PLOS ONE

RESEARCH ARTICLE

Inference in skew generalized t-link models for clustered binary outcome via a parameter-expanded EM algorithm

Chénangnon Frédéric Tovissodé[1]*, Aliou Diop[2], Romain Glèlè Kakaï[1]

1 Laboratoire de Biomathématiques et d'Estimations Forestières, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, Abomey-Calavi, Bénin, **2** Laboratoire d'Etudes et Recherches en Statistiques et Développement, Université Gaston Berger de Saint-Louis, Saint-Louis, Sénégal

* chenangnon@gmail.com

Citation: Tovissodé CF, Diop A, Glèlè Kakaï R (2021) Inference in skew generalized t-link models for clustered binary outcome via a parameter-expanded EM algorithm. PLoS ONE 16(4): e0249604. https://doi.org/10.1371/journal.pone.0249604

Editor: Luca Citi, University of Essex, UNITED KINGDOM

Received: August 15, 2020

Accepted: March 19, 2021

Published: April 6, 2021

Data Availability Statement: All relevant data are within the paper and its Supporting information files.

Funding: CFT is grateful to the Centre d'Excellence Africain en Sciences Mathématiques et Applications (CEA-SMA, https://ceasma-benin.org/) for funding his work. CFT was also financially supported by the African German Network of Excellence in Science (AGNES), through the "AGNES mobility grant for young scientists from sub Saharan Africa" (https://agnes-h.org/). The

Abstract

Binary Generalized Linear Mixed Model (GLMM) is the most common method used by researchers to analyze clustered binary data in biological and social sciences. The traditional approach to GLMMs causes substantial bias in estimates due to steady shape of logistic and normal distribution assumptions thereby resulting into wrong and misleading decisions. This study brings forward an approach governed by skew generalized t distributions that belong to a class of potentially skewed and heavy tailed distributions. Interestingly, both the traditional logistic and probit mixed models, as well as other available methods can be utilized within the skew generalized t-link model (SGTLM) frame. We have taken advantage of the Expectation-Maximization algorithm accelerated via parameter-expansion for model fitting. We evaluated the performance of this approach to GLMMs through a simulation experiment by varying sample size and data distribution. Our findings indicated that the proposed methodology outperforms competing approaches in estimating population parameters and predicting random effects, when the traditional link and normality assumptions are violated. In addition, empirical standard errors and information criteria proved useful for detecting spurious skewness and avoiding complex models for probit data. An application with respiratory infection data points out to the superiority of the SGTLM which turns to be the most adequate model. In future, studies should focus on integrating the demonstrated flexibility in other generalized linear mixed models to enhance robust modeling.

Introduction

Binary outcomes are prominent in many applied sciences, including but not limited to biological and social sciences. Moreover, in cross sectional as well as panel studies, dichotomous responses are often naturally grouped by sampling techniques or some properties of the sampling units [1]. The preferred modern method to analyze clustered binary data is through the Generalized Linear Mixed Model (GLMM) framework [2]. Indeed, when a binary outcome

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

has been recorded repeatedly or in the presence of latent factors, GLMMs allow accounting explicitly for over-dispersion and correlation within clusters using random effects.

Let Y_{ij} denote the binary outcome (0 or 1) of the j^{th} measurement ($j = 1, 2, \cdots, n_i$) and $\mathbf{Y}_i$ the collection of responses from the i^{th} cluster, *i.e.* $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$, $i = 1, 2, \cdots, n$. In terms of an underlying latent continuous random vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{in_i})^\top$ and random effects $\mathbf{b}_i = (b_{i1}, \cdots, b_{iq})^\top$, the mixed probit model (PM) assumes that Y_{ij} are conditionally independent and given as [3]:

$$\begin{aligned} Y_{ij} = I_{(\mathbf{0},\infty)}(Z_{ij}), \mathbf{Z}_i|\mathbf{b}_i &\overset{ind}{\sim} \mathcal{N}_{n_i}(\boldsymbol{\eta}_i, \mathbf{I}_{n_i}) \\ \mathbf{b}_i &\overset{ind}{\sim} \mathcal{N}_q(\mathbf{0}, \mathbf{D}) \end{aligned} \tag{1}$$

where $I_A(x)$ is the indicator function which equals to 1 if $x \in A$ and 0 otherwise; $\boldsymbol{\eta}_i$ is the n_i*−vector* of linear predictors, $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{in_i})^\top = \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i$; $\boldsymbol{\beta}$ is the p*−vector* of fixed effects; $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i})^\top$ and $\mathbf{W}_i = (\mathbf{W}_{i1}, \ldots, \mathbf{W}_{in_i})^\top$ are respectively known $n_i \times p$ and $n_i \times q$ matrices of covariates with $\mathbf{X}_{ij} = (X_{ij1}, \cdots, X_{ijp})^\top$ and $\boldsymbol{W}_{ij} = (W_{ij1}, \cdots, W_{ijq})^\top$; $\mathbf{I}_{n_i}$ is the $n_i \times n_i$ identity matrix and $\mathcal{N}_q(\mathbf{0}, \mathbf{D})$ denotes the q-variate normal distribution, with null mean vector and variance-covariance an unknown $q \times q$ matrix $\mathbf{D}$, meant to capture the dependence structure of $\mathbf{Y}_i$. The latent variable Z_{ij} serves for a convenient stochastic representation of the conditional outcome Y_{ij}. Equivalently, one may write $P(Y_{ij} = 1|\mathbf{b}_i) = \Phi(\eta_{ij})$ with $\Phi(\cdot)$ the cumulative distribution function (cdf) of the standard normal distribution, standing as the inverse link function mapping the linear predictor η_{ij} and the predicted probabilities of the outcome Y_{ij}. Combined with the normality assumption on random effects, the systematic use of this link and the well known alternative, the logit link, is somewhat controversial [4, 5].

The link function indeed has a critical role in GLMMs since it heavily impacts estimates, predictions and consequently interpretations [4, 6]. As a result, in the binary generalized linear model literature, aside the logistic and probit models based on the steady shape logistic and normal distributions respectively, there has been increasing efforts to render the link function flexible. Many works have considered heavy tailed link functions, for instance the Semi-Nonparametric [7], