Biometrika (2019), **106**, 4, *pp.* 765–779
Printed in Great Britain
doi: 10.1093/biomet/asz034
Advance Access publication 25 August 2019

Conjugate Bayes for probit regression via unified skew-normal distributions

BY DANIELE DURANTE

Department of Decision Sciences, Bocconi University, Via Röntgen 1, 20136 Milan, Italy
daniele.durante@unibocconi.it

SUMMARY

Regression models for dichotomous data are ubiquitous in statistics. Besides being useful for inference on binary responses, these methods serve as building blocks in more complex formulations, such as density regression, nonparametric classification and graphical models. Within the Bayesian framework, inference proceeds by updating the priors for the coefficients, typically taken to be Gaussians, with the likelihood induced by probit or logit regressions for the responses. In this updating, the apparent absence of a tractable posterior has motivated a variety of computational methods, including Markov chain Monte Carlo routines and algorithms that approximate the posterior. Despite being implemented routinely, Markov chain Monte Carlo strategies have mixing or time-inefficiency issues in large-p and small-n studies, whereas approximate routines fail to capture the skewness typically observed in the posterior. In this article it is proved that the posterior distribution for the probit coefficients has a unified skew-normal kernel under Gaussian priors. This result allows efficient Bayesian inference for a wide class of applications, especially in large-p and small-to-moderate-n settings where state-of-the-art computational methods face notable challenges. These advances are illustrated in a genetic study, and further motivate the development of a wider class of conjugate priors for probit models, along with methods for obtaining independent and identically distributed samples from the unified skew-normal posterior.

Some key words: Bayesian inference; Binary data; Conjugacy; Probit regression; Unified skew-normal distribution.

1. INTRODUCTION

In many fields there is interest in understanding how the probability mass function of a binary response $y \in \{0; 1\}$ varies with a set of observed predictors $x = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ (e.g., Agresti, 2013). Common approaches to addressing this problem assume that y is a Bernoulli variable whose probability parameter changes with a linear combination of the predictors under a probit or logit mapping. In the first case $\mathrm{pr}(y = 1 \mid x, \beta) = \Phi(x^{\mathrm{T}}\beta)$, whereas in the second $\mathrm{pr}(y = 1 \mid x, \beta) = \{1 + \exp(-x^{\mathrm{T}}\beta)\}^{-1}$, and the goal is to perform inference on $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$.

Although frequentist inference for the above class of models is well established (e.g., Agresti, 2013), the Bayesian approach has attracted increasing interest since it allows the borrowing of information, uncertainty quantification, shrinkage and tractable inference via the posterior distribution for the regression coefficients (e.g., Agresti, 2013, § 7.2). Moreover, predictor-dependent models for binary data are useful building blocks in more complex Bayesian formulations, such as density regression (Rodriguez & Dunson, 2011), additive trees (Chipman et al., 2010), nonparametric classification (Rasmussen & Williams, 2006) and graphical models

(Spiegelhalter & Lauritzen, 1990), among others. Although these methods provide popular learning procedures, computational barriers still exist. Indeed, unlike for Bayesian regression with Gaussian data, there are no results on the availability of tractable posteriors for β in regression models for Bernoulli data, under the commonly used Gaussian priors on the coefficients (e.g., Chopin & Ridgway, 2017).

Motivated by the above issue, several computational methods have been proposed for Bayesian regression with binary response data. Popular routines involve data-augmentation strategies relying on hierarchical representations that provide conjugate full conditional distributions, within a Markov chain Monte Carlo framework (Albert & Chib, 1993; Holmes & Held, 2006; Frühwirth-Schnatter & Frühwirth, 2007; Polson et al., 2013). Although routinely implemented, these methods yield poor convergence and mixing in practice, especially for imbalanced datasets (Johndrow et al., 2019). One solution is to employ alternative strategies such as carefully tuned or adaptive Metropolis–Hastings (Haario et al., 2001; Roberts & Rosenthal, 2001) and more recent generalizations of Hamiltonian Monte Carlo, such as the no-U-turn sampler of Hoffman & Gelman (2014). Both of these approaches guarantee computational advantages in large-n and moderate-p settings, relative to data-augmentation Markov chain Monte Carlo. However, when p is large, Metropolis–Hastings has difficulties in exploring the parameter space, whereas Hamiltonian Monte Carlo tends to be expensive (Chopin & Ridgway, 2017). Laplace approximations, variational Bayes and expectation propagation can scale up the computations, but usually provide Gaussian approximations that affect the quality of inference when the posterior is skewed. This is a common situation in regression for binary data (Kuss & Rasmussen, 2005) and, as will be discussed later in this article, is a property inherent to the posterior. See Chopin & Ridgway (2017) for a thorough discussion and comparison of the aforementioned approaches with a focus on probit regression.

Although providing state-of-the-art methods in Bayesian regression for binary response data, the strategies discussed above are still suboptimal compared to situations in which the posterior belongs to a known and tractable class of random variables. Indeed, the latter case could facilitate the calculation of several quantities relevant to posterior inference, without relying on Monte Carlo methods. In this article it is proved that in the case of probit regression models which have Gaussian priors for the coefficients, the posterior belongs to the class of unified skew-normal distributions (Arellano-Valle & Azzalini, 2006). Such variables have already appeared in probit models for obtaining flexible link functions via skewed latent data, instead of Gaussian ones (e.g., Bazán et al., 2006). However, that is a different situation from the one considered in the present article.

To the best of the author's knowledge, the main result of this article has not previously been reported in the literature, and it can contribute to some important advances. In fact, as will be shown later, the unified skew-normal posterior guarantees closure properties (Arellano-Valle & Azzalini, 2006) in addition to explicit formulas for the marginal, joint and conditional posteriors, along with predictive distributions and marginal likelihoods for model selection. These quantities involve cumulative distribution functions $\Phi_n(\cdot)$ of n-variate Gaussians, and hence can be efficiently evaluated in small-to-moderate-n settings by recent minimax tilting methods (Botev, 2017). The methods of Botev (2017) are also useful for obtaining independent samples from the posterior by exploiting a representation of the unified skew-normal distribution via a linear combination of p-variate Gaussians and n-variate truncated Gaussians. These results are valid in any dimension, and provide key methodological advances that could motivate future theoretical studies and facilitate a formal understanding of the skewness typically observed in the posterior. However, the associated inference strategies require evaluation of $\Phi_n(\cdot)$ or sampling from n-variate truncated Gaussians, and therefore are of practical utility in studies with small-to-moderate n, typically of the order of

a few hundred, and any, even huge, p. This scenario is the most challenging for current Markov chain Monte Carlo routines (e.g., Chopin & Ridgway, 2017), and in fact the methods for posterior inference arising from the new unified skew-normal result of this paper can significantly improve state-of-the-art algorithms in such settings, thus filling in an important computational gap; see § 2.4 for a more detailed discussion of these aspects.

2. Posterior inference in probit regression via unified skew-normal distributions

2.1. *Unified skew-normal distribution*

Before deriving the unified skew-normal posterior induced by Gaussian priors for the β coefficients in a probit regression, let us first introduce the unified skew-normal distribution. Recalling Arellano-Valle & Azzalini (2006), this random variable unifies different generalizations of the multivariate skew-normal distribution $z \sim \text{SN}_p(\xi, \Omega, \alpha)$ (Azzalini & Dalla Valle, 1996), whose density $2\phi_p(z-\xi;\Omega)\Phi\{\alpha^{\text{T}}\omega^{-1}(z-\xi)\}$ is obtained by modifying that of a p-variate Gaussian $N_p(\xi,\Omega)$ with the cumulative distribution function of a standard normal evaluated at $\alpha^{\text{T}}\omega^{-1}(z-\xi)$, where ω is a $p\times p$ diagonal matrix containing the square roots of the diagonal elements in Ω. This strategy introduces skewness into $N_p(\xi,\Omega)$ that is controlled by $\alpha = (\alpha_1,\ldots,\alpha_p)^{\text{T}} \in \mathbb{R}^p$, with $\xi = (\xi_1,\ldots,\xi_p)^{\text{T}} \in \mathbb{R}^p$ and Ω driving location and variability, respectively (e.g., Arellano-Valle & Azzalini, 2006). Indeed, when $\alpha = 0_p$ the multivariate skew-normal distribution coincides with $N_p(\xi,\Omega)$, whereas setting $p = 1$ yields a univariate skew-normal distribution $\text{SN}(\xi,\omega^2,\alpha)$ (Azzalini, 1985).

