Bayesian modeling of university first-year students' grades after placement test

To appear in the *Journal of Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–15

Bayesian modeling of university first-year students' grades after placement test

Antonio Canale[a*] Euloge Clovis Kenne Pagui[b] and Bruno Scarpa[b]

[a]*Department of Economics and Statistics, University of Turin and Collegio Carlo Alberto, Italy*;
[b]*Department of Statistical Sciences, University of Padua, Italy*

(*2015*)

University first-year students grades are naturally correlated with the scores obtained at placement tests. Often this characteristic leads the university grades in the first exams to be asymmetrically distributed. Motivated by the analysis of grades of the basic Statistics examination of first-year students, we discuss informative priors for the shape parameter of the skew-normal model, a class of distribution which account for several degree of asymmetry. Our proposed prior leads to closed-form full-conditional posterior distributions, particularly useful in Markov Chain Montecarlo simulation. A Gibbs sampling algorithm is discussed for the joint vector of parameters and the method is applied to a real dataset from the School of Economics, University of Padua, Italy. Our analysis reveals that the correlation between the placement test and the grades of first-year students leads to a measurable positive skewness of the distribution of the university grades.

Keywords: Gibbs sampling; Informative prior; Skew-normal distribution; Unified skew-normal distribution

Classification codes: 62F15; 62E15

1. Introduction

The relationships between placement tests and students' success in college has been widely studied and discussed in the literature. Although there has been a strong debate on which kind of tests better predict students' future performance, particularly for the SAT standardised tests [e.g., 16], it is clear that the results of such test is positively correlated with the students' performance. Indeed, both essay and multiple choice tests have been shown to correlate with grades in college courses [12] and particularly when the results of such preliminary tests are graded or have academic consequences such as admission to the school or not [22].

In many countries, such as in Italy, university grades are numerical values, thus we can assume that the results of placement tests and the grades of a specific first-year exam are normally distributed and correlated. Consider the case in which the placement test admits a student to the first year only if such student's score is above a given threshold denoting the mean of the distribution of the test score. From a probabilistic point of view, we need to consider a selection mechanism which starts from bivariate normal distribution with correlation δ. Then, the domain of one of the components is restricted to be greater than its mean, and this component in marginalised out. This representation has been already utilized, particularly in the context of psychometric tests [e.g., 4, 10].

*Corresponding author. Email: antonio.canale@unito.it

The above probabilistic construction is one of the many stochastic representations of the so called skew-normal model introduced by Azzalini in [5]. A univariate skew-normal random variable, say $Y \sim SN(\xi, \omega, \alpha)$, has probability density function

$$f(y; \xi, \omega, \alpha) = \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left(\alpha\frac{y-\xi}{\omega}\right), \quad y \in \mathrm{R}, \tag{1}$$

where ξ, ω and $\alpha = \delta/\sqrt{1-\delta^2}$ are location, scale and shape parameters, respectively, δ is the latent correlation of the original bivariate normal, and ϕ and Φ are, respectively, the probability density function and the cumulative distribution function of a Gaussian distribution. Clearly, if $\alpha = 0$, we are back at the Gaussian distribution. The skew-normal class of models has been widely generalised and extended by many authors such as [6, 7, 9, 11, 17] among others. A commendable work of unification of some of the proposals is made by [1], in which the unified skew-normal (SUN) class of distribution is introduced.

Our motivating data refers to first-year undergraduate students for the program in Economics at the University of Padua (Italy). We want to model the distribution of these students' grades in the first class of Statistics, one of the main mandatory first-year courses. In order to be admitted to Economics, students are required to pass a preliminary placement test. In this situation, representation (1) may be adopted. While estimating the distribution of students' grades, we can also estimate, as a byproduct, the latent correlation between the unobserved placement test score and the grades in Statistics, which can be used to evaluate the accuracy of the placement test to predict the students' performance. We likely expect that this correlation is positive, and this is one of the extra information we have. In this example, we also know the grades of previous years. Thus, we may want to use the information that the distribution of Statistics grades is skewed to the right or has a mean around a given value. The Bayesian approach of inference easily allows us to include prior information in our analysis; within this framework, we propose an informative prior for the skew-normal shape parameter.

As discussed in the recent monograph by Azzalini and Capitanio [8], the maximum likelihood estimation of α poses some intrinsic problems. For example, in specific cases, the likelihood function does not have a maximum in the interior of the parameter space. The Bayesian approach has been shown to overcome these problems and some objective Bayesian procedures have been proposed to estimate α [see e.g. 13, 21]. However, in many circumstances, as in our case, prior information is available. With a subjective perspective, in [2] the authors propose a conjugate prior, given skew-normal likelihood with fixed location and scale parameters. Despite the closed form of the posterior distribution, the authors state that their class of distributions is not closed under sampling and they do not discuss tools for posterior computation. We address this problem by discussing an informative prior for the shape parameter of the skew-normal distribution, leading to closed-form full-conditional posterior distribution.

In the next section, we discuss a prior for α, assuming ξ, ω to be fixed and focusing on the univariate model (1). Prior elicitation and an extension to the multivariate case are also extensively discussed. In Section 3 we exploit one of the possible stochastic representations of the skew-normal model and discuss an easy sampling method, particularly useful in Markov Chain Monte Carlo (MCMC) approximation of the posterior. The results are then extended to the case in which we assign an independent normal inverse-gamma prior to ξ and ω. Section 4 compares the results of our prior with Jeffreys' non informative prior in a simulation experiment. In Section 5, we analyse the data on grades in the first-year examination of Statistics by undergraduate students of the School of Economics, University of Padua, Italy, in 2003. The paper ends with a final

discussion.

2. Likelihood and prior specifications

Let us first assume that ξ, ω are known and, without loss of generality, that $\xi = 0$ and $\omega = 1$. The likelihood of model (1) for an iid sample $y = (y_1, \ldots, y_n)$ of size n is

$$L(\alpha) = \prod_{i=1}^{n} 2\phi(y_i)\Phi(\alpha y_i).$$

We assume *a priori* that the parameter α is skew-normal distributed, i.e.,

$$\alpha \sim \pi(\alpha), \;\; \pi(\alpha) = \frac{2}{\psi_0}\phi\left(\frac{\alpha-\alpha_0}{\psi_0}\right)\Phi\left(\lambda_0\frac{\alpha-\alpha_0}{\psi_0}\right), \tag{2}$$

where α_0 and ψ_0 are location and scale hyperparameters, respectively and λ_0 is a shape hyperparameter reflecting our beliefs on the direction of skewness. The posterior distribution for α is

$$\begin{aligned}\pi(\alpha; y) &\propto \phi\left(\frac{\alpha-\alpha_0}{\psi_0}\right)\Phi\left(\lambda_0\frac{\alpha-\alpha_0}{\psi_0}\right)\prod_{i=1}^{n}\Phi(\alpha y_i) \\ &\propto \phi\left(\frac{\alpha-\alpha_0}{\psi_0}\right)\Phi_{n+1}\left(\begin{bmatrix} y\alpha_0 \\ 0\end{bmatrix} + \begin{bmatrix} y \\ \lambda_0/\psi_0\end{bmatrix}(\alpha-\alpha_0); I_{n+1}\right).\end{aligned} \tag{3}$$

The above equation, once normalised, belongs to the SUN class of distributions discussed in [1] and, more precisely,

$$\alpha|y \sim SUN_{1,n+1}(\alpha_0, \gamma, \psi_0, 1, \Delta, \Gamma) \tag{4}$$

where $\Delta = [\delta_i]_{i=1,\ldots,n+1}$ is the vector of size $n+1$ containing $\delta_i = \psi_0 z_i(\psi_0^2 z_i^2 + 1)^{-1/2}$ with $z = (\psi_0 y^T, \lambda_0)^T$, $\gamma = (\Delta_{1:n}\alpha_0\psi_0^{-1}, 0)$, is the vector of size $n+1$ containing the first n entries of Δ and a zero, and $\Gamma = I - D(\Delta)^2 + \Delta\Delta^T$, where $D(V)$ is a diagonal matrix, the elements of which coincide with those of vector V. Algebraic details on how to obtain such quantities are given in the Appendix. The posterior mean and variance may be obtained from the cumulant generating function expression presented in [1]. Two interesting practical cases are obtained by considering $\lambda_0 = 0$ and $\alpha_0 = 0$.

When $\lambda_0 = 0$ the prior distribution is normal and its parameters may be chosen to center the prior on a particular guess for α. In this case, easy algebra leads to

$$\begin{aligned}\mathrm{E}[\alpha; y] &= \alpha_0 + \zeta_1(\alpha_0/\psi_0 1_n; \tilde{\Gamma}) \\ \mathrm{Var}[\alpha; y] &= \psi_0^2 + \zeta_2(\alpha_0/\psi_0 1_n; \tilde{\Gamma}),\end{aligned}$$

where 1_n is a $n \times 1$ vector of ones, $\zeta_k(x; \Sigma)$ is the kth derivative of $\log(2\Phi_n(x; \Sigma))$ with $x \in \mathrm{R}^n$, and the matrix $\tilde{\Gamma}$ is a positive semidefinite matrix with $1/\delta_i^2$ on the diagonal and 1 in all off-diagonal elements obtained as $\tilde{\Gamma} = D(\Delta)^{-1}\Gamma D(\Delta)^{-1}$.

Having $\alpha_0 = 0$, is equivalent to have rough prior information only on the skewness side of the distribution of the data. Indeed, assuming positive or negative values for the shape hyperparameter λ_0, it puts more prior mass on the positive or negative semi-axis,

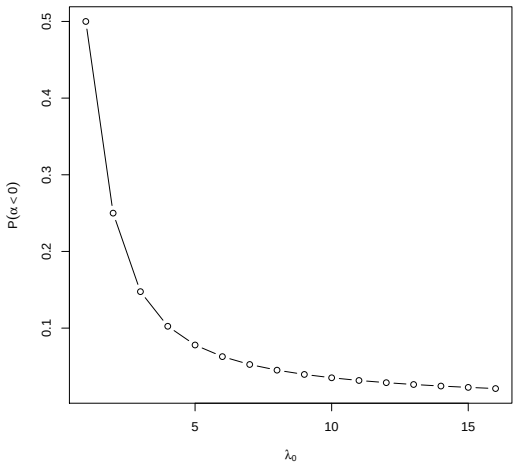


Figure 1. Probabilities mass of the occurrence of negative values of α for different choices of λ_0, with $\alpha_0 = 0$ and $\psi_0 = 1$.

respectively. The posterior mean and variance in the this case turn out to be

$$\begin{aligned} \mathrm{E}[\alpha; y] &= \zeta_1(0_n; \tilde{\Gamma}), \\ \mathrm{Var}[\alpha; y] &= \psi_0^2 + \zeta_2(0_n; \tilde{\Gamma}), \end{aligned}$$

where 0_n is a $n \times 1$ vector of zeros.

In both cases, the explicit expressions for the mean and variance of the posterior distribution are tedious to calculate since they involve the calculation of $\Phi_n(x; \Gamma)$, an n-dimensional integral numerically unstable even for moderate n. However the above expressions have a nice interpretation, as in both cases, posterior mean and variance are the sum of the prior expectation and variance and a data-driven quantity.

2.1 *Prior elicitation*

As we already introduced in Section 1, prior information are typically available when analysing university grades. Thus, it is of substantial interest to discuss the elicitation of the prior's hyperparameters. For example, the sign of the skewness of the grades distribution is expected to be positive before analysing data, and mild to moderate knowledge on it can be easily incorporated by using (2) centred in zero, that is with $\alpha_0 = 0$. Clearly, a positive value of λ_0 leads to a skew prior assigning low probability mass to negative values of α. To quantify the impact of choosing λ_0 in hypothesizing the direction of skewness in this context, we plot in Figure 1 the prior probability of negative α, $\Pr(\alpha \leq 0)$, for different choices of positive λ_0. It is evident that a very low prior mass (less than 0.05) is assumed when $\lambda_0 \geq 7$. At the same time, the choice of ψ_0 affects the concentration of mass around zero or on the chosen half real line. For example, a large ψ_0 jointly with a high positive λ_0 corresponds to a prior belief of positive skewness but mild knowledge on the actual values of α.

Often stronger prior beliefs are available on the moments of the data generating distribution. Known relations between the parameters of the model and the first four moments allows one to incorporate these prior beliefs into the model. Azzalini [5] showed that, con-

ditional on the parameters,

$$\begin{aligned}
&\mathrm{E}[Y] = \xi - \omega b\delta,\\
&\mathrm{Var}[Y] = \omega^2\{1-(b\delta)^2\},\\
&\gamma_1[Y] = ((4-\boldsymbol{\pi})/2)\mathrm{sign}(\alpha)\left[\{\mathrm{E}[Y]\}^2/\mathrm{Var}[Y]\right]^{3/2},\\
&\gamma_2[Y] = 2(\boldsymbol{\pi}-3)\left[\{\mathrm{E}[Y]\}^2/\mathrm{Var}[Y]\right]^{2},
\end{aligned}$$

where b is equal to $\sqrt{2/\boldsymbol{\pi}}$, with $\boldsymbol{\pi}$ being the mathematical constant, and γ_1 and γ_2 are the third and the fourth standardised cumulants, representing the skewness and the kurtosis of the distribution, respectively; from these expressions, given the first four standardised cumulants, a single α can be obtained. Thus, one can elicit prior hyperparameters so that the expected skewness of the data matches the prior belief. The uncertainty about α varies according to the prior variance ψ_0 which can be large or small for high and low uncertainty respectively.

2.2 *Multivariate extension*

Let us suppose that the interest lies in estimating the joint distribution of all first-year grades. To this end a natural extension of model (1) is the multivariate skew-normal [see, e.g., 8]. A d-variate skew-normal random variable $Y \sim SN_d(\xi, \Omega, \alpha)$ has probability density function

$$f(\mathbf{y};\boldsymbol{\xi},\Omega,\boldsymbol{\alpha}) = 2\phi_d(\mathbf{y}-\boldsymbol{\xi};\Omega)\Phi(\boldsymbol{\alpha}^T\omega^{-1}(\mathbf{y}-\boldsymbol{\xi})), \tag{5}$$

where $\boldsymbol{\xi}$ is a d-dimensional location parameter, Ω is a $d\times d$ positive semidefinite symmetric matrix with diagonal elements $\omega_1^2,\ldots,\omega_d^2$, $\omega = \mathrm{diag}(\Omega)$, where $\mathrm{diag}(A)$ is the diagonal matrix with the elements of the diagonal of A, and $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_d)^T$ is a d-dimensional shape vector.

The generalisation of our approach in the multivariate context is straightforward. Consider the multivariate likelihood arising from an iid sample $\mathbf{y} = (\mathbf{y}_1^T,\ldots,\mathbf{y}_n^T)^T$ of size n from the d-variate skew-normal (5) with standardised marginals, correlation matrix Ω, and vector of means zero, namely

$$L(\boldsymbol{\alpha}) = \prod_{i=1}^{n} 2\phi_d(\mathbf{y}_i;\Omega)\Phi(\boldsymbol{\alpha}^T\mathbf{y}_i).$$

If we assume that the marginal distributions of the components of $\boldsymbol{\alpha}$ are all skew

normal as in (2), the posterior distribution is

$$\begin{aligned}\pi(\boldsymbol{\alpha};\mathbf{y}) &\propto \phi_d\Big(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0;\Psi\Big)\Phi\left(\boldsymbol{\lambda}D(\psi)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)\right)\prod_{i=1}^{n}\Phi(\boldsymbol{\alpha}^T\mathbf{y}_i)\\ &\propto \phi_d\Big(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0;\Psi\Big)\Phi\Big(\boldsymbol{\lambda}D(\psi)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)\Big)\Phi_n\left(\begin{bmatrix}\boldsymbol{\alpha}^T\mathbf{y}_1\\ \vdots\\ \boldsymbol{\alpha}^T\mathbf{y}_n\end{bmatrix};I_n\right)\\ &\propto \phi_d\Big(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0;\Psi\Big)\Phi_{n+1}\left(\begin{bmatrix}\mathbf{y}^T\boldsymbol{\alpha}\\ \boldsymbol{\lambda}D(\boldsymbol{\psi})^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)\end{bmatrix};I_{n+1}\right). \qquad (6)\end{aligned}$$

3. Posterior computation

In this section, we first introduce an efficient algorithm to simulate the full conditional of α, given ξ and ω, then we specify prior distributions for the location and scale of the skew-normal and propose an efficient Gibbs sampler for the joint vector of parameters.

3.1 *A stochastic representation*

In the following Lemma 3.1 we recall a result on a stochastic representation of the SUN family, introduced in Section 2.1 of [1]. We then exploit this result and use it as an efficient simulation algorithm for drawing observations from posterior distributions (4) or (6).

LEMMA 3.1 *(Arellano-Valle and Azzalini, 2006) Let* $V_0 \sim LTN_q(-\gamma;0,\Gamma)$, $V_1 \sim N(0,\Omega)$ *with* V_0 *independent of* V_1 *and the notation* $LTN_d(\tau;\mu,\Sigma)$ *denotes a d-variate normal distribution with mean* μ *and variance-covariance matrix* Σ *truncated at* τ *from below. If*

$$Y = \xi + \omega(\Delta\Gamma^{-1}V_0 + \sqrt{1-\Delta^T\Gamma^{-1}\Delta}V_1),$$

then $Y \sim SUN_{1,q}(\xi,\gamma,\omega,1,\Delta,\Gamma)$.

