific to each model) with parameters $\hat{\mu}_{\beta^*}$ and $0.2\hat{\Sigma}_{\beta^*}$, where $\hat{\mu}_{\beta^*}$ and $\hat{\Sigma}_{\beta^*}$ are the estimated posterior median and variance-covariance matrix of the corresponding β^*, respectively. This choice was adopted in order to avoid numerical problems (*e.g.* very small or large values of the l_{ji}'s).

D. Documentation for the R code

Bayesian inference is implemented through the Markov chain Monte Carlo (MCMC) sampler described in Subsection 4.2, under the priors presented in Subsection 4.1. Inference was implemented in R[1] version 3.0.1. The code is freely available at
http://www.warwick.ac.uk/go/msteel/steel_homepage/software/university_codes.zip.
This includes the MCMC algorithm and the Bayesian variable selection methods described in the paper. Before using this code, the following libraries must be installed in R: BayesLogit, MASS, mvtnorm, Matrix and compiler. All of these are freely available from standard R repositories and are loaded in R when "Internal_Codes.R" is executed. The last two libraries speed up matrix calculations and the "for" loops, respectively. Table 2 explains the notation used throughout the code. The implementation was based on three-dimensional arrays, with the third dimension representing the event type.

Table 2: Notation used throughout the R code

Variable name	Description
n	Total number of multinomial outcomes (*i.e.* $\sum t_i$ across all students)
t0	Number of period-indicators δ_{rt} (for each cause)
k	Number of effects (t0 + number of covariate effects)
CATEGORIES	Number of possible outcomes, excluding censoring (equal to 3 for the PUC dataset)
inc	Vector containing covariate indicators $\gamma_1, \ldots, \gamma_{k*}$
X.Period	Design matrix related to the period-indicators δ_{rt}'s only. Dimension n $\times$ t0
Y	Vector of outcomes. Dimension: n $\times 1$
X	Design matrix, including the binary indicators (denoted by Z in the paper). Dimension: n $\times$ k
beta	β^* (period-indicators and covariates effects for all event types)
prior	Choice of hyper prior for g_r: *(i)* Benchmark-Beta or *(ii)* Hyper-g/n [see Ley and Steel, 2012]
fix.g	If TRUE, $g_1, \ldots, g_{\mathcal{R}}$ are fixed. Default value: FALSE
mean.beta	Prior mean for $\{\beta_1^*, \ldots, \beta_{\mathcal{R}}^*\}$. Dimension: $1\times$ k $\times$ CATEGORIES
prec.delta	Precision matrix for $(\delta_{r1}, \ldots, \delta_{rt_0})'$. Dimension: t0 $\times$ t0
df.delta	Degrees of freedom for prior of $(\delta_{r1}, \ldots, \delta_{rt_0})'$. Default value: 1
N	Total number of MCMC iterations
thin	Thinning period for MCMC algorithm
burn	Burn-in period for MCMC algorithm
beta0	Starting value for $\{\beta_1^*, \ldots, \beta_{\mathcal{R}}^*\}$. Dimension: $1\times$ k $\times$ CATEGORIES
logg0	Starting value for $\{\log(g_1), \ldots, \log(g_{\mathcal{R}})\}$. Dimension: $1\times$ CATEGORIES
ls.g0	Starting value for the logarithm of the proposal variance used in Metropolis-Hastings updates of $\log(g_1), \ldots, \log(g_{\mathcal{R}})$. Dimension: $1\times$ CATEGORIES
ar	Optimal acceptance rate for the adaptive Metropolis-Hastings updates. Default value: 0.44

The code is separated into two files. The file "Internal_Codes.R" contains functions that are required for the implementation but the user is not expected to directly interact with these. These functions must be loaded in R before doing any calculations. The remaining functions are contained in the file "User_Codes.R". In the following, a short description of these functions is provided. Their use is illustrated in the file "Example.R" using a simulated dataset.

- X.design. Creates a design matrix based on X.Period and covariate inclusion indicators inc. This function is based on the structure of the PUC dataset and must be modified if analyzing a different dataset.

[1] Copyright (C) The R Foundation for Statistical Computing.

- `ind.var`. Indicates active covariate effects based on covariate inclusion indicators `inc`. This function is based on the structure of the PUC dataset and must be modified if analyzing a different dataset.

- `MCMC.MLOG`. Adaptive Metropolis-within-Gibbs algorithm [Roberts and Rosenthal, 2009] for the competing risks Proportional Odds model used throughout the paper. If not fixed, univariate Gaussian random walk proposals are implemented for $\log(g_1), \ldots, \log(g_{\mathcal{R}})$. Arguments: `N`, `thin`, `Y`, `X`, `t0`, `beta0`, `mean.beta`, `prec.delta`, `df.delta`, `logg0`, `ls.g0`, `prior`, `ar` and `fix.g`. The output is a list containing the following elements: `beta` MCMC sample of β^* (array of dimension (`N/thin+1`) $\times$ `k` $\times$ `CATEGORIES`), `logg` MCMC sample of $\log(g_1), \ldots, \log(g_{\mathcal{R}})$ (dimension (`N/thin+1`) $\times$ `CATEGORIES`), `ls.g` stored values for the logarithm of the proposal variances for $\log(g_1), \ldots, \log(g_{\mathcal{R}})$ (dimension (`N/thin+1`) $\times$ `CATEGORIES`) and `lambda` MCMC sample of $\lambda_1, \ldots, \lambda_{\mathcal{R}}$), which are defined in equation (9) in the paper (dimension (`N/thin+1`) $\times$ `CATEGORIES`). Recording `ls.g` allows the user to evaluate if the adaptive variances have been stabilized. Overall acceptance rates are printed in the `R` console (if appropriate). This value should be close to the optimal acceptance rate `ar`.

- `DIC.MLOG`. Computes the DIC [Spiegelhalter et al., 2002] taking `Y`, `X` and a MCMC sample of β^* `chain.beta` as inputs. It is based on the deviance funcion $D(\beta^*) = -2\log(\prod_{i=1}^{n} L_i)$, where L_i is defined in equation (3) in the paper, but also incorporates a penalization factor for the complexity of the model. DIC is defined as

$$DIC \equiv \mathrm{E}(D(\beta^*)|\text{data}) + p_D = \mathrm{E}(D(\beta^*)|\text{data}) + [\mathrm{E}(D(\beta^*)|\text{data}) - D(\hat{\beta^*})], \tag{5}$$

  where $\hat{\beta^*}$ is the posterior median of β^* (existence of the posterior mean is not guaranteed) and p_D can be interpreted as the effective number of parameters of the model. This function returns a single number which is a Monte Carlo estimate of the DIC. The effective and actual number of model parameters are printed in the `R` console.

- `CaseDeletion.MLOG`. Leave-one-out cross validation analysis. The function returns a matrix with `n` rows. Its first column contains the logarithm of the $\text{CPO}_i = P(y_i|y_{-i})$ for $i = 1, \ldots, n$ [Geisser and Eddy, 1979], where y_i denotes the observations for individual i and y_{-i} is the observed data for the other individuals. This is computed as $P(y_i|y_{-i}) = [\mathrm{E}\,(1/L_i)\,|\text{data}]^{-1}$. The second and third columns contain the KL divergence between $\pi(\theta|d_{-i})$ and $\pi(\theta|d)$ and its calibration index p_i [Cho et al., 2009], respectively. The later can be used in order to evaluate the existence of influential observations. The Pseudo Marginal Likelihood (PsML) predictive criterion is defined as $\text{PsML} = \prod_{i=1}^{n} \text{CPO}_i$. The logarithm of PsML can be computed as the sum of `logCPO` across the `n` observations.

- `LMLBS.MLOG`. Marginal likelihood estimator (logarithmic scale) using Bridge Sampling [Meng and Wong, 1996, Meng and Schilling, 2002] based on an MCMC sample of β^*. It returns 5,000 iterations of the iterative estimator described in Section C of this Supplementary Material. This allows the user to assess if the estimator has stabilized, in which case the final value is the log marginal likelihood estimate.

References

M.H. Chen and Q.M. Shao. Estimating ratios of normalizing constants for densities with different dimensions. *Statistica Sinica*, 7:607–630, 1997.

H. Cho, J. G. Ibrahim, D. Sinha, and H. Zhu. Bayesian case influence diagnostics for survival models. *Biometrics*, 65:116–124, 2009.

S. Geisser and W.F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74:153–160, 1979.

E. Ley and M.F.J. Steel. Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics*, 171:251–266, 2012.

X.L. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11: 552–586, 2002.

X.L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

C.P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2nd edition, 2007.

G.O. Roberts and J.S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18:349–367, 2009.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, B*, 64:583–640, 2002.