# Tractable Bayes of Skew-Elliptical Link Models for Correlated Binary Data

Zhongwei Zhang[*], Reinaldo B. Arellano-Valle[†], Marc G. Genton[*], Raphaël Huser[*]

**Abstract**

Correlated binary response data with covariates are ubiquitous in longitudinal or spatial studies. Among the existing statistical models the most well-known one for this type of data is the multivariate probit model, which uses a Gaussian link to model dependence at the latent level. However, a symmetric link may not be appropriate if the data are highly imbalanced. Here, we propose a multivariate skew-elliptical link model for correlated binary responses, which includes the multivariate probit model as a special case. Furthermore, we perform Bayesian inference for this new model and prove that the regression coefficients have a closed-form unified skew-elliptical posterior. The new methodology is illustrated by application to COVID-19 data from three different counties of the state of California, USA. By jointly modeling extreme spikes in weekly new cases, our results show that the spatial dependence cannot be neglected. Furthermore, the results also show that the skewed latent structure of our proposed model improves the flexibility of the multivariate probit model and provides better fit to our highly imbalanced dataset.

*Keywords:* Asymmetric link model; Correlated binary data; COVID-19 pandemic; Markov Chain Monte Carlo; Tractable Bayes; Unified skew-elliptical distribution.

# 1 Introduction

Correlated binary response data with covariates frequently arise in longitudinal (Fitzmaurice et al., 1995, 2008) or spatial studies (Heagerty & Lele, 1998; Lin & Clayton, 2005). For instance, in longitudinal studies, the disease status (i.e., diseased or not diseased) is measured over time on the same person. Similarly, in a panel study of income dynamics, the employment status information may be collected over time from the same survey participant. The multivariate probit model (Ashford & Sowden, 1970; Chib & Greenberg, 1998) is well-known for this type of data, as it describes the dependence between binary variables by a latent Gaussian link, which allows for flexible modelling of dependence, has straightforward interpretation of the parameters and is easily amenable to Bayesian inference.

[*] CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

[†] Departamento of Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile

A symmetric link, however, does not always provide the best fit to a given dataset; see Chen et al. (1999), Kim et al. (2008) for some examples. In this case, the link might be misspecified, yielding substantial bias in the mean response estimates (Czado & Santner, 1992). Chen et al. (1999) used the rate at which the probability of a given binary response variable approaches 0 and 1 to guide the selection of a symmetric or asymmetric link. In other words, if the binary response data are highly imbalanced, the rate of the probability of the random variable approaching 0 is typically very different from the one approaching 1, so that an asymmetric link might be preferred than a symmetric link. Motivated by this observation, a variety of flexible asymmetric link models have been proposed for univariate binary response data. Surprisingly, to the best of our knowledge, no multivariate asymmetric link models have previously been proposed in the literature for correlated binary responses. The purpose of this paper is to fill this important gap by proposing a flexible multivariate skew-elliptical link model for correlated binary responses, which includes the multivariate probit model as a special case and allows for fast and accurate Bayesian inference; see Section 2.2 for details on the multivariate skew-elliptical distribution, used in our model as a key building block.

Durante (2019) has proved that for the univariate probit model with Gaussian priors, the posterior of the regression coefficients belongs to the class of unified skew-normal distributions (Arellano-Valle & Azzalini, 2006). This result has led to a similar result for the multinomial probit model (Fasano & Durante, 2020) and a closed-form predictive probability in probit models with Gaussian process priors (Cao et al., 2020a). In this paper we also consider Bayesian inference for our new multivariate model and prove that the posterior of the regression coefficients belongs to the unified skew-elliptical family. The closed-form and tractable posterior for the regression coefficients facilitates inference by using an algorithm which does not rely on data-augmentation, and thus avoids the convergence and mixing issues of the classical data-augmentation algorithms for probit models; see Johndrow et al. (2019) for a discussion of this issue.

We illustrate the new methodology by application to COVID-19 pandemic data from three different counties of the state of California, USA. By jointly modeling the occurrences of extreme spikes in weekly new infected cases using our new model, we can estimate the underlying spatial dependence structure, which might provide helpful quantitative insights into the transmission modes of the virus and help authorities mitigate its spread. Furthermore, our model has additional skewness parameters compared to the multivariate probit model, which improves its flexibility and makes it more appropriate for modeling our highly imbalanced dataset.

# 2 Preliminaries: Skew-Elliptical and Unified Skew-Elliptical Distributions

## 2.1 The Skew-Elliptical Distribution

The skew-elliptical distribution, originally proposed by Azzalini & Capitanio (1999), was formulated by multiplying an elliptical density with a skewing function. Branco & Dey (2001) proposed a new formulation of the skew-elliptical distribution by means of a conditioning mechanism. The close relationship between these two formulations is established in Azzalini & Capitanio (2003). Thanks to the construction in terms of a conditioning mechanism, the formulation in Branco & Dey (2001) has led to many attractive properties of this class of distribution, such as existence of stochastic representation and closeness under marginalization and affine transformation. Fang (2003) considered a slightly wider class of distributions than Branco & Dey (2001) by adding an extra truncation parameter, which was later called the extended skew-elliptical distribution in Arellano-Valle & Genton (2010), and showed that this new distribution is closed under marginalization, affine transformation and also conditioning.

Here we adopt a slightly different parametrization than Fang (2003) with the truncation parameter taken as 0 and consider only skew-elliptical random vectors which possess densities. Let $g^{(d+1)}$ be a density generator for a $(d+1)$-dimensional elliptical random vector that satisfies

$$\int_0^\infty r^{(d+1)/2-1} g^{(d+1)}(r)\mathrm{d}r = \Gamma((d+1)/2)\pi^{-(d+1)/2},$$

then a $d$-dimensional random vector $\boldsymbol{X}$ has a skew-elliptical distribution with location parameter vector $\boldsymbol{\xi} \in \mathbb{R}^d$, positive-definite scale matrix $\Sigma \in \mathbb{R}^{d\times d}$, skewness parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^d$, and density generator $g^{(d+1)}$, if its density function is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{2}{\sqrt{|\Sigma|}} g^{(d)}\big((\boldsymbol{x}-\boldsymbol{\xi})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\big) G\big(\boldsymbol{\alpha}^\top \sigma^{-1}(\boldsymbol{x}-\boldsymbol{\xi}); g_{q(\boldsymbol{x})}\big), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{1}$$

where $\sigma = \mathrm{diag}(\Sigma)^{1/2} \in \mathbb{R}^{d\times d}$, $q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\xi})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\xi})$, $g^{(d)}$ is the $d$-dimensional marginal density generator induced by $g^{(d+1)}$, and $G(\cdot; g_{q(\boldsymbol{x})})$ is the cumulative distribution function of the univariate elliptical distribution with mean 0, scale 1, and conditional density generator $g_{q(\boldsymbol{x})}(s) = g^{(d+1)}(s + q(\boldsymbol{x}))/g^{(d)}(q(\boldsymbol{x}))$. We write $\boldsymbol{X} \sim \mathcal{SE}_d(\boldsymbol{\xi}, \Sigma, \boldsymbol{\alpha}, g^{(d+1)})$. When $\boldsymbol{\alpha} = \boldsymbol{0}$, the skew-elliptical distribution reduces to an elliptical distribution.

The skew-elliptical distribution has two stochastic representations, i.e., a convolution-type representation and a conditioning-type representation; see Equations (10) and (19) in Fang (2003). The former is useful for random sampling, and the latter allows us to express its cumulative distribution function in the following simple form

$$F(\boldsymbol{x}) = 2G_{d+1}(\boldsymbol{x}_* - \boldsymbol{\xi}_*; \Sigma_*, g^{(d+1)}), \tag{2}$$

with $\boldsymbol{x}_* = (0, \boldsymbol{x}^\top)^\top$, $\boldsymbol{\xi}_* = (0, \boldsymbol{\xi}^\top)^\top$ and

$$\Sigma_* = \begin{pmatrix} 1 & -\boldsymbol{\delta}^\top \sigma \\ -\sigma \boldsymbol{\delta} & \Sigma \end{pmatrix},$$

where $\sigma = \mathrm{diag}(\Sigma)^{1/2} \in \mathbb{R}^{d\times d}$, $\boldsymbol{\delta} = (1+\boldsymbol{\alpha}^\top \bar{\Sigma} \boldsymbol{\alpha})^{-1/2} \bar{\Sigma} \boldsymbol{\alpha}$ with $\bar{\Sigma}$ being the correlation matrix corresponding to $\Sigma$, i.e., $\Sigma = \sigma \bar{\Sigma} \sigma$, and $G_{d+1}(\boldsymbol{x}_* - \boldsymbol{\xi}_*; \Sigma_*, g^{(d+1)})$ denotes the cumulative distribution function of the $(d+1)$-variate elliptical distribution with location vector $\boldsymbol{\xi}_* \in \mathbb{R}^{d+1}$, positive-definite covariance matrix $\Sigma_* \in \mathbb{R}^{(d+1)\times(d+1)}$, and density generator $g^{(d+1)}$. The positive definiteness of $\Sigma_*$ implies that the admissible parameters of $(\Sigma, \alpha)$ are such that the matrix $\bar{\Sigma} - \boldsymbol{\delta}\boldsymbol{\delta}^\top$ is positive definite.

A prominent subclass of the skew-elliptical distribution is the skew-normal distribution (Azzalini, 1985; Azzalini & Dalla Valle, 1996). Specifically, when $g^{(d+1)}$ is the $(d+1)$-variate normal density generator, the density function of $\boldsymbol{X}$ is

$$f(\boldsymbol{x}) = 2\phi_d(\boldsymbol{x} - \boldsymbol{\xi}; \Sigma)\Phi\big(\boldsymbol{\alpha}^\top \sigma^{-1}(\boldsymbol{x} - \boldsymbol{\xi})\big), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

where $\phi_d(\boldsymbol{x} - \boldsymbol{\xi}; \Sigma)$ denotes the probability density function of the $d$-variate Gaussian distribution with mean vector $\boldsymbol{\xi}$ and covariance matrix $\Sigma$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. We denote this distribution as $\boldsymbol{X} \sim \mathcal{SN}_d(\boldsymbol{\xi}, \Sigma, \boldsymbol{\alpha})$. When $\boldsymbol{\alpha} = \mathbf{0}$, it reduces to the $d$-dimensional normal distribution, $\mathcal{N}_d(\boldsymbol{\xi}, \Sigma)$, and when $d = 1$ it coincides with the univariate skew-normal distribution (Azzalini, 1985).

When $g^{(d+1)}$ is the $d$-variate Student's $t$ density generator with $\nu$ degrees of freedom, we get another important subclass of the skew-elliptical distribution, i.e., the skew-$t$ distribution (Branco & Dey, 2001; Azzalini & Capitanio, 2003; Gupta, 2003). Its density has the following form

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = 2t_d(\boldsymbol{x} - \boldsymbol{\xi}; \Sigma, \nu) T\Big\{\boldsymbol{\alpha}^\top \sigma^{-1}(\boldsymbol{x} - \boldsymbol{\xi})\Big(\frac{\nu + p}{q(\boldsymbol{x}) + \nu}\Big)^{1/2}; \nu + d\Big\}, \quad \boldsymbol{x} \in \mathbb{R}^d,$$

where $t_d(\boldsymbol{x} - \boldsymbol{\xi}; \Sigma, \nu)$ denotes the probability density function of the $d$-variate t distribution with location vector $\boldsymbol{\xi}$, scale matrix $\Sigma$, and degrees of freedom $\nu$, $T(\cdot; \nu + d)$ denotes the univariate $t$ distribution function with degrees of freedom $\nu + d$. We write $\boldsymbol{X} \sim \mathcal{ST}_d(\boldsymbol{\xi}, \Sigma, \boldsymbol{\alpha}, \nu)$. When $\boldsymbol{\alpha} = \mathbf{0}$, it reduces to the $d$-dimensional Student's $t$ distribution, and when $\nu \to \infty$, it tends to the $d$-dimensional skew-normal distribution.

## 2.2 The Unified Skew-Elliptical Distribution

An extension of the skew-elliptical distribution is the unified skew elliptical distribution (Arellano-Valle & Genton, 2010), which aims to gain more flexibility by unifying various skew-elliptical families under the same model. Specifically, a $d$-dimensional random vector $\boldsymbol{X}$ has a unified skew elliptical distribution, denoted by $\boldsymbol{X} \sim \mathcal{SUE}_{d,m}(\boldsymbol{\xi}, \Sigma, \Lambda, \boldsymbol{\tau}, \Gamma, g^{(d+m)})$, if its density function is

$$f(\boldsymbol{x}) = \frac{g^{(d)}\big((\boldsymbol{x}-\boldsymbol{\xi})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\big)}{\sqrt{|\Sigma|}} \frac{G_m\big(\boldsymbol{\tau} + \Lambda\sigma^{-1}(\boldsymbol{x}-\boldsymbol{\xi}); \Gamma, g^{(m)}_{q(\boldsymbol{x})}\big)}{G_m(\boldsymbol{\tau}; \Gamma + \Lambda\bar{\Sigma}\Lambda^\top, g^{(m)})}, \quad \boldsymbol{x} \in \mathbb{R}^d,$$

where $q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\xi})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\xi})$, $g^{(d+m)}$ is a $(d+m)$-variate elliptical density generator, $g^{(d)}$ and $g^{(m)}$ are its $d$-variate and $m$-variate marginal density generators, respectively, $g^{(m)}_{q(\boldsymbol{x})}(s) = g^{(d+m)}\{s + q(\boldsymbol{x})\}/g^{(d)}\{q(\boldsymbol{x})\}$, $\boldsymbol{\xi} \in \mathbb{R}^d$ is a location parameter vector, $\boldsymbol{\tau} \in \mathbb{R}^d$ introduces additional flexibility to capture skewness, $\Gamma \in \mathbb{R}^{m\times m}$ is a correlation matrix, and $\Lambda \in \mathbb{R}^{m\times d}$ encompasses the main effect on the skewness. When $m = 1$, it reduces to the extended skew-elliptical distribution (Fang, 2003), and if we further have $\boldsymbol{\tau} = \mathbf{0}$, it reduces to the skew-elliptical distribution (1).

Similar to the skew-elliptical distribution, the unified skew elliptical distribution also has two special subclasses, i.e., the unified skew-normal distribution (Arellano-Valle & Azzalini, 2006) and the unified skew-$t$ distribution. When $g^{(d+m)}$ is the $(d+m)$-variate normal density generator, we get the unified skew-normal distribution with density

$$f(\boldsymbol{x}) = \phi_d(\boldsymbol{x}-\boldsymbol{\xi}; \Sigma)\frac{\Phi_m\big(\boldsymbol{\tau} + \Lambda\sigma^{-1}(\boldsymbol{x}-\boldsymbol{\xi}); \Gamma\big)}{\Phi_m(\boldsymbol{\tau}; \Gamma + \Lambda\bar{\Sigma}\Lambda^\top)}, \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{3}$$

where $\Phi_m(\cdot; \Gamma)$ denotes the centered $m$-dimensional normal distribution function with covariance matrix $\Gamma$. We write $\boldsymbol{X} \sim \mathcal{SUN}_{d,m}(\boldsymbol{\xi}, \Sigma, \Lambda, \boldsymbol{\tau}, \Gamma)$. The definition (3) is equivalent to the one in Arellano-Valle & Azzalini (2006) with a different parametrization. When $g^{(d+m)}$ is the $(d+m)$-variate Student's $t$ density generator with $\nu$ degrees of freedom, we get the unified skew-$t$ distribution with density

$$f(\boldsymbol{x}) = t_d(\boldsymbol{x}-\boldsymbol{\xi}; \Sigma, \nu)\frac{T_m\Big(\big(\boldsymbol{\tau} + \Lambda\sigma^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\big)\big(\frac{\nu+p}{q(\boldsymbol{x})+\nu}\big)^{1/2}; \Gamma, \nu+d\Big)}{T_m(\boldsymbol{\tau}; \Gamma + \Lambda\bar{\Sigma}\Lambda^\top, \nu)}, \quad \boldsymbol{x} \in \mathbb{R}^d,$$

where $T_m(\cdot; \Gamma, \nu+d)$ denotes the centered $m$-dimensional Student's t distribution function with dispersion matrix $\Gamma$ and degrees of freedom $\nu + d$. We write $\boldsymbol{X} \sim \mathcal{SUT}_{d,m}(\boldsymbol{\xi}, \Sigma, \Lambda, \nu, \boldsymbol{\tau}, \Gamma)$.

# 3 Posterior Inference for the Skew-Elliptical Link Model

## 3.1 The Skew-Elliptical Link Model

As discussed in Section 1, when modeling correlated binary data, the multivariate probit model uses a Gaussian link to capture dependence at the "latent level". A symmetric link, however, does not always provide the best fit to a given dataset, in particular for binary response data that are highly imbalanced.

In this section we extend the Gaussian link to the multivariate skew-elliptical link, which includes the skew-normal and skew-$t$ links as special cases. Specifically, let $Y_{ij}$ denote a binary 0/1 response on the $i$th observation of the $j$th variable and denote by $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{iM})^\top$ the collection of the $i$th observation on all $M$ variables for $i = 1, \dots, n$. Let $\boldsymbol{Y}_i^* = (Y_{i1}^*, \dots, Y_{iM}^*)^\top$ be a vector of latent variables capturing dependence among the components of $\boldsymbol{Y}_i$, $\boldsymbol{\beta} \in \mathbb{R}^p$ be a vector of regression coefficients, $X_i = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{iM})^\top \in \mathbb{R}^{M \times p}$ be the data matrix for the $i$th observation, and denote $X = (X_1^\top, \dots, X_n^\top)^\top \in \mathbb{R}^{nM \times p}$. Then the multivariate skew-elliptical link model can be expressed as

$$\begin{aligned} Y_{ij} &= \begin{cases} 1, & \text{if } Y_{ij}^* > 0, \\ 0, & \text{otherwise,} \end{cases} \\ \boldsymbol{Y}^* &= (\boldsymbol{Y}_1^{*\top}, \dots, \boldsymbol{Y}_n^{*\top})^\top = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (4) \\ \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\varepsilon} \end{pmatrix} \Bigg| \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)} &\sim \mathcal{SE}_{p+nM}\left( \begin{pmatrix} \boldsymbol{\mu} \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & 0 \\ 0 & \mathrm{I}_n \otimes \Sigma \end{pmatrix}, \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\alpha} \end{pmatrix}, g^{(p+nM+1)} \right), \end{aligned}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location parameter vector, $\Omega \in \mathbb{R}^{p \times p}$ is a positive-definite covariance matrix, $\mathrm{I}_n \in \mathbb{R}^{n \times n}$ is the identity matrix, $\otimes$ denotes the Kronecker product, $\Sigma \in \mathbb{R}^{M \times M}$ is a positive definite covariance matrix, $\boldsymbol{\alpha} \in \mathbb{R}^{nM}$ is a skewness parameter vector, and $g^{(p+nM+1)}$ is a $(p + nM + 1)$-variate elliptical density generator.

The multivariate probit model (Chib & Greenberg, 1998) assumes that the covariates are not shared by the $M$ variables $Y_{i1}, \dots, Y_{iM}$. In that case, $\boldsymbol{\beta}$ can be understood as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_M^\top)^\top$, where $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$ with $\sum_{j=1}^M p_j = p$ is the regression coefficients for the $j$-th variable $Y_{1j}, \dots, Y_{nj}$, and $\boldsymbol{x}_{ij}$ is understood as the vector $\boldsymbol{x}_{ij} = (\boldsymbol{x}_{ij1}^\top, \dots, \boldsymbol{x}_{ijM}^\top)^\top$ with $\boldsymbol{x}_{ijk} = \mathbf{0}$ for $k \neq j$, so that $\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} = \boldsymbol{x}_{ijj}^\top \boldsymbol{\beta}_j$. This notation of expanded vector $\boldsymbol{\beta}$ and $\boldsymbol{x}_{ij}$ simplifies the expression of our model (4).

To better understand the assumption on the joint distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ in the model (4), we express it in a different way. Using Proposition 2 in Fang (2003), an equivalent assumption is that

$$\boldsymbol{\beta} \mid g^{(p+nM+1)} \sim \mathcal{SE}_p(\boldsymbol{\mu}, \Omega, \mathbf{0}, g^{(p+1)}), \tag{5}$$

$$\boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)} \sim \mathcal{SE}_{nM}(\mathbf{0}, \mathrm{I}_n \otimes \Sigma, \boldsymbol{\alpha}, g_{q(\boldsymbol{\beta})}^{(nM+1)}), \tag{6}$$

where $q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\mu})^\top \Omega^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu})$, $g_{q(\boldsymbol{\beta})}^{(nM+1)}(s) = g^{(p+nM+1)}(s + q(\boldsymbol{\beta}))/g^{(p)}(q(\boldsymbol{\beta}))$, and $g^{(p)}$, $g^{(p+1)}$ are the $p$- and $(p+1)$-variate marginal density generators induced by the same generator $g^{(p+nM+1)}$, respectively. Assumption (5) may be understood as the prior for $\boldsymbol{\beta}$, while (6) is the distributional assumption for the latent data vector $\boldsymbol{Y}^*$. From (6) we observe that $\beta$ and $\varepsilon$ are dependent, but they are conditionally independent given $q(\beta)$. This weak dependence between them is broken when $g^{(p+nM+1)}$ is the normal density generator. Specifically, when $g^{(p+nM+1)}$ is the $(p + nM + 1)$-variate normal density generator, (5) becomes the typical Gaussian prior, $\mathcal{N}_p(\boldsymbol{\mu}, \Omega)$, and (6) becomes $\boldsymbol{\varepsilon} \mid \Sigma, \boldsymbol{\alpha} \sim \mathcal{SN}_{nM}(\mathbf{0}, \mathrm{I}_n \otimes \Sigma, \boldsymbol{\alpha})$, which is independent of $\boldsymbol{\beta}$ conditional on $\Sigma$ and $\boldsymbol{\alpha}$. If we further have $\boldsymbol{\alpha} = \mathbf{0}$, then (6) becomes $\boldsymbol{\varepsilon} \mid \Sigma \sim \mathcal{N}_{nM}(\mathbf{0}, \mathrm{I}_n \otimes \Sigma)$ and model (4) reduces to the well-known multivariate probit model (Ashford & Sowden, 1970; Chib & Greenberg, 1998) with a typical Gaussian prior for $\boldsymbol{\beta}$. By assuming a joint distribution for $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, we can gain two major advantages. The first is that we are able to account not only for the dependence between $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, but also for the dependence between the different observations $\boldsymbol{Y}_i, i = 1, \ldots, n$. The second is that this assumption allows us to get a tractable posterior for $\boldsymbol{\beta}$; see Section 3.2 for more details.

From (6) we know that the admissible parameters of $(\Sigma, \alpha)$ are these such that the matrix $\mathrm{I}_n \otimes \bar{\Sigma} - \boldsymbol{\delta}\boldsymbol{\delta}^\top$ is positive definite, where $\bar{\Sigma}$ is the correlation matrix corresponding to $\Sigma$ and $\boldsymbol{\delta} = \big(1 + \boldsymbol{\alpha}^\top(\mathrm{I}_n \otimes \bar{\Sigma})\boldsymbol{\alpha}\big)^{-1/2}(\mathrm{I}_n \otimes \bar{\Sigma})\boldsymbol{\alpha}$. From (4), the joint probability mass function of $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top)^\top = \boldsymbol{y}$, given all the parameters and the data matrix $X$, is

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) = \int_{A_{nM}} \cdots \int_{A_{11}} \frac{2}{|\mathrm{I}_n \otimes \Sigma|^{1/2}} g^{q(\boldsymbol{\beta}),nM}\big(\boldsymbol{t}^\top(\mathrm{I}_n \otimes \Sigma^{-1})\boldsymbol{t}\big) G(\boldsymbol{\alpha}^\top \boldsymbol{t}; g_{q(\boldsymbol{t})}^{q(\boldsymbol{\beta})}) \mathrm{d}\boldsymbol{t}, \tag{7}$$

where $q(\boldsymbol{t}) = \boldsymbol{t}^\top(\mathrm{I}_n \otimes \Sigma^{-1})\boldsymbol{t}$, $g_{q(\boldsymbol{t})}^{q(\boldsymbol{\beta})}(s) = g_{q(\boldsymbol{\beta})}^{(nM+1)}(s + q(\boldsymbol{t}))/g^{q(\boldsymbol{\beta}),nM}(q(\boldsymbol{t}))$, $g^{q(\boldsymbol{\beta}),nM}$ is the $nM$-variate marginal density generator induced by $g_{q(\boldsymbol{\beta})}^{(nM+1)}$, and $A_{ij}, i = 1, \ldots, n, j = 1, \ldots, M$ is the interval

$$A_{ij} = \begin{cases} (-\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}, \infty), & \text{if } y_{ij} = 1, \\ (-\infty, \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}], & \text{if } y_{ij} = 0. \end{cases}$$

Although the joint probability (7) involves multidimensional integration over a constrained space, we show in the following section that it can be substantially simplified.

## 3.2 Unified Skew Elliptical Posterior for the Regression Coefficients

In this section we prove that for the multivariate skew-elliptical link model (4), the regression coefficients parameter $\boldsymbol{\beta}$ has a unified skew elliptical posterior. To prove this result, we first simplify the joint probability mass function $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)})$ of the observed data in the following lemma.

**Lemma 1.** *The joint probability mass function $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)})$ based on (4) can be simplified to*

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) = 2G_{nM+1}(D_*\boldsymbol{\beta}; \Sigma_*, g^{(nM+1)}_{q(\boldsymbol{\beta})}),$$

*where $D = \operatorname{diag}(2\boldsymbol{y} - \mathbf{1}_{nM}) \in \mathbb{R}^{nM\times nM}$ with $\mathbf{1}_{nM} \in \mathbb{R}^{nM}$ being the vector of* 1*s, $D_* = \big(\mathbf{0}_p, (DX)^\top\big)^\top \in \mathbb{R}^{(nM+1)\times p}$, $\mathbf{0}_p \in \mathbb{R}^p$ is a vector of* 0*s, and*

$$\Sigma_* = \begin{pmatrix} 1 & -\boldsymbol{\delta}^\top D(\mathrm{I}_n \otimes \sigma) \\ -(\mathrm{I}_n \otimes \sigma)D\boldsymbol{\delta} & D(\mathrm{I}_n \otimes \Sigma)D \end{pmatrix} \in \mathbb{R}^{(nM+1)\times(nM+1)}$$

*with $\boldsymbol{\delta} \in \mathbb{R}^{nM}, \boldsymbol{\delta} = \big(1 + \boldsymbol{\alpha}^\top(\mathrm{I}_n \otimes \bar{\Sigma})\boldsymbol{\alpha}\big)^{-1/2}(\mathrm{I}_n \otimes \bar{\Sigma})\boldsymbol{\alpha}$, $\sigma = \operatorname{diag}(\Sigma)^{1/2} \in \mathbb{R}^{d\times d}$ and $\bar{\Sigma}$ being the correlation matrix corresponding to $\Sigma$, i.e., $\Sigma = \sigma\bar{\Sigma}\sigma$.*

*Proof.* Since a diagonal matrix $\operatorname{diag}(\boldsymbol{x})$ with $\boldsymbol{x} \in \{-1, 1\}^{nM}$ has the property

$$\operatorname{diag}(\boldsymbol{x})\boldsymbol{x} = \mathbf{1}_{nM}, \text{ and } \big(\operatorname{diag}(\boldsymbol{x})\big)^{-1} = \operatorname{diag}(\boldsymbol{x}),$$

we have

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) &= \Pr(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) \\ &= \Pr(2\boldsymbol{Y} - \mathbf{1}_{nM} = 2\boldsymbol{y} - \mathbf{1}_{nM} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) \\ &= \Pr\big(D(2\boldsymbol{Y} - \mathbf{1}_{nM}) = \mathbf{1}_{nM} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}\big) \\ &= \Pr\big(D\boldsymbol{Y}^* > \mathbf{0} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}\big) \\ &= \Pr\big(-D\boldsymbol{\varepsilon} - DX\boldsymbol{\beta} < \mathbf{0} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}\big). \end{aligned}$$

By (6), $\boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)} \sim \mathcal{SE}_{nM}(\mathbf{0}, \mathrm{I}_n \otimes \Sigma, \boldsymbol{\alpha}, g^{(nM+1)}_{q(\boldsymbol{\beta})})$. Using Proposition 1 in Fang (2003), we know that

$$(-D\boldsymbol{\varepsilon} - DX\boldsymbol{\beta}) \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)} \sim \mathcal{SE}_{nM}(-DX\boldsymbol{\beta}, D(\mathrm{I}_n \otimes \Sigma)D, D\boldsymbol{\alpha}, g^{(nM+1)}_{q(\boldsymbol{\beta})}).$$

Using (2), we finally get

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) = 2G_{nM+1}(D_*\boldsymbol{\beta}; \Sigma_*, g^{(nM+1)}_{q(\boldsymbol{\beta})}).$$

□

Now we are ready to present our main result that the posterior distribution of $\boldsymbol{\beta}$ coincides with a unified skew elliptical distribution.

**Theorem 1.** *Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_n^\top)^\top$ be observations from the multivariate skew-elliptical link model (4) and $X = (X_1^\top, \dots, X_n^\top)^\top$ be the corresponding data matrix. Then*

$$(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) \sim \mathcal{SUE}_{p,nM+1}(\boldsymbol{\mu}_{post}, \Omega_{post}, \Lambda_{post}, \boldsymbol{\tau}_{post}, \Gamma_{post}, g^{(p+nM+1)}),$$

*with posterior parameters*

$$\boldsymbol{\mu}_{post} = \boldsymbol{\mu},\ \Omega_{post} = \Omega,\ \Lambda_{post} = \sigma_*^{-1} D_* \omega,\ \boldsymbol{\tau}_{post} = \sigma_*^{-1} D_* \boldsymbol{\mu},\ \Gamma_{post} = \bar{\Sigma}_*,$$

*where $D_* \in \mathbb{R}^{(nM+1)\times p}$ and $\Sigma_* \in \mathbb{R}^{(nM+1)\times(nM+1)}$ are the matrices defined in Lemma 1, $\sigma_* = \mathrm{diag}(\Sigma_*)^{1/2} \in \mathbb{R}^{(nM+1)\times(nM+1)}$, $\bar{\Sigma}_*$ is the correlation matrix corresponding to $\Sigma_*$, i.e., $\Sigma_* = \sigma_* \bar{\Sigma}_* \sigma_*$, and $\omega = \mathrm{diag}(\Omega)^{1/2} \in \mathbb{R}^{p\times p}$.*

*Proof.* The posterior density of the coefficients $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) \cdot p(\boldsymbol{\beta} \mid g^{(p+nM+1)}).$$

Using Lemma 1 and the assumption (5), we have

$$\begin{aligned}
&p(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) \\
&\propto G_{nM+1}(D_*\boldsymbol{\beta}; \Sigma_*, g^{(nM+1)}_{q(\boldsymbol{\beta})}) \cdot g^{(p)}\big((\boldsymbol{\beta}-\boldsymbol{\mu})^\top \Omega^{-1} (\boldsymbol{\beta}-\boldsymbol{\mu})\big) \\
&= G_{nM+1}(\sigma_*^{-1} D_*\boldsymbol{\beta}; \bar{\Sigma}_*, g^{(nM+1)}_{q(\boldsymbol{\beta})}) \cdot g^{(p)}\big((\boldsymbol{\beta}-\boldsymbol{\mu})^\top \Omega^{-1} (\boldsymbol{\beta}-\boldsymbol{\mu})\big) \\
&= G_{nM+1}\big(\sigma_*^{-1} D_*\boldsymbol{\mu} + \sigma_*^{-1} D_*(\boldsymbol{\beta}-\boldsymbol{\mu}); \bar{\Sigma}_*, g^{(nM+1)}_{q(\boldsymbol{\beta})}\big) \cdot g^{(p)}\big((\boldsymbol{\beta}-\boldsymbol{\mu})^\top \Omega^{-1} (\boldsymbol{\beta}-\boldsymbol{\mu})\big) \\
&= G_{nM+1}\big(\sigma_*^{-1} D_*\boldsymbol{\mu} + \sigma_*^{-1} D_*\omega\omega^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}); \bar{\Sigma}_*, g^{(nM+1)}_{q(\boldsymbol{\beta})}\big) \cdot g^{(p)}\big((\boldsymbol{\beta}-\boldsymbol{\mu})^\top \Omega^{-1} (\boldsymbol{\beta}-\boldsymbol{\mu})\big) \\
&= G_{nM+1}(\boldsymbol{\tau}_{\mathrm{post}} + \Lambda_{\mathrm{post}}\omega^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}); \bar{\Sigma}_*, g^{(nM+1)}_{q(\boldsymbol{\beta})}) \cdot g^{(p)}\big((\boldsymbol{\beta}-\boldsymbol{\mu})^\top \Omega^{-1} (\boldsymbol{\beta}-\boldsymbol{\mu})\big).
\end{aligned}$$

Hence, $(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}, g^{(p+nM+1)}) \sim \mathcal{SUE}_{p,nM+1}(\boldsymbol{\mu}_{post}, \Omega_{post}, \Lambda_{post}, \boldsymbol{\tau}_{post}, \Gamma_{post}, g^{(p+nM+1)})$. □

In Bayesian regression we are mostly interested in the posterior marginals, their moments and more complex functionals such as measures of dependence and credible intervals. Thanks to the fundamental property of the unified skew elliptical distribution that it is closed under marginalization, conditioning and affine transformations, this type of inference is simplified. We refer to Arellano-Valle & Genton (2010) for details on how to obtain the parameters of the marginal distribution, conditional distribution and the distribution after affine transformations. As for the calculation of the posterior moments and credible intervals, numerical integration of the marginal posterior densities can be used. When interest is in the posterior moments, another approach is to use the moment generating function. We refer to Section 5 of Arellano-Valle & Genton (2010) for derivations of the moment generating function and moments of the unified skew elliptical distribution.

### 3.3 Special Case 1: the Skew-Normal Link Model

The skew-normal link model is obtained when $g^{(p+nM+1)}$ in model (4) is the $(p+nM+1)$-variate normal density generator. In this case, the joint distributional assumption of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ becomes

$$\left( \begin{array}{c} \boldsymbol{\beta} \\ \boldsymbol{\varepsilon} \end{array} \right) \Bigg| \Sigma, \boldsymbol{\alpha} \sim \mathcal{SN}_{p+nM} \left( \left( \begin{array}{c} \boldsymbol{\mu} \\ 0 \end{array} \right), \left( \begin{array}{cc} \Omega & 0 \\ 0 & I_n \otimes \Sigma \end{array} \right), \left( \begin{array}{c} \mathbf{0} \\ \boldsymbol{\alpha} \end{array} \right) \right),$$

which is equivalent to assuming

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Omega), \quad \boldsymbol{\varepsilon} \mid \Sigma, \boldsymbol{\alpha} \sim \mathcal{SN}_{nM}(\mathbf{0}, \mathrm{I}_n \otimes \Sigma, \boldsymbol{\alpha}),$$

with the random vectors $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ independent of each other given $\Sigma$ and $\boldsymbol{\alpha}$. This implies that the prior for $\boldsymbol{\beta}$ coincides with the typical weakly informative Gaussian prior, and we use a multivariate SN distribution to model the dependence of the data at the latent level. When the skewness parameter $\boldsymbol{\alpha} = \mathbf{0}$, the skew-normal link model reduces to the well-known multivariate probit model (Ashford & Sowden, 1970; Chib & Greenberg, 1998).

Before analyzing the posterior of the regression coefficients, we first give the explicit expression of the joint probability mass function $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha})$. Taking $g^{(nM+1)}$ in Lemma 1 as the normal density generator, we directly get

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}) = 2\Phi_{nM+1}(D_* \boldsymbol{\beta}; \Sigma_*),$$

where $D_*$ and $\Sigma_*$ are defined in Lemma 1. Similarly to the multivariate probit model, if the covariates are not shared by all the responses, the matrix $\Sigma$ has to be a correlation matrix for identifiability reasons. This can be seen by considering $\Omega = (\omega_{jm}) = C\Sigma C^\top$ with $C = \mathrm{diag}(\Omega)^{1/2}$, $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^\top, \ldots, \tilde{\boldsymbol{\beta}}_M^\top)^\top$ with $\tilde{\boldsymbol{\beta}}_j = \omega_{jj} \boldsymbol{\beta}_j$, and $\tilde{\boldsymbol{\alpha}} = \mathrm{diag}(\mathrm{I}_n \otimes C^{-1})\boldsymbol{\alpha}$, which gives $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}) = p(\boldsymbol{y} \mid \tilde{\boldsymbol{\beta}}, \Omega, \tilde{\boldsymbol{\alpha}})$. If the covariates are shared by all the responses $\boldsymbol{Y}$, then $\Sigma$ does not need to be a correlation matrix. However, if we assume that all the diagonal entries in $\Sigma$ are equal, then $\Sigma$ has to be a correlation matrix because $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}) = p(\boldsymbol{y} \mid b\boldsymbol{\beta}, b^2\Sigma, \boldsymbol{\alpha})$ for any positive number $b$.

When $\boldsymbol{\alpha} = \mathbf{0}$, we get $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma) = \Phi_{nM}(DX\boldsymbol{\beta}; \mathrm{I}_n \otimes \Sigma)$. This result simplifies the calculation of the joint probability of the multivariate skew-normal link model and also the probit model by expressing it in terms of the multivariate normal distribution function. Therefore, existing fast algorithms for the calculation of the multivariate normal probabilities can be utilized especially in high dimensions; see Genton et al. (2018) and Cao et al. (2020b). Now we present the result that for the skew-normal link model the posterior of $\boldsymbol{\beta}$ coincides with a unified skew normal distribution, which directly follows from Theorem 1 by taking $g^{(nM+1)}$ as the $(nM + 1)$-variate normal density generator.

**Corollary 1.** *Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_n^\top)^\top$ be observations from the multivariate skew-normal link model and $X = (X_1^\top, \ldots, X_n^\top)^\top$ be the corresponding data matrix. Then*

$$(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}) \sim \mathcal{SUN}_{p,nM+1}(\boldsymbol{\mu}_{post}, \Omega_{post}, \Lambda_{post}, \boldsymbol{\tau}_{post}, \Gamma_{post}),$$

*where $\boldsymbol{\mu}_{post}, \Omega_{post}, \Lambda_{post}, \boldsymbol{\tau}_{post}, \Gamma_{post}$ are defined in Theorem 1.*

The unified skew normal distribution, a subclass of the unified skew elliptical family, is closed under marginalization, conditioning and affine transformations (Arellano-Valle & Azzalini, 2006; Arellano-Valle & Genton, 2010). This property is useful for certain posterior inferences, such as the posterior marginals or their moments. When interest is in sampling from the posterior distribution, the convolution-type stochastic representation of the unified skew normal random vector is very useful. Specifically, using Equation (8) in Arellano-Valle & Genton (2010), $(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha})$ has the following stochastic representation

$$(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}) \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \boldsymbol{V}_0 + \Omega D_*^\top (D_* \Omega D_*^\top + \Sigma_*)^{-1} s \boldsymbol{V}_1,$$

where $\overset{\mathrm{d}}{=}$ means equality in distribution, $s = \mathrm{diag}(D_* \Omega D_*^\top + \Sigma_*)^{1/2} \in \mathbb{R}^{(nM+1)\times(nM+1)}$, $\boldsymbol{V}_0 \sim \mathcal{N}_p\big(\mathbf{0}, \Omega - \Omega D_*^\top (D_* \Omega D_*^\top + \Sigma_*)^{-1} D_* \Omega\big)$ is independent of $\boldsymbol{V}_1$, which follows a $(nM + 1)$-variate truncated normal distribution with location parameter $\mathbf{0}$, covariance matrix $s^{-1}(D_* \Omega D_*^\top + \Sigma_*) s^{-1}$ and truncated below the level $-s^{-1} D_* \boldsymbol{\mu}$. This stochastic representation facilitates exact simulation from the posterior distribution; see Algorithm 1 of Durante (2019).

## 3.4 Special Case 2: the Skew-$t$ Link Model

When $g^{(p+nM+1)}$ in model (4) is the $(p + nM + 1)$-variate Student's $t$ density generator with $\nu$ degrees of freedom, we get the skew-$t$ link model. Specifically, the joint distributional assumption of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ is

$$\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\varepsilon} \end{pmatrix} \Bigg| \Sigma, \boldsymbol{\alpha}, \nu \sim \mathcal{ST}_{p+nM} \left( \begin{pmatrix} \boldsymbol{\mu} \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & 0 \\ 0 & I_n \otimes \Sigma \end{pmatrix}, \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\alpha} \end{pmatrix}, \nu \right),$$

which is equivalent to assuming

$$\begin{aligned}\boldsymbol{\beta} \mid \nu &\sim \mathcal{T}_p(\boldsymbol{\mu}, \Omega, \nu),\\ \boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, \nu &\sim \mathcal{ST}_{nM}\left(\mathbf{0}, \frac{\nu + (\boldsymbol{\beta} - \boldsymbol{\mu})^\top \Omega^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu})}{\nu + p} \big(\mathrm{I}_n \otimes \Sigma\big), \boldsymbol{\alpha}, \nu + p\right),\end{aligned}$$

where $\mathcal{T}_p(\boldsymbol{\mu}, \Omega, \nu)$ denotes the Student's $t$ distribution with location parameter vector $\boldsymbol{\mu}$, dispersion matrix $\Omega$ and degrees of freedom $\nu$. The nonnegative parameter $\nu$ can be considered as a hyperparameter which controls the dependence between $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$. As $\nu$ increases the dependence decreases, and when $\nu \to \infty$, the skew-$t$ link model tends to the skew-normal link model and the dependence between them vanishes.

By taking $g^{(nM+1)}$ in Lemma 1 as the Student's $t$ density generator with $\nu$ degrees of freedom, we get the following explicit expression of the joint probability

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, \nu) = 2T_{nM+1}\left(\left(\frac{\nu + p}{\nu + (\boldsymbol{\beta} - \boldsymbol{\mu})^\top \Omega^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu})}\right)^{1/2} D_* \boldsymbol{\beta}; \Sigma_*, \nu + p\right). \tag{8}$$

In practice, we typically assume a weakly informative prior for $\boldsymbol{\beta}$, which means $\nu$ is often large and $\Omega$ is often taken as a diagonal matrix with large diagonal entries. This implies that $(\boldsymbol{\beta} - \boldsymbol{\mu})^\top \Omega^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu})$ is often very small compared to $\nu$ and $\nu \approx \nu + (\boldsymbol{\beta} - \boldsymbol{\mu})^\top \Omega^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu})$. Hence, if we assume that the diagonal entries of $\Sigma$ are all equal, then $\Sigma$ needs to be a correlation matrix because $p(\boldsymbol{y} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}, \nu) \approx p(\boldsymbol{y} \mid b\boldsymbol{\beta}, b^2\Sigma, \boldsymbol{\alpha}, \nu)$ for any positive number $b$. We now state the result that for the skew-$t$ link model the posterior of $\boldsymbol{\beta}$ coincides with a unified skew $t$ distribution, which directly follows from Theorem 1 by taking $g^{(nM+1)}$ as the $(nM + 1)$-variate Student's $t$ density generator with $\nu$ degrees of freedom.

**Corollary 2.** *Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_n^\top)^\top$ be observations from the multivariate skew-t link model and $X = (X_1^\top, \dots, X_n^\top)^\top$ be the corresponding data matrix. Then*

$$(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}, \nu) \sim \mathcal{SUT}_{p,nM+1}(\boldsymbol{\mu}_{post}, \Omega_{post}, \Lambda_{post}, \nu, \boldsymbol{\tau}_{post}, \Gamma_{post}),$$

*where $\boldsymbol{\mu}_{post}, \Omega_{post}, \Lambda_{post}, \boldsymbol{\tau}_{post}, \Gamma_{post}$ are defined in Theorem 1.*

Similarly to the unified skew normal distribution, the unified skew $t$ distribution is also closed under marginalization, conditioning and affine transformations (Arellano-Valle & Genton, 2010), which simplifies the inference of the posterior marginals, their moments and functionals such as measures of dependence and credible intervals. Thanks to the stochastic representation of the unified skew $t$

distribution, exact sampling from the distribution of $(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}, \nu)$ is also feasible. Specifically, using Equation (9) in Arellano-Valle & Genton (2010), $(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}, \nu)$ has the stochastic representation

$$(\boldsymbol{\beta} \mid \boldsymbol{y}, \Sigma, \boldsymbol{\alpha}, \nu) \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \left( \frac{\nu + \boldsymbol{U}_1^\top s (D_* \Omega D_*^\top + \Sigma_*)^{-1} s \boldsymbol{U}_1}{\nu + nM + 1} \right)^{1/2} \boldsymbol{U}_0 + \Omega D_*^\top (D_* \Omega D_*^\top + \Sigma_*)^{-1} s \boldsymbol{U}_1, \quad (9)$$

where $s = \mathrm{diag}(D_* \Omega D_*^\top + \Sigma_*)^{1/2} \in \mathbb{R}^{(nM+1)\times(nM+1)}$, $\boldsymbol{U}_0 \sim \mathcal{T}_p\big(\mathbf{0}, \Omega - \Omega D_*^\top (D_* \Omega D_*^\top + \Sigma_*)^{-1} D_* \Omega, \nu + nM + 1\big)$ is independent of $\boldsymbol{U}_1$, which follows a $(nM+1)$-variate truncated $t$ distribution with location parameter vector $\mathbf{0}$, dispersion matrix $s^{-1}(D_* \Omega D_*^\top + \Sigma_*) s^{-1}$, degrees of freedom $\nu$, and truncated below the level $-s^{-1} D_* \boldsymbol{\mu}$.

# 4 Simulation and Empirical Studies

## 4.1 Prior and Posterior for $\alpha$ and $\Sigma$

As the skew-normal link model is a limiting case of the skew-$t$ link model when the degrees of freedom $\nu$ tends to $\infty$, in this section we focus on the skew-$t$ link model and perform a simulation study and a real-data application. To make the model parsimonious, in both the simulation study and empirical study we assume that the skewness parameters are the same across different observations, i.e., $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M, \ldots, \alpha_1, \ldots, \alpha_M)^\top \in \mathbb{R}^{nM}$, and $\Sigma$ is a correlation matrix, i.e., $\Sigma = \bar{\Sigma}$. The assumption of a correlation matrix for $\Sigma$ is not very restrictive because it is approximately equivalent to assuming that all the diagonal entries in $\Sigma$ are equal, as we discussed in Section 3.4. Now we specify the prior and posterior for the skewness parameter $\boldsymbol{\alpha}_s = (\alpha_1, \ldots, \alpha_M)^\top$ and the correlation matrix $\bar{\Sigma}$.

Bayesian modeling of unstructured covariance or correlation matrices is a fundamental and difficult task because of the constraint of positive definiteness and the quadratic increase of the number of parameters with respect to the number of correlated variables. More importantly, it is difficult to specify a prior for them(Gelman et al., 2014). Typical priors for correlation matrices include the marginally uniform prior, the jointly uniform prior (Barnard et al., 2000) and the so-called LKJ prior (Lewandowski et al., 2009).

The marginally uniform prior means that each non-diagonal element in the correlation matrix has a uniform marginal distribution over $[-1, 1]$, whereas the jointly uniform prior means that the correlation matrix has a joint uniform distribution over the compact space of valid correlation matrices. The LKJ prior is recommended in the `R` library `rstan` (R Core Team, 2020) and has the form $\pi(\bar{\Sigma}) \propto |\bar{\Sigma}|^{\eta-1}$, where $|\bar{\Sigma}|$ is the determinant of $\bar{\Sigma}$ and $\eta > 0$ is the shape parameter of the LKJ distribution. The jointly uniform prior is a special case of the LKJ prior when $\eta = 1$.

In this work we adopt the jointly uniform prior for $\bar{\Sigma}$ by setting $\eta = 1$ in the LKJ prior and specify an independent weakly informative Gaussian prior for $\boldsymbol{\alpha}_s$. Then, using Equation (8), the joint posterior of $(\bar{\Sigma}, \boldsymbol{\alpha}_s)$ given the data and the regression coefficients is

$$p(\bar{\Sigma}, \boldsymbol{\alpha}_s \mid \boldsymbol{y}, \boldsymbol{\beta}, \nu) \propto 2T_{nM+1}\left(\left(\frac{\nu + p}{\nu + (\boldsymbol{\beta} - \boldsymbol{\mu})^\top \Omega^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu})}\right)^{1/2} D_*\boldsymbol{\beta}; \Sigma_*, \nu + p\right)\pi(\boldsymbol{\alpha}_s). \qquad (10)$$

We evaluate the multivariate Student's $t$ probability on the right-hand side of (10) using the R library `tlrmvnmvt`, which implements the classic Genz algorithm (Genz & Bretz, 1999, 2002) and exploits a tile-low-rank algorithm (Cao et al., 2020b) to speed up the computation of the multivariate normal and $t$ probabilities. To avoid sampling the correlation matrix from a constrained space, we consider the reparametrization adopted in Smith (2013) and Chin et al. (2020), which re-expresses a correlation matrix in terms of the Cholesky factor of a positive definite matrix$\bar{\Sigma} = \Lambda_{\bar{\Sigma}}^{-1/2} L_{\bar{\Sigma}} L_{\bar{\Sigma}}^\top \Lambda_{\bar{\Sigma}}^{-1/2}$, where $L_{\bar{\Sigma}}$ is a lower triangular matrix and $\Lambda_{\bar{\Sigma}} = \operatorname{diag}(L_{\bar{\Sigma}} L_{\bar{\Sigma}}^\top)$. Here the diagonal entries of $L_{\bar{\Sigma}}$ are set to 1 such that the correspondence between $L_{\bar{\Sigma}}$ and $\bar{\Sigma}$ is one-to-one. We denote the collection of the $M(M-1)/2$ unconstrained parameters in $L_{\bar{\Sigma}} = (l_{ij})$ by $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta} = \{l_{ij} : i > j, i, j = 1, \ldots, M\}$, and the $M(M-1)/2$ constrained parameters in $\bar{\Sigma}$ by $\operatorname{vec}(\bar{\Sigma})$, then using a change of variables we get the posterior of $(\boldsymbol{\theta}, \boldsymbol{\alpha}_s)$ as

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}_s \mid \boldsymbol{y}, \boldsymbol{\beta}, \nu) = p(\bar{\Sigma}, \boldsymbol{\alpha}_s \mid \boldsymbol{y}, \boldsymbol{\beta}, \nu)|J| = p(\bar{\Sigma}, \boldsymbol{\alpha}_s \mid \boldsymbol{y}, \boldsymbol{\beta}, \nu)\prod_{i=1}^{M}\Big(1 + \sum_{j<i} l_{ij}^2\Big)^{-(M+1)/2},$$

where $|J| = |\partial \operatorname{vec}(\bar{\Sigma})/\partial \boldsymbol{\theta}|$ is the determinant of the Jacobian matrix of this transformation.

As direct sampling from the distribution of $\boldsymbol{\theta}, \boldsymbol{\alpha}_s \mid \boldsymbol{y}, \boldsymbol{\beta}, \nu$ is unknown, we propose to use a random walk Metropolis-Hastings algorithm to generate samples from it. Specifically, we first sample $\boldsymbol{\alpha}_s'$ from a proposal distribution with density $q(\cdot \mid \boldsymbol{\alpha}_s)$ and $\boldsymbol{\theta}'$ from a proposal distribution with density $r(\cdot \mid \boldsymbol{\theta})$. Here we take both proposal densities $q$ and $r$ as symmetric normal densities, i.e., $\boldsymbol{\alpha}_s' \mid \boldsymbol{\alpha}_s \sim \mathcal{N}_M(\boldsymbol{\alpha}_s, h_1 \mathrm{I}_M)$ and $\boldsymbol{\theta}' \mid \boldsymbol{\theta} \sim \mathcal{N}_J(\boldsymbol{\theta}, h_2 \mathrm{I}_J)$, $J = M(M-1)/2$. Then the acceptance probability is

$$\alpha((\boldsymbol{\alpha}_s, \boldsymbol{\theta}), (\boldsymbol{\alpha}_s', \boldsymbol{\theta}')) = \min\left\{\frac{p(\boldsymbol{\theta}', \boldsymbol{\alpha}_s' \mid \boldsymbol{y}, X, \boldsymbol{\beta})1((\boldsymbol{\theta}', \boldsymbol{\alpha}_s') \in C)}{p(\boldsymbol{\theta}, \boldsymbol{\alpha}_s \mid \boldsymbol{y}, X, \boldsymbol{\beta})1((\boldsymbol{\theta}, \boldsymbol{\alpha}_s) \in C)}, 1\right\},$$

where $1(\cdot)$ is the indicator function and $C$ is the space of all $(\boldsymbol{\theta}, \boldsymbol{\alpha}_s)$ such that the resulting matrix $\bar{\Sigma} - \boldsymbol{\delta}\boldsymbol{\delta}^\top$ is positive definite with $\boldsymbol{\delta} = (1 + \boldsymbol{\alpha}^\top \bar{\Sigma} \boldsymbol{\alpha})^{-1/2} \bar{\Sigma} \boldsymbol{\alpha}$.

## 4.2 MCMC Sampling Scheme

As sampling from the distribution of $(\boldsymbol{\beta} \mid \boldsymbol{y}, X, \bar{\Sigma}, \boldsymbol{\alpha})$ is feasible using (9) and sampling from the distribution of $(\bar{\Sigma}, \boldsymbol{\alpha} \mid \boldsymbol{y}, X, \boldsymbol{\beta})$ has been described in Section 4.1, we now combine them to construct an MCMC sampler for the multivariate skew-$t$ link model.

**Algorithm 1:** MCMC sampling scheme for the multivariate ST link model

```
Initialization: Set $\boldsymbol{\beta}^{(0)}, \bar{\Sigma}^{(0)}, \boldsymbol{\alpha}^{(0)}$ ;
for iteration $k$ from 1 to $K$ do
    [1] Sample $\boldsymbol{U}_0^{(k)}$ from $\mathcal{T}_p\big(\mathbf{0}, \Omega - \Omega D_*^\top (D_* \Omega D_*^\top + \Sigma_*)^{-1} D_* \Omega, \nu + nM + 1\big)$ (in R use rmvt);
    [2] Sample $\boldsymbol{U}_1^{(k)}$ from a $(nM+1)$-variate truncated $t$ distribution with location parameter
      vector $\mathbf{0}$, dispersion matrix $s^{-1}(D_* \Omega D_*^\top + \Sigma_*) s^{-1}$, degrees of freedom $\nu$, and truncated
      below the level $-s^{-1} D_* \boldsymbol{\mu}$, using the accept-reject algorithm of Botev (2017) (in R use
      mvrandt);
    [3] Compute $\boldsymbol{\beta}^{(k)}$ via
      $\boldsymbol{\beta}^{(k)} = \boldsymbol{\mu} + \big(\frac{\nu + (\boldsymbol{U}_1^{(k)})^\top s (D_* \Omega D_*^\top + \Sigma_*)^{-1} s \boldsymbol{U}_1^{(k)}}{\nu + nM + 1}\big)^{1/2} \boldsymbol{U}_0^{(k)} + \Omega D_*^\top (D_* \Omega D_*^\top + \Sigma_*)^{-1} s \boldsymbol{U}_1^{(k)}$;
    [4] Use the Metropolis-Hastings algorithm described in Section 4.1 to sample $(\boldsymbol{\theta}^{(k)}, \boldsymbol{\alpha}_s^{(k)})$
      from the distribution of $(\boldsymbol{\theta}, \boldsymbol{\alpha}_s \mid \boldsymbol{y}, X, \boldsymbol{\beta}^{(k)})$, then return the resulting $\bar{\Sigma}^{(k)}$ and $\boldsymbol{\alpha}^{(k)}$.
Output: $(\boldsymbol{\beta}^{(1)}, \bar{\Sigma}^{(1)}, \boldsymbol{\alpha}^{(1)}), \ldots, (\boldsymbol{\beta}^{(K)}, \bar{\Sigma}^{(K)}, \boldsymbol{\alpha}^{(K)})$
```

## 4.3 Simulation Study

In this section, we conduct a simulation study to assess the performance of our proposed Algorithm 1. We consider three different scenarios with different values for the degrees of freedom, i.e., $\nu = 5, 10, 20$. In each of the scenarios, we generate a dataset with sample size $n = 50$, regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top = (-1, 0.5, -0.5)^\top$, skewness parameter $\boldsymbol{\alpha}_s = (\alpha_1, \alpha_2, \alpha_3)^\top = (2, 0, -2)^\top$, and dispersion matrix

$$\bar{\Sigma} = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & -0.5 \\ 0 & -0.5 & 1 \end{pmatrix}.$$

The first column of the data matrix $X$ is set to $\mathbf{1}$ to account for the intercept and the remaining entries in $X$ are generated from a standard normal distribution. The intercept $\beta_1$ is chosen as $-1$ so as to obtain a highly-imbalanced dataset $\boldsymbol{y}$ with more than 80% of the observations being equal to 0.

For each of the scenarios, we fix $\nu$ to its true value, and then run Algorithm 1 for 10000 iterations,

Table 1: Posterior estimates for different scenarios in the simulation study

| Scenario | | $\nu = 5$ | | | $\nu = 10$ | | | $\nu = 20$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | True | Est | Sd | 95% CI | Est | Sd | 95% CI | Est | Sd | 95% CI |
| $\beta_1$ | -1 | -1.54 | 0.69 | (-3.26, -0.75) | -1.33 | 0.39 | (-2.27, -0.77) | -1.24 | 0.27 | (-1.86, -0.80) |
| $\beta_2$ | 0.5 | 0.78 | 0.39 | (0.31, 1.75) | 0.55 | 0.22 | (0.21, 1.06) | 0.52 | 0.18 | (0.21, 0.92) |
| $\beta_3$ | -0.5 | -0.46 | 0.28 | (-1.12, -0.10) | -0.76 | 0.27 | (-1.41, -0.35) | -0.71 | 0.21 | (-1.19, -0.37) |
| $\bar{\Sigma}_{12}$ | 0.5 | 0.33 | 0.31 | (-0.35, 0.81) | 0.42 | 0.28 | (-0.18, 0.86) | 0.69 | 0.17 | (0.28, 0.94) |
| $\bar{\Sigma}_{13}$ | 0 | -0.20 | 0.33 | (-0.79, 0.46) | -0.26 | 0.33 | (-0.81, 0.40) | -0.57 | 0.26 | (-0.93, 0.02) |
| $\bar{\Sigma}_{23}$ | -0.5 | -0.38 | 0.32 | (-0.91, 0.29) | -0.54 | 0.31 | (-0.95, 0.21) | -0.52 | 0.30 | (-0.93, 0.20) |
| $\alpha_1$ | 2 | 0.50 | 3.62 | (-5.69, 7.46) | -1.75 | 3.24 | (-6.90, 5.35) | 0.66 | 3.69 | (-8.11, 6.66) |
| $\alpha_2$ | 0 | -2.89 | 3.33 | (-9.81, 2.74) | 0.39 | 3.18 | (-4.78, 8.31) | 2.88 | 4.16 | (-5.15, 11.32) |
| $\alpha_3$ | -2 | 2.62 | 3.72 | (-5.19, 9.20) | -0.49 | 4.59 | (-7.43, 9.79) | -0.81 | 3.02 | (-6.46, 5.53) |

discarding the first 3000 iterations as burn-in. The prior for $\boldsymbol{\beta}$ is taken as $\mathcal{N}_p(0, 25\mathrm{I}_p)$, the prior for $\boldsymbol{\alpha}_s$ is $\mathcal{N}_M(0, 16)$, and the variances $h_1, h_2$ of the proposal normal densities in the Metropolis-Hastings algorithm described in Section 4.1 are taken as $h_1 = h_2 = 0.09$. Table 1 displays the posterior estimates for each of scenarios. The results show that the regression coefficients $\boldsymbol{\beta}$ can be quite well estimated in all scenarios, but the skewness parameter $\boldsymbol{\alpha}_s$ and dispersion matrix $\bar{\Sigma}$ are not easy to estimate. This is expected as both $\boldsymbol{\alpha}_s$ and $\bar{\Sigma}$ determine the dependence between the different binary observations in $\boldsymbol{y}$ and such dependence enforced at the latent level cannot be easily estimated with a relatively small sample size $n$. Moreover, the credible intervals appear to be more narrow for larger degrees of freedom $\nu$, which is due to the weaker dependence among observations (hence, larger effective sample size) implied by larger $\nu$.

## 4.4 Application to COVID-19 Pandemic Data

In this section we illustrate our methodology to COVID-19 pandemic data from different counties of the state of California, USA, freely downloaded from the California open data portal https://data.ca.gov. The dataset contains the number of daily new confirmed cases and deaths from March 18, 2020, to November 24, 2020, in 58 counties of California. There is a clear weekly cyclic pattern in this dataset, i.e., the numbers of new confirmed cases on weekdays are often much larger than those during the weekends. This is possibly due to the fact that people tend to enjoy their weekends and go to the hospital for testing after the weekend, or some testing facilities are closed during weekends. To avoid modeling this artificial cyclic pattern, we aggregate the data and consider the weekly new confirmed cases, resulting in $n = 36$ weekly observations. As $n$ is relatively small, we here only focus on the three most populous counties in California, i.e., Los Angeles, San Diego and Orange. Our goal here is to jointly model the occurrence of extreme spikes in new weekly cases, i.e., "abnormal" weeks with respect to the overall expected trend, and to detect if the spikes are spatially correlated. This informs us about the potential transmission modes of the virus between counties, and whether an outburst in one county may lead to an increased number of cases in another county.

To remove the obvious trend, we apply smoothing splines with five knots to the logarithm of each of the three time series, where the logarithm is used because most epidemics grow approximately exponentially during the initial phase (Ma, 2020). Alternatively, one can try to fit the well-known susceptible-infected-recovered (SIR) model (Anderson & May, 1979; May & Anderson, 1979) to remove the trend. Unfortunately, this is not feasible with our dataset as it does not contain the number of daily recovered cases. Figure 1 displays the observed data for the three counties, the smoothing splines for each time series and the resulting residuals. We then consider a residual point as an extreme spike if it exceeds the empirical 90% quantile of the corresponding time series, and we denote it as 1; otherwise we denote it as 0. In this way, we get three imbalanced binary time series and we aim to model the dependence among them.

We consider three covariates in total, i.e., an intercept, one covariate as time, and another one as the square of time. Following the recommendation of Gelman et al. (2008), we standardize the two temporal predictors in a preliminary step to make them have mean 0 and standard deviation 1. To assess the performance of the multivariate skew-normal link model, we consider six models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5, \mathcal{M}_6$ of different complexity. $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ are the multivariate skew-$t$ link model with $\nu = 5, 10, 20$, respectively, $\mathcal{M}_4$ is the multivariate skew-normal model (i.e., obtained as $\nu \to \infty$), $\mathcal{M}_5$ is the multivariate probit model (obtained with $\nu \to \infty$ and $\alpha = 0$), and $\mathcal{M}_6$ is the independent probit model (obtained with $\nu \to \infty$, $\bar{\Sigma} = \mathrm{I}, \alpha = 0$).

For each of these models, we run the Algorithm 1 for 25000 iterations and remove the first 5000 samples as burn-in. The prior for the regression parameters $\boldsymbol{\beta}$ is specified as $\mathcal{N}_p(\mathbf{0}, 25\mathrm{I}_p)$, and the prior for the skewness parameters is taken as $\mathcal{N}_M(\mathbf{0}, 16\mathrm{I}_M)$. The variances of the proposal densities in the Metropolis-Hastings algorithm are taken as $h_1 = h_2 = 0.09$.

Table 2 summarizes the estimation results for all the models. The results show that the estimate

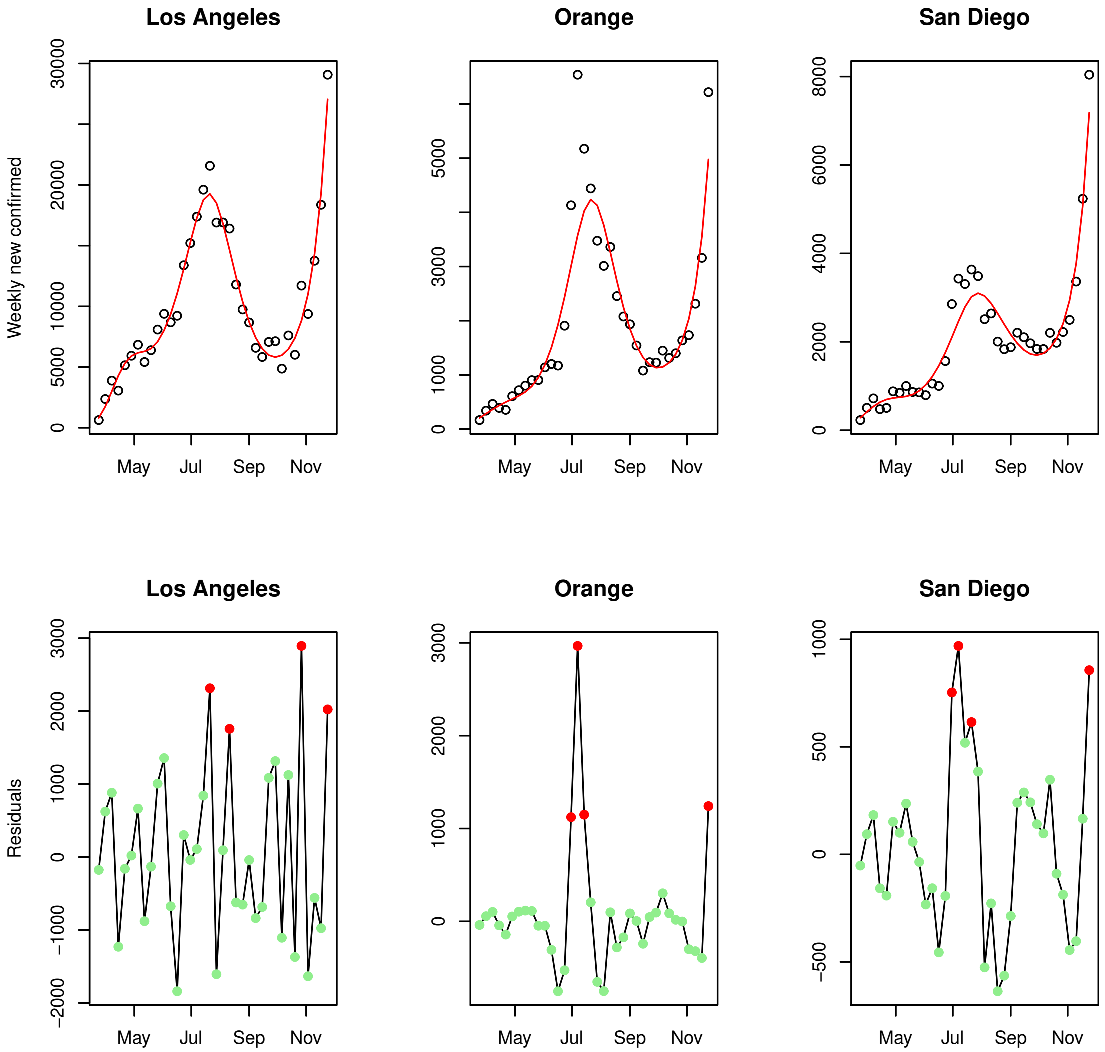

Figure 1: Upper panel: smoothing splines for the time series of weekly new confirmed cases at Los Angeles (left), Orange (middle), and San Diego (right). Lower panel: the residuals obtained as the difference between the original data and the fitted splines, with red points considered as extreme spikes and green points as non-extreme values.

of the intercept for all the models are almost the same and are significantly negative. This is expected as 90% of the observations are 0 and only 10% are 1. We also observe that the confidence intervals for the correlation and skewness parameters are generally quite large (as in the simulation study), implying that they are hard to estimate with only $n = 36$ observations. However, the correlation between the counties of Orange and San Diego, i.e., $\bar{\Sigma}_{23}$, seems to be quite strong, as its posterior mean for models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$, and $\mathcal{M}_5$ is consistently far from zero (with an estimate close to 0.78) and its 95% credible interval always excludes zero. This indicates that these two counties are more connected together in terms of extreme COVID-19 cases than the other pairs of counties considered, which sheds some light into the spread of the epidemic. The extreme occurrences observed in the counties of Los Angeles and San Diego also seem fairly strongly interconnected since the estimate of $\Sigma_{13}$ is also quite high, yet to a milder degree.

To compare the different fitted models, we use the Deviance Information Criterion (DIC) proposed by Spiegelhalter et al. (2002). The DIC is the Bayesian analogue of the Akaike Information Criterion (AIC) and is defined as

$$\mathrm{DIC} = D(\bar{\boldsymbol{\tau}}) + 2p_D,$$

where $\boldsymbol{\tau}$ denotes the collection of all the parameters, $\bar{\boldsymbol{\tau}} = \mathrm{E}[\boldsymbol{\tau} \mid \boldsymbol{y}]$ is its posterior mean, $D(\cdot)$ is a deviance function and $p_D = \mathrm{E}[D(\boldsymbol{\tau}) \mid \boldsymbol{y}] - D(\bar{\boldsymbol{\tau}})$ is the effective number of model parameters. Here we take the deviance function $D(\boldsymbol{\tau})$ as $-2\log p(\boldsymbol{y} \mid \boldsymbol{\beta}, \bar{\Sigma}, \boldsymbol{\alpha}, \nu)$ when the model is the skew-$t$ link model, or $-2\log p(\boldsymbol{y} \mid \boldsymbol{\beta}, \bar{\Sigma}, \boldsymbol{\alpha})$ when the model is the skew-normal link model, and estimate $\mathrm{E}[D(\boldsymbol{\tau}) \mid \boldsymbol{y}]$ by Monte Carlo using the samples generated from Algorithm 1. The smaller the DIC value, the better the model's goodness-of-fit and predictive performance. We refer to Spiegelhalter et al. (2002) for other properties about the DIC measure.

Table 3 reports the estimated DIC values for the six different models. The results show that the multivariate skew-normal model $\mathcal{M}_4$ provides the best fit to the data despite its high complexity, the multivariate probit model $\mathcal{M}_5$ is the second best, and the independent symmetric probit model $\mathcal{M}_6$ is the worst. This has two major implications. The first is that spatial dependence plays an important role in the spread of the epidemic and ignoring the correlation would lead to a poor fit of the extreme spikes. The second is that adding the skewness parameter indeed improves the model's flexibility and can provide a better fit to our highly imbalanced dataset.

Table 2: Posterior estimates for different models fitted in our COVID-19 data application in Section 4.4

| | $\mathcal{M}_1$ | | | $\mathcal{M}_2$ | | | $\mathcal{M}_3$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Est | Sd | 95% CI | Est | Sd | 95% CI | Est | Sd | 95% CI |
| Intercept | -1.67 | 0.77 | (-3.62, -0.76) | -1.58 | 0.49 | (-2.76, -0.86) | -1.48 | 0.37 | (-2.34, -0.89) |
| Time | 1.72 | 1.68 | (-0.67, 5.80) | 1.62 | 1.36 | (-0.54, 4.75) | 1.54 | 1.27 | (-0.60, 4.43) |
| Time$^2$ | -1.19 | 1.38 | (-4.47, 0.99) | -1.13 | 1.15 | (-3.73, 0.81) | -1.07 | 1.08 | (-3.45, 0.82) |
| $\bar{\Sigma}_{12}$ | 0.24 | 0.29 | (-0.36,0.74) | 0.23 | 0.46 | (-0.33, 0.72) | 0.24 | 0.28 | (-0.34, 0.73) |
| $\bar{\Sigma}_{13}$ | 0.54 | 0.24 | (-0.01, 0.89) | 0.51 | 0.23 | (0.02, 0.87) | 0.51 | 0.24 | (-0.03, 0.88) |
| $\bar{\Sigma}_{23}$ | 0.78 | 0.16 | (0.40, 0.97) | 0.76 | 0.15 | (0.40, 0.97) | 0.73 | 0.17 | (0.29 0.95) |
| $\alpha_1$ | 1.86 | 4.69 | (-6.81, 10.71) | -2.04 | 3.13 | (-7.31, 4.47) | -0.26 | 2.94 | (-5.64, 5.54) |
| $\alpha_2$ | -0.82 | 3.49 | (-6.74, 6.66) | -0.45 | 3.13 | (-6.00, 4.89) | 1.05 | 3.41 | (-5.64, 6.82) |
| $\alpha_3$ | -0.42 | 3.02 | (-7.13, 4.91) | 0.05 | 4.37 | (-7.02, 8.75) | -0.20 | 4.83 | (-10.34, 7.70) |
| | $\mathcal{M}_4$ | | | $\mathcal{M}_5$ | | | $\mathcal{M}_6$ | | |
| | Est | Sd | 95% CI | Est | Sd | 95% CI | Est | Sd | 95% CI |
| Intercept | -1.40 | 0.27 | (-1.98, -0.92) | -1.42 | 0.26 | (-1.97, -0.96) | -1.36 | 0.20 | (-1.77, -1.00) |
| Time | 1.47 | 1.18 | (-0.62, 4.05) | 1.47 | 1.18 | (-0.56, 4.05) | 1.23 | 0.89 | (-0.37, 3.13) |
| Time$^2$ | -1.01 | 1.02 | (-3.13, 0.86) | -1.02 | 1.02 | (-3.17, 0.79) | -0.84 | 0.77 | (-2.45, 0.59) |
| $\bar{\Sigma}_{12}$ | 0.27 | 0.28 | (-0.32, 0.76) | 0.25 | 0.30 | (-0.34, 0.77) | | | |
| $\bar{\Sigma}_{13}$ | 0.62 | 0.24 | (0.01, 0.93) | 0.53 | 0.24 | (0.01, 0.90) | | | |
| $\bar{\Sigma}_{23}$ | 0.78 | 0.16 | (0.37, 0.98) | 0.74 | 0.16 | (0.34, 0.95) | | | |
| $\alpha_1$ | 1.65 | 5.16 | (-5.29, 11.63) | | | | | | |
| $\alpha_2$ | -0.39 | 3.08 | (-6.83, 6.43) | | | | | | |
| $\alpha_3$ | 0.39 | 2.54 | (-4.54, 5.53) | | | | | | |

Table 3: Estimated DIC values for the different models fitted in our COVID-19 data application in Section 4.4

| Model | # of parameters | DIC |
|---|---|---|
| $\mathcal{M}_1$ | 9 | 67.93 |
| $\mathcal{M}_2$ | 9 | 68.06 |
| $\mathcal{M}_3$ | 9 | 67.84 |
| $\mathcal{M}_4$ | 9 | **65.77** |
| $\mathcal{M}_5$ | 6 | 67.12 |
| $\mathcal{M}_6$ | 3 | 78.97 |

# 5 Conclusion

Although we here focus on the skew-elliptical link model, the result of a closed-form posterior for the regression coefficients could also be obtained if we consider a more flexible class of distributions for the assumption (6). In fact, if $\boldsymbol{\varepsilon} \mid \boldsymbol{\beta}, \Sigma, \boldsymbol{\alpha}$ has a distribution which is closed under affine transformation, following the proof of Lemma 1 and Theorem 1, one can show that the posterior of $\boldsymbol{\beta}$ coincides with a fundamental skew distribution (Arellano-Valle & Genton, 2005). This novel result opens up new avenues for the development of skewed link models for correlated binary data.

There are various directions for future research. As the number of observations in our dataset is relatively small, we chose not to consider too many covariates and restricted the number of counties. An interesting extension of our real data application would be to consider a larger dataset with more informative covariates, such as daily weather information or population migration between different counties. Adding such extra covariates could potentially fit the data better and provide a more detailed and informed explanation of the spread of epidemic. Another interesting methodological extension is to improve Algorithm 1. As we used the accept-reject algorithm of Botev (2017) within Algorithm 1 to sample from a multivariate truncated $t$ distribution, its lack of scalability to higher dimensions is inevitably inherited. Therefore, more efficient algorithms to sample from high-dimensional truncated normal and $t$ distributions would significantly improve the speed of algorithm 1. Finally, in the simulation study and data application we chose to fix the degrees of freedom $\nu$ to three different values to facilitate inference. If one has many more observations and a more efficient algorithm to sample from high-dimensional truncated $t$ distribution, one can alternatively include the estimation of $\nu$ in the

Metropolis-Hastings algorithm described in Section 4.1.

# References

Anderson, R. M., & May, R. M. (1979). Population biology of infectious diseases: Part i. *Nature*, *280*(5721), 361–367.

Arellano-Valle, R. B., & Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, *33*, 561–574.

Arellano-Valle, R. B., & Genton, M. G. (2005). On fundamental skew distributions. *Journal of Multivariate Analysis*, *96*(1), 93–116.

Arellano-Valle, R. B., & Genton, M. G. (2010). Multivariate unified skew-elliptical distributions. *Chilean Journal of Statistics*, *1*(1), 17–33.

Ashford, J. R., & Sowden, R. R. (1970). Multivariate probit analysis. *Biometrics*, *26*(3), 535–546.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178.

Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistics Society (Series B)*, *61*(3), 579–602.

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistics Society (Series B)*, *65*(2), 367–389.

Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, *83*(4), 715–726.

Barnard, J., McCullloch, R., & Meng, X. (2000). Modeling covariance matries in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281–1311.

Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society (Series B)*, *79*, 125–148.

Branco, M. D., & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, *79*, 99–113.

Cao, J., Durante, D., & Genton, M. G. (2020a). Scalable computation of predictive probabilities in probit models with gaussian process priors. Available from https://arxiv.org/abs/2009.01471.

Cao, J., Genton, M. G., Keyes, D. E., & Turkiyyah, G. M. (2020b). Exploiting low rank covariance structures for computing high-dimensional normal and student-t probabilities. *Statistics and Computing*, to appear.

Chen, M.-H., Dey, D. K., & Shao, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, *94*(448), 1172–1186.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361.

Chin, V., Gunawan, D., Fiebig, D. G., Kohn, R., & Sisson, S. A. (2020). Efficient data augmentation for multivariate probit model with panel data: an application to general practitioner decision making about contraceptives. *Journal of the Royal Statistical Society (Series C)*, *69*(2), 277–300.

Czado, C., & Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, *33*, 213–231.

Durante, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, *106*(4), 765–779.

Fang, B. Q. (2003). The skew elliptical distributions and their quadratic forms. *Journal of Multivariate Analysis*, *87*, 298–314.

Fasano, A., & Durante, D. (2020). A class of conjugate priors for multinomial probit models which includes the multivariate normal one. Available from https://arxiv.org/abs/2007.06944.

Fitzmaurice, G. M., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal Data Analysis*. Chapman and Hall/CRC Press.

Fitzmaurice, G. M., Molenberghs, G., & Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society (Series B)*, *57*(4), 691–704.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis*. CRC Press, third ed.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Genton, M. G., Keyes, D. E., & Turkiyyah, G. (2018). Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *27*(2), 268–277.

Genz, A., & Bretz, F. (1999). Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, *63*, 361–378.

Genz, A., & Bretz, F. (2002). Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, *11*(4), 950–971.

Gupta, A. K. (2003). Multivariate skew t-distribution. *Statistics: A Journal of Theoretical and Applied Statistics*, *37*(4), 359–363.

Heagerty, P. J., & Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, *93*(443), 1099–1111.

Johndrow, J. E., Smith, A., Pillai, N., & Dunson, D. B. (2019). MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, *114*(527), 1394–1403.

Kim, S., Chen, M.-H., & Dey, D. K. (2008). Flexible generalized t-link models for binary response data. *Biometrika*, *95*(1), 93–106.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*, 1989–2001.

Lin, P.-S., & Clayton, M. K. (2005). Analysis of binary spatial data by quasi-likelihood estimating equations. *The Annals of Statistics*, *33*(2), 542–555.

Ma, J. (2020). Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, *5*, 129–141.

May, R. M., & Anderson, R. M. (1979). Population biology of infectious diseases: Part ii. *Nature*, *280*(5722), 455–461.

R Core Team (2020). R: A language and environment for statistical computing.
URL `https://www.R-project.org/`

Smith, M. S. (2013). Bayesian approaches to copula modelling. In P. Damien, P. Dellaportas, N. G. Polson, & D. A. Stephens (Eds.) *Bayesian Theory and Applications*, chap. 17, (pp. 336–358). Oxford University Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society (Series B)*, *64*, 583–639.