Simulations from the model above can be easily done relying on efficient sampling algorithms for multivariate truncated Gaussian distribution. Recent results in this direction involves slice sampler [20] or Hamiltonian Monte Carlo [23] algorithms. From our experience, the slice sampling algorithm is faster than the Hamiltonian Monte Carlo approach and thus we use the former approach henceforth. The bottleneck of the convolution of Lemma 3.1 is represented by the inverse of the $n \times n$ matrix Γ. To perform a general matrix inversion, it is well-known that $O(n^3)$ operations are required. However, given the particular expression for Γ, a closed form for its inverse is available. Using the Sherman-Morrison formula [e.g., 19, p. 50], we can write

$$\begin{aligned}\Gamma^{-1} &= \left(I - D(\Delta)^2 + \Delta\Delta^T\right)^{-1}\\ &= D(\Delta') - \frac{1}{1+\sum_{i=1}^n \delta_i^2(1-\delta_i^2)^{-1}}D(\Delta')\Delta\Delta^T D(\Delta')\\ &= D(\Delta') - \frac{1}{1+\sum_{i=1}^n \delta_i^2(1-\delta_i^2)^{-1}}\tilde{\Delta},\end{aligned}$$

where Δ' is the vector of size n with entries $(1-\delta_i^2)^{-1}$ and $\tilde{\Delta}$ is an $n \times n$ matrix with elements $\tilde{\delta}_{ij} = \delta_i\delta_j(1-\delta_i^2)^{-1}(1-\delta_j^2)^{-1}$, for $i, j = 1, \ldots, n$. Note that this expression is not valid in general for the SUN model but it is a consequence of the prior specification. This is the first time (to our knowledge) that such expression is discussed.

A particular case of Lemma 3.1 refers to skew-normal distribution. In this case we can simulate a skew-normal random variable $X \sim SN(\xi, \omega, \lambda)$ with its hierarchical representation in which, conditionally on X_0, a realisation from a half normal distribution, X is normal with mean $\xi + \omega\delta X_0$ and variance $(1-\delta^2)\omega^2$.

3.2 *An efficient Gibbs sampler for the whole parameter vector*

For inference on the complete vector of the parameters, we specify an independent normal inverse gamma distribution for the location and scale parameter and the prior distribution (2) for the shape parameter. Specifically, we let the prior distribution for the whole vector of the parameters of model (1) be

$$\pi(\xi, \omega, \alpha) = N(\xi; \xi_0, \kappa\omega^2) \times \text{I-Ga}(\omega^2; a, b) \times \pi(\alpha), \tag{7}$$

where I-Ga$(\cdot; a, b)$ denotes the inverse gamma distribution with mean $b/(a-1)$ and variance $b^2/\{(a-1)^2(a-2)\}$, π is the prior (2), with suitable hyperparameter vector.

A particular case of Lemma 3.1 suggests us to introduce independent standard normal latent variables $\eta_1, \ldots, \eta_n$. Conditionally on such latent variables, we can consider the generic i-th observation as being normally distributed with mean $\xi + \omega\delta|\eta_i|$ and variance $(1-\delta^2)\omega^2$. Thanks to this interpretation we gain conjugacy for the location and scale parameters. This last argument allows us to build an efficient Gibbs sampling algorithm which iterates through the following steps:

- Update η_i from its full conditional posterior distribution

$$\eta_i \sim TN_0(\delta(y_i - \xi), \omega^2(1-\delta^2))$$

  where δ is $\alpha/\sqrt{\alpha^2+1}$ and $TN_\tau(\mu, \sigma^2)$ is a mean μ variance σ^2 normal truncated below τ.
- Sample (ξ, ω) from

$$N\left(\hat{\mu}, \hat{\kappa}\omega^2\right) \text{I-Ga}(a + (n+1)/2, b + \hat{b})$$

  where

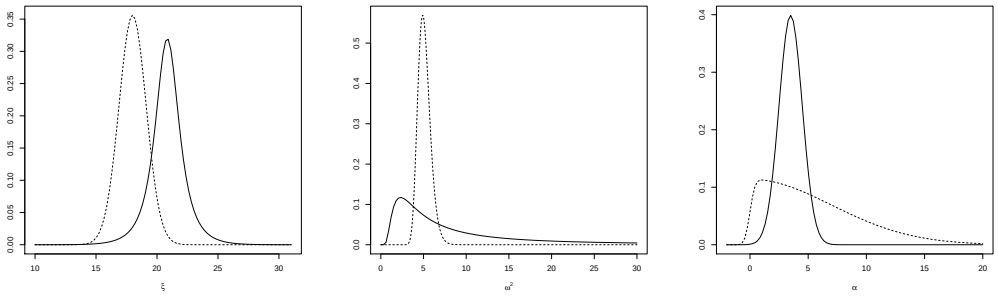
$$\hat{\mu} = \frac{\kappa\sum_{i=1}^n (y_i - \delta\eta_i) + (1-\delta^2)\xi_0}{n\kappa + (1-\delta^2)},$$

$$\hat{\kappa} = \frac{\kappa(1-\delta^2)}{n\kappa + (1-\delta^2)},$$

$$\hat{b} = \frac{1}{2(1-\delta^2)}\left\{\delta^2\sum_{i=1}^n \eta_i^2 - 2\delta\sum_{i=1}^n \eta_i(y_i - \xi) + \sum_{i=1}^n (y_i - \xi)^2 + \frac{1-\delta^2}{\kappa}(\xi - \xi_0)^2\right\}$$

- Sample α from

$$\alpha \sim \pi(\alpha|y^*)$$



(a) (b) (c)

Figure 2. Marginal prior distribution with first (dashed line) and second (continuous line) prior elicitation for ξ (a), ω^2 (b) and α (c). First elicitation: $\alpha_0 = 0$, $\psi_0 = 7$, $\lambda_0 = 20$. Second elicitation: prior means of parameters matching sample quantities calculated on a different dataset.

where $y_i^* = (y_i - \xi)/\omega$ for $i = 1, \ldots, n$, and $\pi(\alpha|y)$, is the posterior (4).

4. Simulation

To assess the performance of the proposed model, we analyzed simulated data in which the true values of the parameters were known. The data were chosen to have behavior similar to that of the real dataset analyzed in Section 5. More precisely, we simulate a sample of size $n = 50$ from a $SN(22, 3, 5)$. For three different choices of prior information, we run our proposed Gibbs sampler and, after a burn-in of 2000 iterations, we collect 10,000 MCMC samples.

Given the selection mechanism described in the introduction, we expect a positive correlation between the results of placement tests and the Statistics examinations, and thus expect skewness to the right. Hence, as a first analysis we choose as prior a skew-normal distribution with location parameter $\alpha_0 = 0$, scale parameter $\psi_0 = 7$, and shape parameter $\lambda_0 = 20$. We expect that the average grade for the examination will around 20–21. With the already mentioned relations between the central moments and the parameters of the skew-normal distribution, this information can be described by a normal-inverse-gamma prior for the skew-normal location and scale parameters with hyperparameters $\xi_0 = 21$, $\kappa = 0.25$, $a = 50$, and $b = 250$. With these choices, we assign a prior probability of about 95% for values of the location parameter between 19 and 24 and about 90% to variance between 3 and 6.

As a second analysis, we mimic the real data situation in which data on the previous year's examinations are known. Hence, we generate a different random sample of the same size and from the same distribution as the original sample, by presuming that it describes a previous year's examination results and compute the three first central moments of such a sample. As discussed in Section 2.1 we elicit prior hyperparameters in order to match our expectations to the previous year sample quantities. Figure 2 shows the marginal priors for the three parameters.

Convergence and mixing are diagnosed by monitoring the traceplots of the three parameters; mixing is adequate in each case and the Geweke's diagnostics suggest very rapid convergence.

To compare our results with a non-informative approach within the Bayesian framework, we use Jeffreys' prior for the parameters by setting the prior probability of (ξ, ω) as proportional to $1/\omega$ and using the prior obtained by Liseo and Loperfido [21] for the shape parameters. As pointed out by the above authors, this prior for the location and scale parameters given α is the conditional reference prior. To compute posterior summaries,

Table 1. Posterior means and 95% credible intervals for the simulated sample. First elicitation: $\alpha_0 = 0$, $\psi_0 = 7$, $\lambda_0 = 20$. Second elicitation: prior means of parameters matching sample quantities calculated on an independent sample (see Section 2.1).

Prior	ξ	ω	α
First elicitation	22.059 (21.727, 22.440)	2.249 (1.965, 2.589)	5.131 (1.936, 12.389)
Second elicitation	22.106 (21.765, 22.447)	2.465 (1.917, 3.195)	3.329 (1.976, 4.901)
Jeffreys prior	21.912 (21.450, 22.766)	2.588 (1.869, 3.359)	25.694 (1.318, 168.063)

we implement a blocked Gibbs sampler with sub-steps composed of Metropolis-Hastings. In this case, the burn-in is longer than for the informative proposals, and we discard the first $5,000$ iterations but still collect $10,000$ MCMC samples. Convergence and mixing are diagnosed by monitoring the traceplots of the three parameters.

Table 1 reports the posterior means and 95% credible intervals of the parameters. Substantial improvements relative to the non informative approach can be appreciated both in terms of point estimate and in terms of credible intervals. Under the non-informative prior credible intervals are wider than the relative intervals with our informative priors.

5. Density estimation of university grades

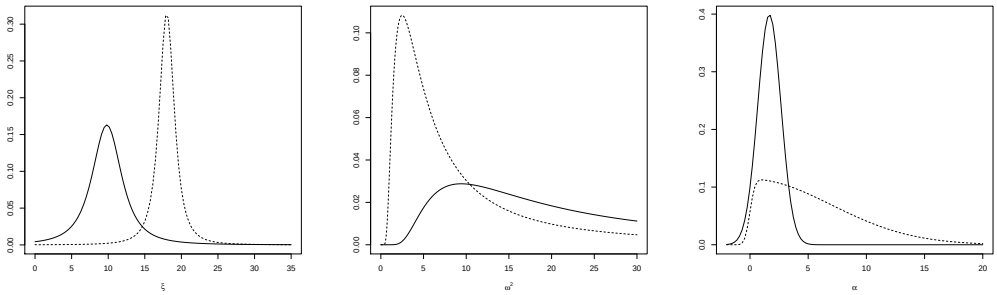
We analyzed data on grades on first-year undergraduate students in Economics at the University of Padua (Italy). Available data refer to 79 students who took the examination for the basic exam of "Statistics" at the first session of July 2003. Of these students, 54 also took the "Business organisation" exam at the same session. In order to be enrolled in the program in Economics, students need to pass a placement test at the beginning of the academic year, which give to each of them a numerical evaluation. The test consists in simple questions of basic math, logic and problem solving. The distribution of the scores of the placement test is typically symmetric, with a small number of very good and very low scores, so that it can be supposed to be Gaussian. We also assume that the grades of the Statistics exam, without any selection mechanism, are normally distributed. Thus, the skew distribution for the results of these exams is a reasonable assumption.

5.1 *Univariate skew-normal*

We first perform a univariate analysis considering only the results of the "Statistics" exam only.

We start by assuming that the correlation between the results of the placement test and the Statistics examination is positive. Thus, we choose $\alpha_0 = 0$, $\psi_0 = 7$, and $\lambda_0 = 20$, which is equivalent to putting less than 0.02 prior mass below zero, i.e. we strongly believe that α parameter is positive. This choice leads to a prior expectation for α of 5.58. We choose the hyperparameters for normal-inverse gamma $\xi_0 = 18$, $\kappa = 0.01$, $a = 1$, and $b = 5$, which lead to an expectation for ω of 1.58. These choices for prior parameters correspond to assuming that *a priori* data have first, second and third central standardised moments of 19.24, 0.98, and 0.88, respectively.

As a second analysis, we also consider the data on the past year examination, so that we can use a normally distributed prior putting $\lambda_0 = 0$. We elicit prior hyperparameters in order to match prior expectations to the previous year sample quantities. The sample mean, variance and skewness of the past year's examinations are 22.68, 13.72 and 0.35, respectively, which correspond to location, scale and shape parameters of 9.81, 18.82 and



(a) (b) (c)

Figure 3. Marginal prior distribution with first (dashed line) and second (continuous line) prior elicitation for ξ (a), ω^2 (b) and α (c). First elicitation: $\alpha_0 = 0$, $\psi_0 = 7$, $\lambda_0 = 20$. Second elicitation: prior means of parameters matching sample quantities calculated on previous year sample.

1.67, respectively. We then center our prior to have means matching those quantities, i.e. $\xi_0 = 9.81$, $\kappa = 0.25$, $a = 1$, $b = 18.82$, $\psi_0 = 1$, and $\alpha_0 = 1.67$.

The resulting prior distributions are somehow different. In Figure 3, the marginal priors for the three parameters are plotted for both the distributions. The third panel of the figure shows that the marginal prior for α is more concentrated around its mode assuming the second elicitation of the prior rather than the first one. The first two panels show that the two inverse-gamma distributions are centred on very different values, leading to marginal priors for the location parameters, that is a three-parameter t distribution, with different prior variability.

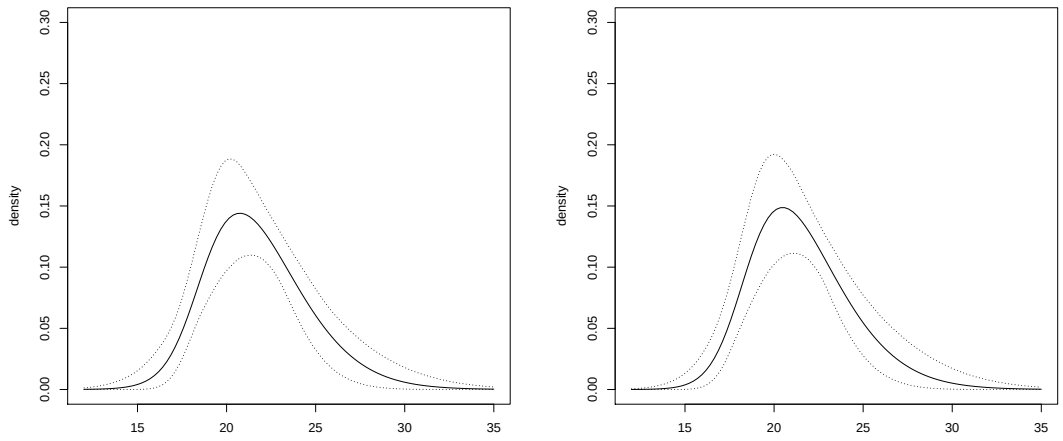
We run our Gibbs sampler for 12,000 iterations, discarding the first 2,000 as burn-in in both cases. The parameters values are monitored to gauge rates of apparent convergence and mixing. The traceplots of the parameters show excellent mixing and rapid convergence. Results are shown in Table 2 and Figure 4.

Both the final posterior densities have modes around 21 and similar variability and skewness although the prior for the parameters were different. The posterior distribution obtained from the first prior elicitation has slightly larger posterior variability than that obtained using the second elicitation, as shown by the width of the credible intervals for the parameters. The use of the previous year's data to elicit hyperparameters is clearly more informative than simply assume positive skewness.

Given the already discussed selection mechanism it is of interest to discuss the relation between the placement test and the grade in Statistics. Using the relation $\delta = \alpha/\sqrt{\alpha^2 + 1}$, we can produce point and interval estimates of the correlation coefficient δ between the placement test and the grade in Statistics, a quantity that is more interpretable. This can be simply achieved applying the transformation to the MCMC sample of α, which behaves as a sample from the conditional posterior of δ. The posterior mean of δ under the first elicitation is 0.91 with 95% credible interval equal to (0.57, 0.98) and, under the second elicitation is 0.92 with 95% credible interval equal to (0.77, 0.97). In both cases, the correlation is estimated positive and very high, with again the second, more informative, elicitation leading to a narrower credible interval. This high value suggests that the placement test is a good way to select the best students, at least for what concerns their performance in the Statistics exam.

Table 2. Posterior means and 95% credible bands for university grade dataset: First elicitation: $\alpha_0 = 0$, $\psi_0 = 7$, $\lambda_0 = 20$. Second elicitation: prior means of parameters matching sample quantities calculated on previous year sample (see Section 2.1).

Prior	ξ	ω	α
First elicitation	18.817 (17.886, 20.229)	4.163 (3.094, 5.767)	2.361 (0.693, 4.556)
Second elicitation	18.495 (17.688, 19.426)	4.176 (3.125, 5.728)	2.508 (1.224, 4.042)



(a) (b)

Figure 4. Posterior mean density (black lines) and 95% credible bands (dotted lines) for the first (a) and second (b) specification of the prior.

5.2 *Multivariate skew-normal*

We now perform a bivariate analysis considering jointly the results of the "Statistics" and "Business Organization" exams. The set of students that passed both "Statistics" and "Business Organization" is a subset of that analysed in Section 5.1 of dimension $n = 54$. Here, for the marginal prior specification of both variables, we elicit hyperparameters in order to match prior expectations to the previous year sample quantities. For "Statistics" the choices reflect the ones of Section 5.1, while for "Business Organization" the sample mean, variance and skewness of the past year's examinations are 23.64, 10.35, and -0.023 respectively, which correspond to location, scale and shape parameters of 27.52, 11.05, and -0.49, respectively. Note that the skewness for the grade of "Business Organization", in the previous year, is very close to zero, suggesting an absence of correlation of this outcome with the preliminary placement test. Therefore, our prior specification for model (5) consists in

$$\begin{aligned}
&\boldsymbol{\xi} \sim N\left(\begin{pmatrix} 9.81 \\ 27.52 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right), \\
&\Omega = \omega^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \ \rho \sim U(-1, 1), \ \omega^2 \sim \text{I-Ga}(\omega^2; 1, 15) \\
&\boldsymbol{\alpha} \sim N\left(\begin{pmatrix} 1.67 \\ -0.49 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right),
\end{aligned}$$

where a noninformative uniform prior density is assumed for the correlation between the results of the two exams.