Motivated by the success of the above formulation (e.g., Azzalini & Capitanio, 1999), several extensions have been proposed to capture further properties. Two important generalizations are obtained by adding another parameter γ to $\Phi\{\alpha^{\text{T}}\omega^{-1}(z-\xi)\}$ and by allowing the skewness-inducing mechanism to be multivariate. The first modification leads to the multivariate extended skew-normal distribution (Arnold & Beaver, 2000; Arnold et al., 2002). The second provides the closed skew-normal family (González-Farías et al., 2004; Gupta et al., 2004), which incorporates a skewness matrix $\Delta \in \mathbb{R}^{p\times n}$ and an $n\times n$ full-rank scale Γ into $\Phi_n(\cdot)$. Besides increasing flexibility, these extensions confer closure properties for marginals, conditionals and joint distributions, thus producing a general class. Arellano-Valle & Azzalini (2006) unified these generalizations within a single and tractable unified skew-normal representation, obtaining the density function

$$\phi_p(z-\xi;\Omega)\frac{\Phi_n\{\gamma+\Delta^{\text{T}}\bar{\Omega}^{-1}\omega^{-1}(z-\xi);\Gamma-\Delta^{\text{T}}\bar{\Omega}^{-1}\Delta\}}{\Phi_n(\gamma;\Gamma)} \tag{1}$$

for $z \sim \text{SUN}_{p,n}(\xi,\Omega,\Delta,\gamma,\Gamma)$. In (1), $\phi_p(z-\xi;\Omega)$ represents the density of a p-variate Gaussian with expectation $\xi = (\xi_1,\ldots,\xi_p)^{\text{T}} \in \mathbb{R}^p$ and $p\times p$ variance-covariance matrix $\Omega = \omega\bar{\Omega}\omega$, a quadratic combination of a correlation matrix $\bar{\Omega}$ and a diagonal matrix ω containing the square root of the diagonal elements of Ω. The quantities $\Phi_n\{\gamma+\Delta^{\text{T}}\bar{\Omega}^{-1}\omega^{-1}(z-\xi);\Gamma-\Delta^{\text{T}}\bar{\Omega}^{-1}\Delta\}$ and $\Phi_n(\gamma;\Gamma)$ are the cumulative distribution functions of the multivariate Gaussians $N_n(0_n,\Gamma-\Delta^{\text{T}}\bar{\Omega}^{-1}\Delta)$ and $N_n(0_n,\Gamma)$ evaluated at $\gamma+\Delta^{\text{T}}\bar{\Omega}^{-1}\omega^{-1}(z-\xi)$ and γ, respectively, with the $p\times n$ matrix Δ having the main effect on skewness. In fact, when Δ is zero, (1) coincides with the density of the $N_p(\xi,\Omega)$ distribution. The vector $\gamma \in \mathbb{R}^n$ introduces additional flexibility in departures from normality, consistent with the multivariate extended skew-normal distribution; see Arellano-Valle & Azzalini (2006) and Azzalini & Capitanio (2014, § 7.1.2) for details.

Arellano-Valle & Azzalini (2006) also added a further condition which constrains the $(n+p)\times(n+p)$ matrix Ω^*, with blocks $\Omega^*_{[11]} = \Gamma$, $\Omega^*_{[22]} = \bar{\Omega}$ and $\Omega^*_{[21]} = \Omega^{*\text{T}}_{[12]} = \Delta$, to be a full-rank

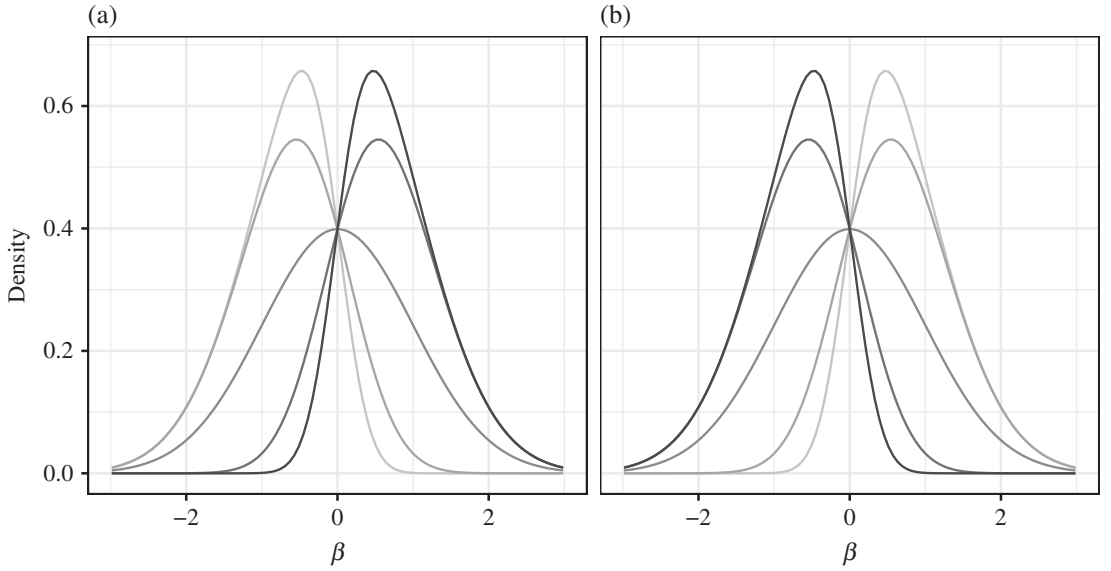

Fig. 1. Density of a $\text{SUN}_{1,1}\{0, 1, (2y-1)x(x^2+1)^{-1/2}, 0, 1\}$ posterior for β, for different combinations of x and y values: (a) $y = 1$; (b) $y = 0$. Shading ranges from light to dark grey as $x \in \{-3; -1.5; 0; 1.5; 3\}$ goes from -3 to 3.

correlation matrix. As will be clarified in § 2.2, this identifiability restriction is not required in the Bayesian setting. In fact, the parameters of the unified skew-normal posterior for the coefficients β are functions of the observed data and the prespecified hyperparameters of the Gaussian prior, thus avoiding identifiability issues. Nonetheless, this parameterization will be maintained in the rest of the article so that the classical results for the unified skew-normal distribution can be applied directly, and also to ensure identifiability of the prior when the findings for the Gaussian case are generalized to the entire class of unified skew-normal priors. In the following subsections it is proved that the posterior for the β coefficients in a probit model with Gaussian priors is a unified skew-normal distribution and the consequences of this result for posterior inference are studied.

2.2. *Unified skew-normal posterior for Bayesian probit regression with Gaussian priors*

To introduce the general case consisting of n observations from a probit model with Gaussian prior $\pi(\beta) = \phi_p(\beta - \xi; \Omega)$, first consider a simple setting with a single data point y and one covariate x, such that $(y \mid x, \beta) \sim \text{Ber}\{\Phi(x\beta)\}$ and $\beta \sim N(0, 1)$. Although this scenario is uncommon in practice, it provides key insights into the role of $x \in \mathbb{R}$ and $y \in \{0; 1\}$ in driving departures from normality in the posterior distribution. Indeed, according to Lemma 1, $(\beta \mid y, x)$ has a unified skew-normal density when $\pi(\beta) = \phi(\beta)$. All proofs are given in the Appendix.

LEMMA 1. *Let* $(y \mid x, \beta) \sim \text{Ber}\{\Phi(x\beta)\}$ *and set* $\pi(\beta) = \phi(\beta; 1) = \phi(\beta)$. *Then* $(\beta \mid y, x) \sim \text{SUN}_{1,1}\{0, 1, (2y-1)x(x^2+1)^{-1/2}, 0, 1\}$ *for every* $x \in \mathbb{R}$ *and* $y \in \{0; 1\}$.

Figure 1 displays the density function of the unified skew-normal posterior for β in the illustrative example, for different values of x and y. As expected, $(2y-1)x(x^2+1)^{-1/2}$ controls skewness. Indeed, the larger $|x|$ is, the more skewness is observed in the posterior. This skewness is either positive or negative depending on the sign of $(2y-1)x$. To clarify these results, the unified skew-normal distribution in Lemma 1 can be also re-expressed as a basic $\text{SN}\{0, 1, (2y-1)x\}$.

The above result holds more generally for independent response data $y_1, \ldots, y_n$ from a probit model $(y_i \mid x_i, \beta) \sim \text{Ber}\{\Phi(x_i^{\mathrm{T}}\beta)\}$ $(i = 1, \ldots, n)$, where $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ denotes the vector of covariates for unit i and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$ the associated coefficients. Indeed, Theorem 1 states that when β has a Gaussian prior $\beta \sim N_p(\xi, \Omega)$ with mean $\xi \in \mathbb{R}^p$ and full-rank variance-covariance matrix $\Omega = \omega\bar{\Omega}\omega$, the posterior coincides with a unified skew-normal distribution.

THEOREM 1. *If $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ comprises conditionally independent binary response data from a probit model $(y_i \mid x_i, \beta) \sim \text{Ber}\{\Phi(x_i^{\mathrm{T}}\beta)\}$ $(i = 1, \ldots, n)$ and $\beta \sim N_p(\xi, \Omega)$, then*

$$(\beta \mid y, X) \sim \text{SUN}_{p,n}(\xi_{\text{post}}, \Omega_{\text{post}}, \Delta_{\text{post}}, \gamma_{\text{post}}, \Gamma_{\text{post}}), \tag{2}$$

with posterior parameters

$$\xi_{\text{post}} = \xi, \quad \Omega_{\text{post}} = \Omega, \quad \Delta_{\text{post}} = \bar{\Omega}\omega D^{\mathrm{T}} s^{-1},$$
$$\gamma_{\text{post}} = s^{-1}D\xi, \quad \Gamma_{\text{post}} = s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}$$

for any $n \times p$ data matrix $D = \text{diag}(2y_1 - 1, \ldots, 2y_n - 1)X$ and any $n \times n$ positive diagonal matrix $s = \text{diag}\{(d_1^{\mathrm{T}}\Omega d_1 + 1)^{1/2}, \ldots, (d_n^{\mathrm{T}}\Omega d_n + 1)^{1/2}\}$; here the generic vector d_i^{T} is the ith row of D, X is the design matrix, and I_n denotes the $n \times n$ identity matrix.

Adapting (1) to the results in Theorem 1, the density function of the unified skew-normal posterior can easily be derived, after minor mathematical simplifications, as

$$\pi(\beta \mid y, X) = \phi_p(\beta - \xi; \Omega)\frac{\Phi_n(s^{-1}D\beta; s^{-1}s^{-1})}{\Phi_n\{s^{-1}D\xi; s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}\}}, \tag{3}$$

where $\Phi_n\{s^{-1}D\xi; s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}\}$ defines the normalizing constant. To clarify the role of the prior parameters ξ and Ω, as well as the effect of the data y and X, consider a constructive representation of the posterior. In particular, by adapting known results from unified skew-normal distributions (Azzalini & Capitanio, 2014, § 7.1.2) to the specific posterior in Theorem 1, it can be shown that $(\beta \mid y, X)$ has the stochastic representation in Corollary 1.

COROLLARY 1. *If $(\beta \mid y, X)$ has the unified skew-normal distribution from Theorem 1, then*

$$(\beta \mid y, X) \stackrel{\text{d}}{=} \xi + \omega\{V_0 + \bar{\Omega}\omega D^{\mathrm{T}}(D\Omega D^{\mathrm{T}} + I_n)^{-1}sV_1\} \quad (V_0 \perp V_1), \tag{4}$$

where $V_0 \sim N_p\{0_p, \bar{\Omega} - \bar{\Omega}\omega D^{\mathrm{T}}(D\Omega D^{\mathrm{T}} + I_n)^{-1}D\omega\bar{\Omega}\}$ and V_1 is from an n-variate truncated normal distribution with mean 0_n, covariance matrix $s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}$ and truncation below $-s^{-1}D\xi$.

Based on (4), ξ has a primary effect on the location, but also plays a role in controlling departures from normality since it appears in the truncation $s^{-1}D\xi$. On the other hand, the prior variance-covariance matrix Ω mainly affects scale, via ω, and posterior dependence among the β parameters, while also contributing to the weight matrix assigned to the multivariate truncated Gaussian V_1, along with its variability. Finally, the data in D play a major role in controlling departures from normality. Indeed, if D has elements close to 0, the multivariate truncated Gaussian V_1 has negligible importance compared to the multivariate Gaussian V_0 in (4).

2.3. *Inference, prediction and variable selection under unified skew-normal posteriors*

A primary focus in Bayesian regression studies is on marginal posteriors $(\beta_j \mid y, X)$ $(j = 1, \ldots, p)$, their associated moments, and more complex functionals such as measures of posterior dependence and credible intervals or regions. A fundamental property of unified skew-normal distributions, which facilitates this type of inference, is that the class of variables is closed under marginalization, linear combinations and conditioning (Arellano-Valle & Azzalini, 2006; Azzalini & Capitanio, 2014). More specifically, adapting the derivations in Arellano-Valle & Azzalini (2006) to Theorem 1, each marginal posterior is still from a unified skew-normal distribution for every β_j $(j = 1, \ldots, p)$. In particular, $(\beta_j \mid y, X) \sim \text{SUN}_{1,n}(\xi_{\text{post}j}, \Omega_{\text{post}jj}, \Delta_{\text{post}j}, \gamma_{\text{post}}, \Gamma_{\text{post}})$, where $\Delta_{\text{post}j}$ denotes the jth row of $\bar{\Omega}\omega D^{\mathrm{T}} s^{-1}$, $\xi_{\text{post}j}$ the jth element of the prior mean vector ξ and $\Omega_{\text{post}jj}$ the (j,j) entry in Ω, while γ_{post} and Γ_{post} coincide with those already defined in Theorem 1. A similar result holds for subvectors of coefficients $(\beta_{\mathcal{J}} \mid y, X)$ with $\mathcal{J} \subset \{1; \ldots; p\}$, linear combinations $(a + A^{\mathrm{T}}\beta \mid y, X)$, and conditional posteriors $(\beta_{\mathcal{J}} \mid y, X, \beta_{\mathcal{J}^*})$ with $\mathcal{J} \subset \{1; \ldots; p\}$ and $\mathcal{J}^* \subset \{1; \ldots; p\}$ such that $\mathcal{J} \cap \mathcal{J}^* = \emptyset$. Azzalini & Capitanio (2014) provide details on how to obtain the parameters of these unified skew-normal distributions from simple transformations of those in Theorem 1. Certain linear combinations of β, such as $x_i^{\mathrm{T}}\beta$, are of particular interest.

The above results facilitate graphical representation of marginal and joint posteriors, as well as calculation of posterior moments and credible intervals for the probit coefficients via one-dimensional integrals involving the marginal posterior densities. This can be done by numerical integration (e.g., Quarteroni et al., 2010) whenever it is possible to evaluate $\Phi_n(\cdot)$ with efficiency and accuracy. When interest is in posterior moments, another approach is to obtain the quantities via direct derivation of the moment generating function. Indeed, adapting the result in Arellano-Valle & Azzalini (2006) to (2), a similar strategy can be used to study functionals of the posterior, provided that $(\beta \mid y, X)$ has moment generating function

$$M(t) = \exp(\xi^{\mathrm{T}} t + 0.5 t^{\mathrm{T}} \Omega t) \frac{\Phi_n\{s^{-1}D\xi + s^{-1}D\Omega t; s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}\}}{\Phi_n\{s^{-1}D\xi; s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}\}} \quad (t \in \mathbb{R}^p). \tag{5}$$

Exploiting (5) and adapting the derivations in Azzalini & Bacchieri (2010) to the unified skew-normal distribution in Theorem 1, the posterior mean of β is

$$E(\beta \mid y, X) = \xi + \frac{1}{\Phi_n\{s^{-1}D\xi; s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}\}} \Omega D^{\mathrm{T}} s^{-1} \eta, \tag{6}$$

where η represents an $n \times 1$ vector whose generic ith component is $\phi(\bar{\gamma}_i)\Phi_{n-1}(\bar{\gamma}_{-i} - \bar{\Gamma}_{-i}\bar{\gamma}_i; \bar{\Gamma}_{-i,-i} - \bar{\Gamma}_{-i}\bar{\Gamma}_{-i}^{\mathrm{T}})$, with $\bar{\gamma}_i$ and $\bar{\gamma}_{-i}$ denoting, respectively, the ith element of $s^{-1}D\xi = \gamma_{\text{post}}$ and the $(n-1) \times 1$ vector obtained by removing the ith entry in γ_{post}. Similarly, $\bar{\Gamma}_{-i,-i}$ defines the submatrix of $s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1} = \Gamma_{\text{post}}$ without the ith row and column, and $\bar{\Gamma}_{-i}$ is the ith column of Γ_{post} with the ith row element removed. Computing the expectation using (6) is more efficient than numerical integration, since it requires only the calculation of $n + 1$ cumulative distribution functions, which typically involves far fewer evaluations of $\Phi_n(\cdot)$ than would be needed in numerical integration of the marginal posteriors. However, obtaining expressions for higher-order marginal and joint moments via direct derivation of (5) requires tedious calculations (Gupta et al., 2013), hence motivating the use of Monte Carlo methods based on samples from the posterior, as discussed in § 2.4. See also Gupta et al. (2013) and Azzalini & Bacchieri (2010) for expressions of the variance-covariance matrix and the cumulative distribution function of a generic unified skew-normal distribution. Both quantities, appropriately computed under the parameters in Theorem 1, are useful in posterior inference.

Although inference on the posterior distribution of β is often of interest, prediction of a future response $y_{\text{new}} \in \{0; 1\}$ given the associated covariates $x_{\text{new}} \in \mathbb{R}^p$ and the current data (y, X) is a primary goal in applications of probit models to classification. Within the Bayesian framework, this task requires derivation of the posterior predictive distribution $(y_{\text{new}} \mid y, X, x_{\text{new}})$, which in the binary case is simply a Bernoulli distribution with parameter $\text{pr}(y_{\text{new}} = 1 \mid y, X, x_{\text{new}}) = \int \Phi(x_{\text{new}}^{\text{T}}\beta)\pi(\beta \mid y, X)\,\text{d}\beta$. Corollary 2 makes this probability available in explicit form.

COROLLARY 2. *If* $(y_i \mid x_i, \beta) \sim \text{Ber}\{\Phi(x_i^{\text{T}}\beta)\}$ $(i = 1, \ldots, n)$ *and* $\beta \sim N_p(\xi, \Omega)$, *then*

$$\text{pr}(y_{\text{new}} = 1 \mid y, X, x_{\text{new}}) = \frac{\Phi_{n+1}\{s_{\text{new}}^{-1}D_{\text{new}}\xi; s_{\text{new}}^{-1}(D_{\text{new}}\Omega D_{\text{new}}^{\text{T}} + I_{n+1})s_{\text{new}}^{-1}\}}{\Phi_n\{s^{-1}D\xi; s^{-1}(D\Omega D^{\text{T}} + I_n)s^{-1}\}}, \tag{7}$$

where D_{new} *represents the* $(n + 1) \times p$ *matrix obtained by adding a last row* $d_{\text{new}}^{\text{T}} = x_{\text{new}}^{\text{T}}$ *to* D, *and* $s_{\text{new}} = \text{diag}\{(d_1^{\text{T}}\Omega d_1 + 1)^{1/2}, \ldots, (d_n^{\text{T}}\Omega d_n + 1)^{1/2}, (d_{\text{new}}^{\text{T}}\Omega d_{\text{new}} + 1)^{1/2}\}$.

An advantage of (7) over Markov chain Monte Carlo strategies (e.g., Albert & Chib, 1993; Holmes & Held, 2006; Frühwirth-Schnatter & Frühwirth, 2007; Polson et al., 2013) is that prediction does not require Monte Carlo integration for $\int \Phi(x_{\text{new}}^{\text{T}}\beta)\pi(\beta \mid y, X)\,\text{d}\beta$ via sampling of β from the posterior, and hence the computational burden does not depend on p. As will be discussed in § 2.4, this result is especially useful in large-p and small-to-moderate-n situations.

The above derivations are further helpful in obtaining explicit methods for performing Bayesian selection among models $\mathcal{M}_1, \ldots, \mathcal{M}_K$ characterizing, in general, different subsets $\mathcal{J}_1, \ldots, \mathcal{J}_K$ of covariates entering the linear predictor. Although there are different strategies for model selection (e.g., O'Hara & Sillanpää, 2009), the general approach is to formally define prior probabilities $\text{pr}(\mathcal{M}_1), \ldots, \text{pr}(\mathcal{M}_K)$ for the set of models and then rank them by the posterior probabilities $\text{pr}(\mathcal{M}_k \mid y, X) \propto \text{pr}(\mathcal{M}_k) \int \text{pr}(y \mid \mathcal{M}_k, X, \beta_{\mathcal{J}_k})\pi(\beta_{\mathcal{J}_k} \mid \mathcal{M}_k)\,\text{d}\beta_{\mathcal{J}_k}$ for $k = 1, \ldots, K$ (Chipman et al., 2001; Forte et al., 2018). Clearly, the principal issue in this task is the calculation of $\int \text{pr}(y \mid \mathcal{M}_k, X, \beta_{\mathcal{J}_k})\pi(\beta_{\mathcal{J}_k} \mid \mathcal{M}_k)\,\text{d}\beta_{\mathcal{J}_k}$, which may be intractable in the absence of conjugacy, thus requiring Monte Carlo integration or approximations (e.g., Kass & Raftery, 1995). This procedure can also be implemented in probit models by using the methods in § 1, but the same issues arise with regard to inference and computational performance as discussed previously. Corollary 3 instead provides an explicit formula for the marginal likelihood in probit models with Gaussian priors, which can be easily evaluated, especially in the large-p and small-to-moderate-n settings of interest in such studies.

COROLLARY 3. *Let* $\mathcal{M}_k$ *define the probit regression for* $y_1, \ldots, y_n$, *including only the covariates with indices in the subset* $\mathcal{J}_k \subset \{1; \ldots; p\}$. *Moreover, assume* $(\beta_{\mathcal{J}_k} \mid \mathcal{M}_k) \sim N_{p_k}(\xi_k, \Omega_k)$, *where* $p_k = |\mathcal{J}_k|$, *and* $\beta_{\mathcal{J}_k} \in \mathbb{R}^{p_k}$ *are the probit coefficients for the covariates in model* $\mathcal{M}_k$. *Then*

$$\int \text{pr}(y \mid \mathcal{M}_k, X, \beta_{\mathcal{J}_k})\pi(\beta_{\mathcal{J}_k} \mid \mathcal{M}_k)\,\text{d}\beta_{\mathcal{J}_k} = \Phi_n\{s_k^{-1}D_k\xi_k; s_k^{-1}(D_k\Omega_k D_k^{\text{T}} + I_n)s_k^{-1}\} \tag{8}$$

for each model $\mathcal{M}_k$ $(k = 1, \ldots, K)$, *with* $D_k = \text{diag}(2y_1 - 1, \ldots, 2y_n - 1)X_k \in \mathbb{R}^{n \times p_k}$, $s_k = \text{diag}\{(d_{1k}^{\text{T}}\Omega_k d_{1k} + 1)^{1/2}, \ldots, (d_{nk}^{\text{T}}\Omega_k d_{nk} + 1)^{1/2}\} \in \mathbb{R}_+^{n \times n}$, *and* $X_k \in \mathbb{R}^{n \times p_k}$ *denoting the* $n \times p_k$ *design matrix of covariates with indices in* $\mathcal{J}_k$.

Equation (8) is additionally useful for computing Bayes factors (e.g., Kass & Raftery, 1995) and in performing model averaging (Hoeting et al., 1999) without sampling from the posterior.

2.4. *Computational considerations and sampling procedures*

All the inference methods outlined in § 2.3 can, in principle, proceed via direct strategies without sampling from the posterior, thus improving upon existing procedures in large-p applications. The only barrier, which is relevant to the large-n case, is the evaluation of $\Phi_n(\cdot)$. Quasi-randomized Monte Carlo (Genz, 1992; Genz & Bretz, 2009) allows accurate calculation of $\Phi_n(\cdot)$ for small n, and has recently been improved via minimax tilting (Botev, 2017) to ensure effective evaluation of $\Phi_n(\cdot)$ in moderate-n settings. This procedure, available in the `R` library `TruncatedNormal` (R Development Core Team, 2019), has a rare vanishing asymptotic relative error, thus allowing tractable inference without sampling from the posterior in studies involving typically a few hundreds of units. The strategy is also useful in larger-n applications where few evaluations of $\Phi_n(\cdot)$ are required, for example in the prediction of not many outcomes and selection among a few models. However, for more general inferential tasks requiring many evaluations of $\Phi_n(\cdot)$, such as numerical integration, calculation of moments and high-dimensional prediction or model selection, inference without sampling from the posterior may face nonnegligible increases in computational time as n becomes larger; see Botev (2017, § 5) for details on scalability in the evaluation of $\Phi_n(\cdot)$. In this situation, sampling from the posterior provides a tractable strategy for obtaining numerical evaluations of generic functionals $E\{g(\beta) \mid y, X\} = \int g(\beta)\pi(\beta \mid y, X)\,\mathrm{d}\beta$ via Monte Carlo integration. Indeed, the availability of a large number R of replicates from the unified skew-normal posterior allows fast and accurate approximation of $E\{g(\beta) \mid y, X\}$ via $\sum_{r=1}^{R} g(\beta^{(r)})/R$.

Popular routines for the above purpose require data-augmentation Markov chain Monte Carlo (e.g., Albert & Chib, 1993; Holmes & Held, 2006; Frühwirth-Schnatter & Frühwirth, 2007; Polson et al., 2013), which has poor performance, especially in imbalanced high-dimensional settings (Johndrow et al., 2019). This issue can be addressed by Algorithm 1, which combines the stochastic representation of the unified skew-normal posterior in Corollary 1 with a new scheme proposed by Botev (2017) for obtaining independent samples from multivariate truncated Gaussians.

Algorithm 1. Exact scheme for drawing independent samples from the posterior in Theorem 1.

for r *from* 1 *to* R do

[1] Sample $V_0^{(r)}$ from $N_p\{0_p, \bar{\Omega} - \bar{\Omega}\omega D^{\mathrm{T}}(D\Omega D^{\mathrm{T}} + I_n)^{-1}D\omega\bar{\Omega}\}$ (in `R` use `rmvnorm`).

[2] Sample $V_1^{(r)}$ from an n-variate truncated Gaussian with mean vector 0_n, covariance matrix $s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}$ and truncation below $-s^{-1}D\xi$, using the accept-reject algorithm of Botev (2017) (in `R` use `mvrandn`).

[3] Compute $\beta^{(r)}$ via $\beta^{(r)} = \xi + \omega\{V_0^{(r)} + \bar{\Omega}\omega D^{\mathrm{T}}(D\Omega D^{\mathrm{T}} + I_n)^{-1}sV_1^{(r)}\}$.

output: $\beta^{(1)}, \ldots, \beta^{(R)}$

The routine of Botev (2017) relies on minimax tilting and accept-reject methods to improve the acceptance rate of classical rejection sampling, while avoiding the convergence and mixing issues of Markov chain Monte Carlo methods. By combining this sampler with classical routines for multivariate Gaussians, Algorithm 1 inherits their good properties, thus improving upon the computational methods discussed in § 1, especially in large-p and small-to-moderate-n applications. Clearly, when n increases and p decreases, sampling from the n-variate truncated Gaussian progressively affects computational time in favour of more efficient Markov chain Monte Carlo strategies that directly explore the p-dimensional parameter space

(e.g., Chopin & Ridgway, 2017). In this situation, a possible way of scaling up the computations is to exploit the structure of Algorithm 1 to perform parallel computing. Another alternative is to leverage the closure properties of the multivariate truncated Gaussian under conditioning (Horrace, 2005), and iteratively block-update subvectors of V_1 whose dimension can still allow efficient sampling by the methods of Botev (2017). Although this hybrid strategy could induce some autocorrelation in the posterior samples of β, the blocking approach typically guarantees improvements in mixing and convergence (e.g., Roberts & Sahu, 1997).

It is also worth mentioning that Botev (2017) applied his accept-reject method to Bayesian probit regression. However, unlike for Algorithm 1, his proposed strategy requires sampling from $(n + p)$-variate truncated Gaussians. Separating these two blocks, as in Algorithm 1, reduces computational complexity and allows parallel computing. A more similar representation can be found in Holmes & Held (2006, § 2.1) and in the documentation of the `R` library `TruncatedNormal` by Botev (2017). In fact, the resulting routines are closely related to Algorithm 1. However, Holmes & Held (2006, § 2.1) and Botev (2017, § 5.4) based their derivations on different arguments, without noticing that the posterior is in fact a unified skew-normal distribution. This last result and its broader implications arguably constitute the most important contribution of the present article.

Finally, Algorithm 1 can also be adapted to sample from a generic unified skew-normal distribution. This strategy can be broadly useful beyond Bayesian inference. An example is a parametric bootstrap for frequentist inference on the unified skew-normal parameters.

2.5. *A class of conjugate unified skew-normal priors for Bayesian probit regression*

The derivations in § 2.2 suggest the more general result outlined in Corollary 4, thereby allowing tractable inference in Bayesian probit regression under more flexible priors for β.

COROLLARY 4. *If* $(y_i \mid x_i, \beta) \sim \text{Ber}\{\Phi(x_i^{\mathrm{T}}\beta)\}$ *independently for* $i = 1, \ldots, n$ *and* β *is assigned a* $\text{SUN}_{p,m}(\xi, \Omega, \Delta, \gamma, \Gamma)$ *prior* (Arellano-Valle & Azzalini, 2006)*, then*

$$(\beta \mid y, X) \sim \text{SUN}_{p,m+n}(\xi_{\text{post}}, \Omega_{\text{post}}, \Delta_{\text{post}}, \gamma_{\text{post}}, \Gamma_{\text{post}}), \tag{9}$$

with updated parameters $\xi_{\text{post}} = \xi$, $\Omega_{\text{post}} = \Omega$, $\Delta_{\text{post}} = (\Delta, \bar{\Omega}\omega D^{\mathrm{T}} s^{-1})$, $\gamma_{\text{post}} = (\gamma^{\mathrm{T}}, \xi^{\mathrm{T}} D^{\mathrm{T}} s^{-1})^{\mathrm{T}}$ *and* Γ_{post} *characterizing an* $(m+n) \times (m+n)$ *full-rank correlation matrix having block entries* $\Gamma_{\text{post}[11]} = \Gamma$, $\Gamma_{\text{post}[22]} = s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}$ *and* $\Gamma_{\text{post}[21]} = \Gamma_{\text{post}[12]}^{\mathrm{T}} = s^{-1}D\omega\Delta$.

According to Corollary 4, tractable inference in Bayesian probit regression is possible under a broader class of priors. Indeed, all the methods discussed in the previous subsections also apply to this more general case, since the posterior in (9) is still a unified skew-normal distribution. This ensures increased flexibility in prior specification, thus allowing departures from normality. Although the general unified skew-normal choice may be uncommon in applied contexts, such a class incorporates several priors of interest, including multivariate Gaussians, independent skew-normal distributions for each $\beta_1, \ldots, \beta_p$, and multivariate skew-normal distributions for β (Arellano-Valle & Azzalini, 2006).

3. EMPIRICAL STUDIES

To evaluate the methods developed in § 2 and compare their performance with the popular strategies for Bayesian inference in probit regression discussed in § 1, consider a freely available dataset on gene expression levels in $n = 74$ normal and cancerous biological tissue samples at

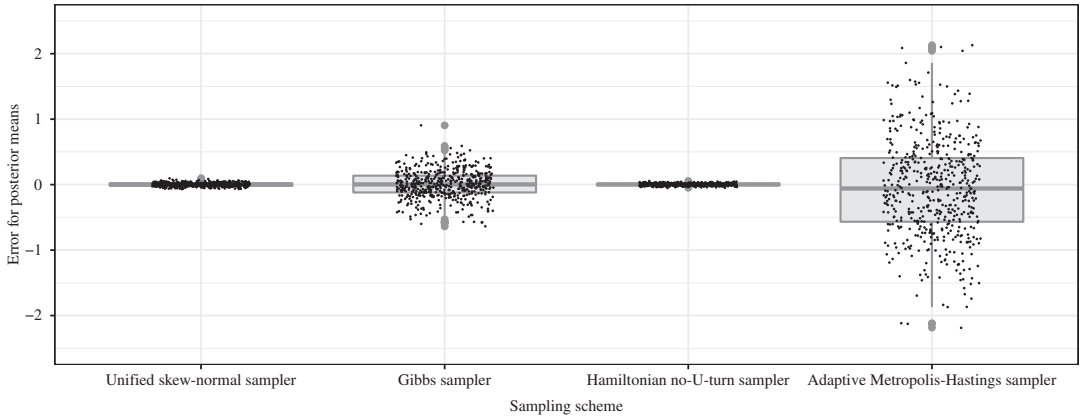

Fig. 2. Performance of moments calculation: boxplots of the differences between the posterior means for the coefficients based on samples from $(\beta \mid y, X)$ and those calculated from (6), for each sampling scheme under study; the dots represent the values of the differences from which each boxplot is derived.

$p - 1 = 516$ different tags (Martinez et al., 2005). A main goal in such applications is to quantify the effects of gene expressions on the probability of cancerous tissue, and to predict the status of new tissue samples as a function of the gene expressions (e.g., Tzanis & Vlahavas, 2007). With this goal in mind, here the focus is on studying the location of the posterior for β and the predictive distribution in the Bayesian model $(y_i \mid x_i, \beta) \sim \text{Ber}\{\Phi(x_i^{\top}\beta)\}$ $(i = 1, \ldots, n)$, with a $\beta \sim N_{517}(0_{517}, 16 \times I_{517})$ prior. In this probit regression, x_i denotes the vector consisting of an intercept term and the gene expressions for tissue i, and y_i is 1 or 0 depending on whether the tissue is cancerous or not, respectively.

The choice of a weakly informative prior for the β coefficients is motivated by the guidelines in Gelman et al. (2008) and by similar implementations in Botev (2017) and Chopin & Ridgway (2017). In line with these works, the gene expressions at the 516 different tags were standardized to have mean zero and standard deviation 0.5. To assess predictive performance, the prior for β is updated with the information of 50 randomly chosen observations, and out-of-sample classification via the posterior predictive distribution is performed on the 24 held-out tissue samples.

Although other datasets could be considered, it is emphasized that state-of-the-art computational methods for probit regression provide valuable strategies in a variety of applications, but face mixing and time-inefficiency challenges in large-p and small-n studies (e.g., Chopin & Ridgway, 2017). As shown in Figs. 2 and 3, as well as in Table 1, the results outlined in § 2 yield notable improvements in such large-p and small-n settings, thus enabling straightforward Bayesian inference in applications where this task had previously been impractical. To clarify these results, the strategies in § 2 are compared with state-of-the-art procedures, including the data-augmentation Gibbs sampler of Albert & Chib (1993), the Hamiltonian no-U-turn sampler of Hoffman & Gelman (2014) and the adaptive Metropolis–Hastings method of Haario et al. (2001). To increase the acceptance rates and efficiency, the starting Gaussian proposal for the Metropolis–Hastings routine was initialized with the mean and rescaled variance-covariance matrix provided by an expectation propagation approximation. Consistent with Chopin & Ridgway (2017) and Roberts & Rosenthal (2001), the scaling factor was set to $2.38^2/p$.

The above Markov chain Monte Carlo algorithms were run for 20 000 iterations after a burn-in of 5000, and can easily be implemented in `R` using libraries `bayesm`, `rstan`, and a combination of `LaplacesDemon` and `EPGLM`. Although certain routines converged more rapidly than others

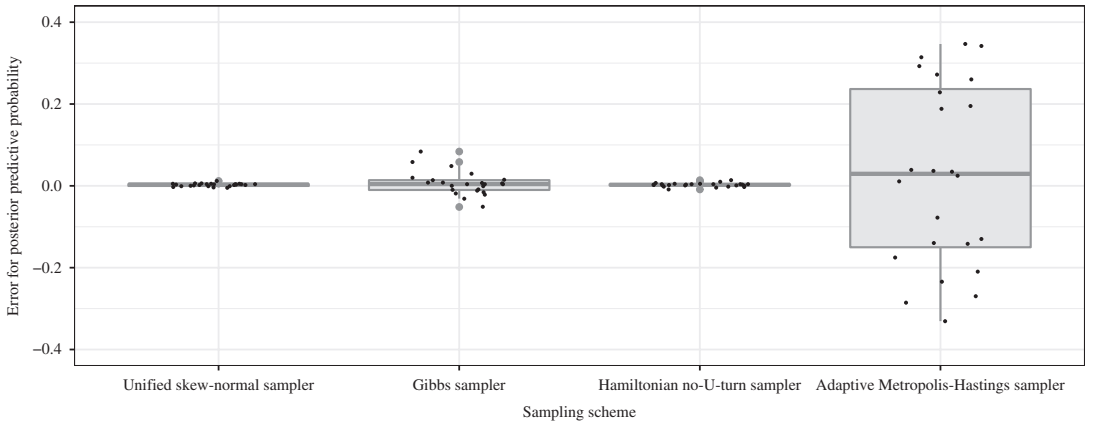

Fig. 3. Predictive performance: boxplots of the differences between the posterior predictive probabilities for the 24 held-out units based on samples from $(\beta \mid y, X)$ and those calculated from (7), for each sampling scheme under study; the dots represent the values of the differences from which each boxplot is derived.

Table 1. *Assessment of computational efficiency: total number of samples from* $(\beta \mid y, X)$ *per second and statistics summarizing the effective sample sizes computed from the produced chains for the coefficients* $\beta_1, \ldots, \beta_{517}$*, for each sampling scheme under analysis*

	Samples per second	Mixing via effective sample sizes		
	Samples of β per second	Minimum	First quartile	Median
Unified skew-normal sampler	886.64	20000.00	20000.00	20000.00
Gibbs sampler	13.48	55.46	2417.38	3687.18
Hamiltonian no-U-turn sampler	15.95	17730.54	20000.00	20000.00
Adaptive Metropolis–Hastings sampler	19.34	28.55	49.22	59.07

and with excellent mixing, the same settings were used for all the algorithms to facilitate comparison. In contrast, the sampling scheme in Algorithm 1 provides independent samples from the exact posterior, and hence requires no burn-in or convergence checks; the code and implementation details are available at `https://github.com/danieledurante/ProbitSUN`.

As seen in Table 1, the Hamiltonian sampler exhibits, in practice, the same mixing as Algorithm 1, but the latter has a significantly faster sampling speed. This could be due to the number of leap-frog steps required at each iteration of the no-U-turn sampler (Chopin & Ridgway, 2017). As expected, the Gibbs sampler and the Metropolis-Hastings algorithm display lower mixing, but have similar or improved running times compared to the Hamiltonian no-U-turn sampler. However, as is clear from Figs. 2 and 3, this reduction in mixing has a direct effect on the accuracy of posterior inference and prediction. In contrast, there is an almost perfect match between Monte Carlo and direct estimates of posterior means and posterior predictive probabilities for Algorithm 1 and the Hamiltonian no-U-turn sampler, although, as already discussed, the Hamiltonian routine is significantly slower than Algorithm 1 in this application. These computational gaps further increase in larger-p settings, with the competing methods becoming rapidly impractical. On the other hand, the inference and sampling methods that rely on the unified skew-normal results have difficulties in scaling with n. This result is confirmed by a voice rehabilitation study presented at `https://github.com/danieledurante/ProbitSUN`. However, in this application, with doubled n and almost halved p, Algorithm 1 remains competitive.

4. FINAL CONSIDERATIONS AND FUTURE DIRECTIONS OF RESEARCH

Although the focus has been on probit regression, the results in this article could also lead to computational gains in more complex formulations that rely on predictor-dependent observed or latent binary data, such as in mixture models for density regression (Rodriguez & Dunson, 2011). For instance, using the results in § 2.3, binary classification via Gaussian processes (Rasmussen & Williams, 2006) could avoid sampling or approximations by exploiting the closure properties of unified skew-normal distributions, especially under conditioning. Moreover, when binary regression serves as a latent dictionary function, sampling the binary data via (7), instead of conditioning on β, could speed up computations. Finally, the novel conjugacy results in § 2.5 open up new avenues for incorporating skewness into prior specification.

There are various directions for future progress. For example, a deeper understanding of the moment generating function of the unified skew-normal distribution could facilitate direct calculation of relevant functionals without the need to sample from $(\beta \mid y, X)$. Along the same lines, improving the methods for efficient evaluation of $\Phi_n(\cdot)$ in large-n applications, by means of data transformations, blocking methods (Chopin, 2011) or recent algorithms (Genton et al., 2018), could expand the range of applications in which direct inference, prediction and model selection are possible, without sampling from $(\beta \mid y, X)$. Obtaining approximations of the exact posterior which preserve the skewness, but allow analytical inference would also be an interesting future research direction. Finally, more detailed studies on particular forms of the prior variance-covariance matrix Ω, such as those associated with g-priors or generalized g-priors (Maruyama & George, 2011) and limiting cases arising from flat priors, could provide novel insights into the effects of specific choices, and potentially lead to simplifications in the unified skew-normal parameters that may ease posterior inference. It is also possible to consider hyperpriors for (ξ, Ω) such as the normal-inverse-Wishart. With this choice, Theorem 1 holds only for the full conditional $(\beta \mid y, X, \xi, \Omega)$. Moreover, $(\xi, \Omega \mid y, X, \beta)$ still has a normal-inverse-Wishart kernel, since (ξ, Ω) only enter the Gaussian prior for β. Hence, although a hierarchical prior would not allow direct sampling from the unified skew-normal posterior, Gibbs sampling methods can easily be applied to this situation.

As discussed in § 1, the availability of an exact posterior with tractable stochastic representations and closure properties can also motivate the development of novel finite-sample and asymptotic theory. Finally, although the studies in § 3 and the discussion in § 2.4 provide general guidelines on the practical usefulness of the results presented in § 2, additional quantitative assessments of scalability and its relation to specific prior settings or dataset structures would certainly be of interest.

ACKNOWLEDGEMENT

The author thanks Adelchi Azzalini, David Dunson and Giacomo Zanella for the stimulating discussions and helpful comments on a first draft of this article. The editor, associate editor and referees are also acknowledged for their useful suggestions. This work was supported by the European Research Council. The author is also affiliated with the Bocconi Institute for Data Sciences and Analytics.

APPENDIX

Proof of Lemma 1. Let $\Phi(x\beta)^y\{1-\Phi(x\beta)\}^{1-y} = \Phi\{(2y-1)x\beta\}$ denote the probability mass function of y in Lemma 1. Direct application of Bayes' rule gives

$$\pi(\beta \mid y, x) \propto \phi(\beta)\Phi\{(2y-1)x\beta\} = \phi(\beta)\Phi\{(2y-1)x(x^2+1)^{-1/2}\beta; (x^2+1)^{-1}\}.$$

Hence, letting $\xi_{\text{post}} = 0$, $\Omega_{\text{post}} = 1$, $\Delta_{\text{post}} = (2y-1)x(x^2+1)^{-1/2}$, $\gamma_{\text{post}} = 0$ and $\Gamma_{\text{post}} = (x^2+1)^{-1} + \Delta_{\text{post}}^{\mathrm{T}}\Delta_{\text{post}} = 1$ yields the kernel of the unified skew-normal distribution in Lemma 1, with correlation matrix Ω^*_{post} having block entries $\Omega^*_{\text{post}[11]} = \Omega^*_{\text{post}[22]} = 1$ and $\Omega^*_{\text{post}[21]} = \Omega^*_{\text{post}[12]} = \Delta_{\text{post}}$. □

Proof of Theorem 1. Adapting the proof of Lemma 1, one can write the joint probability mass function of the responses y as $\prod_{i=1}^{n} \Phi\{(2y_i-1)x_i^{\mathrm{T}}\beta\} = \Phi_n(D\beta; I_n) = \Phi_n(s^{-1}D\beta; s^{-1}s^{-1})$, with D and s defined as in Theorem 1. Combining this likelihood for y with the Gaussian prior for β gives

$$\pi(\beta \mid y, X) \propto \phi_p(\beta-\xi;\Omega)\Phi_n(s^{-1}D\beta; s^{-1}s^{-1}) = \phi_p(\beta-\xi;\Omega)\Phi_n\{s^{-1}D\xi + s^{-1}D(\beta-\xi); s^{-1}s^{-1}\}.$$

To establish the relationship between the above kernel and the unified skew-normal density in (3), rewrite $(\beta-\xi)$ as $\omega\bar{\Omega}\bar{\Omega}^{-1}\omega^{-1}(\beta-\xi)$. Then, letting $\xi_{\text{post}} = \xi$, $\Omega_{\text{post}} = \Omega$, $\Delta_{\text{post}} = \bar{\Omega}\omega D^{\mathrm{T}}s^{-1}$, $\gamma_{\text{post}} = s^{-1}D\xi$ and $\Gamma_{\text{post}} = s^{-1}s^{-1} + \Delta_{\text{post}}^{\mathrm{T}}\bar{\Omega}^{-1}\Delta_{\text{post}} = s^{-1}s^{-1} + s^{-1}D\omega\bar{\Omega}\bar{\Omega}^{-1}\bar{\Omega}\omega D^{\mathrm{T}}s^{-1} = s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}$ leads to the kernel of a unified skew-normal distribution whose parameters coincide with those given in Theorem 1. To conclude the proof it is also necessary to verify that

$$\Omega^*_{\text{post}} = \begin{pmatrix} s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1} & s^{-1}D\omega\bar{\Omega} \\ \bar{\Omega}\omega D^{\mathrm{T}}s^{-1} & \bar{\Omega} \end{pmatrix} = \begin{pmatrix} s^{-1} & 0 \\ 0 & \omega^{-1} \end{pmatrix} \times \begin{pmatrix} D\Omega D^{\mathrm{T}} + I_n & D\Omega \\ \Omega D^{\mathrm{T}} & \Omega \end{pmatrix} \times \begin{pmatrix} s^{-1} & 0 \\ 0 & \omega^{-1} \end{pmatrix}$$

is a full-rank correlation matrix. This is a direct consequence of the fact that Ω^*_{post} coincides with the correlation matrix of the random vector $(z_1^{\mathrm{T}}, z_2^{\mathrm{T}})^{\mathrm{T}}$ where $z_1 = Dz_2 + \epsilon$, with $E(\epsilon) = 0_n$ and $E(\epsilon\epsilon^{\mathrm{T}}) = I_n$, and z_2 is a p-variate random variable with zero mean and positive-definite variance-covariance matrix $E(z_2 z_2^{\mathrm{T}}) = \Omega = \omega\bar{\Omega}\omega$. Finally, $s = \text{diag}\{(d_1^{\mathrm{T}}\Omega d_1 + 1)^{1/2}, \ldots, (d_n^{\mathrm{T}}\Omega d_n + 1)^{1/2}\}$ is the diagonal matrix whose entries are the square roots of the diagonal elements of $E(z_1 z_1^{\mathrm{T}}) = D\Omega D^{\mathrm{T}} + I_n$. □

Proof of Corollary 1. The proof is a simple adaptation of equation (7.4) in Azzalini & Capitanio (2014, § 7.1.2) to the unified skew-normal posterior in Theorem 1. In particular, according to Azzalini & Capitanio (2014, § 7.1.2), the posterior in (2) has the same distribution as the random variable

$$\xi_{\text{post}} + \omega_{\text{post}}(V_0 + \Delta_{\text{post}}\Gamma_{\text{post}}^{-1}V_1),$$

where $V_0 \sim N_p(0_p, \bar{\Omega}_{\text{post}} - \Delta_{\text{post}}\Gamma_{\text{post}}^{-1}\Delta_{\text{post}}^{\mathrm{T}})$ and V_1 is from an n-variate truncated normal with mean 0_n, covariance matrix Γ_{post} and truncation below $-\gamma_{\text{post}}$. Replacing the posterior parameters in this stochastic representation with their expressions in Theorem 1 concludes the proof. To clarify this final claim, it is sufficient to re-express $\Delta_{\text{post}}\Gamma_{\text{post}}^{-1}$ and $\Delta_{\text{post}}\Gamma_{\text{post}}^{-1}\Delta_{\text{post}}^{\mathrm{T}}$ as $\bar{\Omega}\omega D^{\mathrm{T}}s^{-1}s(D\Omega D^{\mathrm{T}}+I_n)^{-1}s = \bar{\Omega}\omega D^{\mathrm{T}}(D\Omega D^{\mathrm{T}}+I_n)^{-1}s$ and $\bar{\Omega}\omega D^{\mathrm{T}}s^{-1}s(D\Omega D^{\mathrm{T}} + I_n)^{-1}ss^{-1}D\omega\bar{\Omega} = \bar{\Omega}\omega D^{\mathrm{T}}(D\Omega D^{\mathrm{T}} + I_n)^{-1}D\omega\bar{\Omega}$, respectively. □

Proof of Corollary 2. Recalling the expression for $\pi(\beta \mid y, X)$ in (3), the posterior predictive probability $\text{pr}(y_{\text{new}} = 1 \mid y, X, x_{\text{new}}) = \int \Phi(x_{\text{new}}^{\mathrm{T}}\beta)\pi(\beta \mid y, X)\,\mathrm{d}\beta$ can be expressed as

$$\Phi_n\{s^{-1}D\xi; s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}\}^{-1} \int \phi_p(\beta-\xi;\Omega)\Phi(x_{\text{new}}^{\mathrm{T}}\beta)\Phi_n(s^{-1}D\beta; s^{-1}s^{-1})\,\mathrm{d}\beta.$$

Using the proof of Theorem 1, the quantity inside the above integral can be re-expressed as $\phi_p(\beta-\xi;\Omega)\Phi_{n+1}(s_{\text{new}}^{-1}D_{\text{new}}\beta; s_{\text{new}}^{-1}s_{\text{new}}^{-1})$, with D_{new} and s_{new} defined as in Corollary 2. Upon comparing this function with the density of the unified skew-normal posterior defined in (3), it immediately follows that $\phi_p(\beta-\xi;\Omega)\Phi_{n+1}(s_{\text{new}}^{-1}D_{\text{new}}\beta; s_{\text{new}}^{-1}s_{\text{new}}^{-1})$ is the kernel of a unified skew-normal distribution with normalizing constant $\int \phi_p(\beta-\xi;\Omega)\Phi(x_{\text{new}}^{\mathrm{T}}\beta)\Phi_n(s^{-1}D\beta; s^{-1}s^{-1})\,\mathrm{d}\beta = \Phi_{n+1}\{s_{\text{new}}^{-1}D_{\text{new}}\xi; s_{\text{new}}^{-1}(D_{\text{new}}\Omega D_{\text{new}}^{\mathrm{T}} + I_{n+1})s_{\text{new}}^{-1}\}$. Substituting this into the above formula for the posterior predictive probability $\text{pr}(y_{\text{new}} = 1 \mid y, X, x_{\text{new}})$ completes the proof. □

Proof of Corollary 3. Recalling Theorem 1, the integral $\int \text{pr}(y \mid \mathcal{M}_k, X, \beta_{\mathcal{J}_k})\pi(\beta_{\mathcal{J}_k} \mid \mathcal{M}_k)\,\text{d}\beta_{\mathcal{J}_k}$ is the normalizing constant of the posterior for $\beta_{\mathcal{J}_k}$ in model $\mathcal{M}_k$. Hence, adapting (3) to model $\mathcal{M}_k$ leads to $\int \text{pr}(y \mid \mathcal{M}_k, X, \beta_{\mathcal{J}_k})\pi(\beta_{\mathcal{J}_k} \mid \mathcal{M}_k)\,\text{d}\beta_{\mathcal{J}_k} = \Phi_n\{s_k^{-1}D_k\xi_k; s_k^{-1}(D_k\Omega_k D_k^{\mathrm{T}} + I_n)s_k^{-1}\}$. □

Proof of Corollary 4. To prove Corollary 4, it suffices to generalize Theorem 1. In particular, adapting the proof of Theorem 1 to the case in which $\beta \sim \text{SUN}_{p,m}(\xi, \Omega, \Delta, \gamma, \Gamma)$ yields

$$\pi(\beta \mid y, X) \propto \phi_p(\beta - \xi; \Omega)\Phi_n(s^{-1}D\beta; s^{-1}s^{-1})\Phi_m\{\gamma + \Delta^{\mathrm{T}}\bar{\Omega}^{-1}\omega^{-1}(\beta - \xi); \Gamma - \Delta^{\mathrm{T}}\bar{\Omega}^{-1}\Delta\}.$$

To proceed with the calculations, first recall the proof of Theorem 1 and write $\Phi_n(s^{-1}D\beta; s^{-1}s^{-1})$ as $\Phi_n\{s^{-1}D\xi + (\bar{\Omega}\omega D^{\mathrm{T}}s^{-1})^{\mathrm{T}}\bar{\Omega}^{-1}\omega^{-1}(\beta - \xi); s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1} - s^{-1}D\omega\bar{\Omega}\bar{\Omega}^{-1}\bar{\Omega}\omega D^{\mathrm{T}}s^{-1}\}$.