We calculate the posterior distribution via straightforward MCMC algorithm. We run the algorithm for 12,000 iterations, discarding the first 2,000 as burn-in. The parameters values are monitored to gauge rates of apparent convergence and mixing. The traceplots of the parameters show good mixing and fast convergence. The posterior means of the parameters (and their 95% credible intervals) are 20.17 (19.01, 21.52), 26.68 (25.15, 27.75), 1.48 (.53, 2.77), -1.56 (-2.71, -0.52), -0.04 (-0.36, 0.31), 13.76 (10.24, 18.16) for ξ_1, ξ_2, α_1, α_2, ρ, and ω^2, respectively. Using the posterior mean parameters we obtain the contours plot reported in Figure 5. The two exams have different mean evaluation (with "Statistics" having a lower mark) with non crossing credible intervals. The posterior mean of the skewness parameter for "Statistics" is positive, yet lower than that obtained in Section 5.1. This is not inconsistent since we are using here almost half of the data used in the previous section. An interesting evidence is related to the second skewness parameter, which posterior mean is negative with the 95% credible intervals not containing zero. This suggests that the results of "Business Organization" is likely to be negatively correlated to the preliminary placement test. The reason may lies in the fact that the mathematical and logical skills tested in the preliminary placement test, are related to "Statistics" but not strongly related to the content of "Business Organization". This is also evident from the posterior mean of ρ, the correlation parameter between "Statistics" and "Business Organization", which credible interval is almost symmetric around zero.

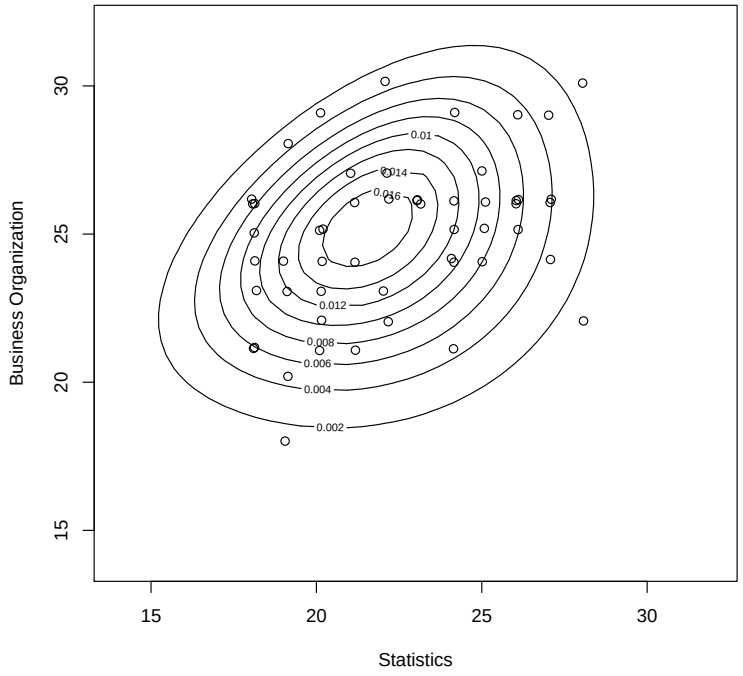


Figure 5. Contour plot of the estimated bivariate skew-normal density with paramters the estimated posterior mean parameters. Dots represent the original data points.

6. Discussion

The selection mechanisms of many university courses, may lead to have skew distributions of university grades. When grades are numerically evaluated, the skew-normal seems an appropriate model to describe the distribution of the grades. Since university placement tests are positively correlated with the students' results, the expected skewness is also expected to be positive. Given the above, an informative prior distribution for the shape parameter of skew-normal distribution has been discussed. The induced posterior is in closed form and belongs to the SUN family of distributions. We described an efficient, easy, and reliable sampling algorithm related to a stochastic representation of the skew-normal model which uses recent advances in sampling from multivariate truncated Gaussian distribution. A Gibbs sampling algorithm for the joint vector of the parameters has also been introduced. The application to first-year undergraduate students grades in Statistics of the Economic program of the University of Padua, shows that, as expected, the more information are embedded into the prior distribution, the more precise the final estimates are. This suggests to use informative priors when prior information are available.

A possible extension consists in assuming that the distribution of the grades is an extended skew-normal (ESN), a model firstly introduced in the pioneering paper of Azzalini [5] and subsequently extensively studied in [3, 14, 15]. The latter generalises the skew-normal in assuming that the latent normal component is restricted to be greater than a general value and not necessarily to its mean. This generalisation can be successfully applied in the contexts at hand when the selection threshold is not constrained to be equal to the mean of the distribution of the placement test.

Acknowledgements

The authors thank Eric Battistin for generously providing the data. This research was partially supported by the University of Padua CPDA121180/12 grant.

References

[1] R.B. Arellano-Valle and A. Azzalini, *On the unification of families of skew-normal distributions*, Scandinavian Journal of Statistics 33 (2006), pp. 561–574.
[2] R.B. Arellano-Valle, M.G. Genton, and R.H. Loschi, *Shape mixture of multivariate skew-normal distributions*, Journal of Multivariate Analysis 100 (2009), pp. 91–101.
[3] B.C. Arnold and R.J. Beaver, *Hidden truncation model*, Sankhyā, series A 62 (2000), pp. 22–35.
[4] B.C. Arnold, R.J. Beaver, R.A. Groeneveld, and W.Q. Meeker, *The non truncated marginal of a truncated bivariate normal distribution*, Psychometrika 58 (1993), pp. 471–488.
[5] A. Azzalini, *A class of distributions which includes the normal ones*, Scandinavian Journal of Statistics 12 (1985), pp. 171–178.
[6] A. Azzalini and A. Capitanio, *Statistical applications of the multivariate skew-normal distribution*, Journal of the Royal Statistical Society series B 61 (1999), pp. 579–602.
[7] A. Azzalini and A. Capitanio, *Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution*, Journal of the Royal Statistical Society series B 65 (2003), pp. 367–389.
[8] A. Azzalini and A. Capitanio, *The Skew-Normal and Related Families*, Cambridge University Press, New York, 2014.
[9] A. Azzalini and A. Dalla Valle, *The multivariate skew-normal distribution*, Biometrika 83 (1996), pp. 715–726.
[10] Z.W. Birnbaum, *Effect of linear truncation on a multinormal population*, The Annals of Mathematical Statistics 21 (1950), pp. 272–279.

[11] M. Branco and D. Dey, *A general class of multivariate skew-elliptical distributions*, Journal of Multivariate Analysis 79 (2001), pp. 93–113.
[12] B. Bridgeman and C. Lewis, *The relationship of essay and multiple-choice scores with grades in college courses*, Journal of Educational Measurement 31 (1994), pp. 37–50.
[13] S. Cabras, W. Racugno, M. Castellanos, and L. Ventura, *A matching prior for the shape parameter of the skew-normal distribution*, Scandinavian Journal of Statistics 39 (2012), pp. 236–247.
[14] A. Canale, *Statistical aspects of the scalar extended skew-normal distribution*, Metron LXIX (2011), pp. 279–295.
[15] A. Capitanio, A. Azzalini, and E. Stanghellini, *Graphical models for skew-normal variates*, Scandinavian Journal of Statistics 30 (2003), pp. 129–144.
[16] S. Geiser and R. Studley, *UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California*, Educational Assessment 8 (2002), pp. 1–26.
[17] M. Genton and N. Loperfido, *Generalized skew-elliptical distributions and their quadratic forms*, Ann. Inst. Statist. Math. 57 (2005), pp. 389–401.
[18] J. Geweke, *Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments*, in *Bayesian Statistics 4*, Oxford: Oxford University Press, 1992.
[19] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, 1989.
[20] M.W. Liechty and J. Lu, *Multivariate normal slice sampling.*, Journal of Computational and Graphical Statistics 19 (2010), pp. 281–294.
[21] B. Liseo and N. Loperfido, *A note on reference priors for the scalar skew-normal distribution*, Journal of Statistical planning and inference 136 (2006), pp. 373–389.
[22] A.R. Napoli and L.A. Raymond, *How reliable are our asessment data?: A comparison of the reliability of data produced in graded and un-graded conditions*, Research in Higher Education 45 (2004), pp. 921–929.
[23] A. Pakman and L. Paninski, *Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians*, Journal of Computational and Graphical Statistics 23 (2014), pp. 518–542.

Appendix

To explain the relations between equations (3) and (4), let first introduce the density of Z, where $Z \sim SUN_{m,d}(\xi, \gamma, \omega, \Omega, \Delta, \Gamma)$, which is

$$f(Z; \xi, \gamma, \omega, \Omega, \Delta, \Gamma) = \phi_d\left(z - \xi; \omega\Omega\omega\right) \frac{\Phi_m(\gamma + \Delta\Omega^{-1}\omega^{-1}(z-\xi); \Gamma - \Delta\Omega^{-1}\Delta^T)}{\Phi_m(\gamma; \Gamma)}, \quad (8)$$

where $\Phi_d(\cdot; \Sigma)$ is the cumulative distribution function of a d-variate Gaussian distribution with variance covariance matrix Σ, Ω, Γ, and $\Omega^* = ((\Gamma, \Delta)^T, (\Delta^T, \Omega)^T)$ are correlations matrices, and ω is a $d \times d$ diagonal matrix.

In order to match (3) with the above SUN parametrization, we set

$$\xi \leftarrow \alpha_0$$
$$\omega \leftarrow \psi_0$$

Under these assumptions, with $d = 1$ and $m = n + 1$, equations (8) can be written as

$$\phi_d\left(\frac{\alpha - \alpha_0}{\psi_0}\right) \Phi_{n+1}\left(\mathrm{diag}\{\delta_i^{-1}\}\gamma + \frac{\alpha - \alpha_0}{\psi_0} 1_{n+1}; \mathrm{diag}\{\delta_i^{-2}\}(\Gamma - \Delta\Delta^T)\right) / \Phi_{n+1}(\gamma; \Gamma),$$

where 1_d is a vector of size d of ones. With similar steps, we also rewrite the $n+1$-variate

cdf of equation (3) as

$$\Phi_{n+1}\left(\begin{bmatrix} y\alpha_0 \\ 0 \end{bmatrix} + \begin{bmatrix} y \\ \lambda_0\psi_0^{-1} \end{bmatrix}(\alpha-\alpha_0); I_{n+1}\right) =$$
$$\Phi_{n+1}\left(\alpha_0\psi_0^{-1}\begin{bmatrix} 1_n \\ 0 \end{bmatrix} + \frac{\alpha-\alpha_0}{\psi_0}1_{n+1}; \mathrm{diag}\{z_i^{-2}\}\right),$$

where $z_i = y_i\psi_0$ for $i = 1,\ldots,n$, $z_{n+1} = \lambda_0$. Then, in order to obtain the parameters involved in SUN density we need to elicit Γ and Δ, so that $(\Gamma - \Delta\Delta^T)\mathrm{diag}\{\delta_i^{-2}\} = \psi_0^{-2}\mathrm{diag}\{z_i^{-2}\}$ and γ so that $\mathrm{diag}\{\delta_i^{-1}\}\gamma = \alpha_0\psi_0^{-1}(1_n^T, 0)$.

Since $\mathrm{diag}\{1/\delta_i\}\Delta\Delta^T\mathrm{diag}\{1/\delta_i\}$ is a $n+1 \times n+1$ matrix of ones, and

$$\mathrm{diag}\{1/\delta_i\}\Gamma\mathrm{diag}\{1/\delta_i\} = \left[\frac{\Gamma_{ij}}{\delta_i\delta_j}\right]_{i,j=1,\ldots,n+1},$$

we require $\gamma_{ij} = \delta_i\delta_j$ for the off-diagonal elements of Γ. Hence, recalling that Γ must be a correlation matrix, we have for each $i = 1,\ldots,n+1$, $z_i^{-2} = (1-\delta_i^2)/\delta_i^2$, which leads to define

$$\delta_i \leftarrow \frac{\psi_0 z_i}{\sqrt{\psi_0^2 z_i^2 + 1}}, \quad \Delta = [\delta_i]_{i=1,\ldots,n+1},$$
$$\Gamma \leftarrow I - D(\Delta)^2 + \Delta\Delta^T,$$
$$\gamma_i \leftarrow \delta_i\alpha_0\psi_0^{-1}, \text{for } i = 1,\ldots,n, \gamma_{n+1} \leftarrow 0,$$

where $D(\Delta)$ is again the diagonal matrix which diagonal elements coincide with those of Δ.