Let us now define the appropriate parameters for which the above kernel coincides with that of the unified skew-normal posterior in Corollary 4. This is easily accomplished by setting $\xi_{\text{post}} = \xi$, $\Omega_{\text{post}} = \Omega$, $\Delta_{\text{post}} = (\Delta, \bar{\Omega}\omega D^{\mathrm{T}}s^{-1})$ and $\gamma_{\text{post}} = (\gamma^{\mathrm{T}}, \xi^{\mathrm{T}}D^{\mathrm{T}}s^{-1})^{\mathrm{T}}$, and taking Γ_{post} to be a full-rank correlation matrix with blocks $\Gamma_{\text{post}[11]} = \Gamma$, $\Gamma_{\text{post}[22]} = s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1}$ and $\Gamma_{\text{post}[21]} = \Gamma_{\text{post}[12]}^{\mathrm{T}} = s^{-1}D\omega\Delta$. As in Theorem 1, it is also necessary to verify that

$$\begin{pmatrix} \Gamma & \Delta^{\mathrm{T}}\omega D^{\mathrm{T}}s^{-1} & \Delta^{\mathrm{T}} \\ s^{-1}D\omega\Delta & s^{-1}(D\Omega D^{\mathrm{T}} + I_n)s^{-1} & s^{-1}D\omega\bar{\Omega} \\ \Delta & \bar{\Omega}\omega D^{\mathrm{T}}s^{-1} & \bar{\Omega} \end{pmatrix}$$

is a full-rank correlation matrix. This follows from minor modifications of the proof of Theorem 1. In fact, the $(m + p) \times (m + p)$ prior matrix Ω^* with block entries $\Omega^*_{[11]} = \Gamma$, $\Omega^*_{[22]} = \bar{\Omega} = \omega^{-1}\Omega\omega^{-1}$ and $\Omega^*_{[21]} = \Omega^{*\mathrm{T}}_{[12]} = \Delta$ is, by definition, a full-rank correlation matrix. □

REFERENCES

AGRESTI, A. (2013). *Categorical Data Analysis*. Hoboken, New Jersey: Wiley, 3rd ed.

ALBERT, J. H. & CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.* **88**, 669–79.

ARELLANO-VALLE, R. B. & AZZALINI, A. (2006). On the unification of families of skew-normal distributions. *Scand. J. Statist.* **33**, 561–74.

ARNOLD, B. C. & BEAVER, R. J. (2000). Hidden truncation models. *Sankhyā* **62**, 23–35.

ARNOLD, B. C., BEAVER, R. J., AZZALINI, A., BALAKRISHNAN, N., BHAUMIK, A., DEY, D., CUADRAS, C. & SARABIA, J. M. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test* **11**, 7–54.

AZZALINI, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171–8.

AZZALINI, A. & BACCHIERI, A. (2010). A prospective combination of phase II and phase III in drug development. *Metron* **68**, 347–69.

AZZALINI, A. & CAPITANIO, A. (1999). Statistical applications of the multivariate skew normal distribution. *J. R. Statist. Soc.* B **61**, 579–602.

AZZALINI, A. & CAPITANIO, A. (2014). *The Skew-Normal and Related Families*. Cambridge: Cambridge University Press.

AZZALINI, A. & DALLA VALLE, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–26.

BAZÁN, J. L., BRANCO, M. D. & BOLFARINE, H. (2006). A skew item response model. *Bayesian Anal.* **1**, 861–92.

BOTEV, Z. (2017). The normal law under linear restrictions: Simulation and estimation via minimax tilting. *J. R. Statist. Soc.* B **79**, 125–48.

CHIPMAN, H. A., GEORGE, E. I. & MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Statist.* **4**, 266–98.

CHIPMAN, H. A., GEORGE, E. I., MCCULLOCH, R. E., CLYDE, M., FOSTER, D. P. & STINE, R. A. (2001). The practical implementation of Bayesian model selection. In *Model Selection*. Lecture Notes–Monograph Series, vol. 38. Hayward, California: Institute of Mathematical Statistics, pp. 65–116.

CHOPIN, N. (2011). Fast simulation of truncated Gaussian distributions. *Statist. Comp.* **21**, 275–88.

CHOPIN, N. & RIDGWAY, J. (2017). Leave Pima Indians alone: Binary regression as a benchmark for Bayesian computation. *Statist. Sci.* **32**, 64–87.

FORTE, A., GARCIA-DONATO, G. & STEEL, M. (2018). Methods and tools for Bayesian variable selection and model averaging in normal linear regression. *Int. Statist. Rev.* **86**, 237–58.

Frühwirth-Schnatter, S. & Frühwirth, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Comp. Statist. Data Anal.* **51**, 3509–28.
Gelman, A., Jakulin, A., Pittau, M. G. & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Statist.* **2**, 1360–83.
Genton, M. G., Keyes, D. E. & Turkiyyah, G. (2018). Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities. *J. Comp. Graph. Statist.* **27**, 268–77.
Genz, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comp. Graph. Statist.* **1**, 141–9.
Genz, A. & Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Berlin: Springer.
González-Farías, G., Domínguez-Molina, A. & Gupta, A. K. (2004). Additive properties of skew normal random vectors. *J. Statist. Plan. Infer.* **126**, 521–34.
Gupta, A. K., Aziz, M. A. & Ning, W. (2013). On some properties of the unified skew-normal distribution. *J. Statist. Theory Pract.* **7**, 480–95.
Gupta, A. K., González-Farías, G. & Domínguez-Molina, A. (2004). A multivariate skew normal distribution. *J. Mult. Anal.* **89**, 181–90.
Haario, H., Saksman, E. & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–42.
Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14**, 382–401.
Hoffman, M. D. & Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–623.
Holmes, C. C. & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.* **1**, 145–68.
Horrace, W. C. (2005). Some results on the multivariate truncated normal distribution. *J. Mult. Anal.* **94**, 209–21.
Johndrow, J. E., Smith, A., Pillai, N. & Dunson, D. B. (2019). MCMC for imbalanced categorical data. *J. Am. Statist. Assoc.* **114**, 1394–403.
Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.
Kuss, M. & Rasmussen, C. E. (2005). Assessing approximate inference for binary Gaussian process classification. *J. Mach. Learn. Res.* **6**, 1679–704.
Martinez, R., Christen, R., Pasquier, C. & Pasquier, N. (2005). Exploratory analysis of cancer SAGE data. In *Proc. ECML–PKDD Discovery Challenge Workshop (Porto, Portugal)*. Berlin: Springer, pp. 72–7.
Maruyama, Y. & George, E. (2011). Fully Bayes factors with a generalized *g*-prior. *Ann. Statist.* **39**, 2740–65.
O'Hara, R. B. & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **4**, 85–117.
Polson, N. G., Scott, J. G. & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Am. Statist. Assoc.* **108**, 1339–49.
Quarteroni, A., Sacco, R. & Saleri, F. (2010). *Numerical Mathematics*. Berlin: Springer, 2nd ed.
R Development Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press.
Roberts, G. O. & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.* **16**, 351–67.
Roberts, G. O. & Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Statist. Soc.* B **59**, 291–317.
Rodriguez, A. & Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6**, 145–78.
Spiegelhalter, D. J. & Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20**, 579–605.
Tzanis, G. & Vlahavas, I. (2007). Accurate classification of SAGE data based on frequent patterns of gene expression. In *Proc. 19th IEEE Int. Conf. Tools with Artificial Intelligence (Patras, Greece)*. Washington, DC: IEEE, pp. 96–100.

[*Received on* 18 *June* 2018. *Editorial decision on* 2 *January* 2019]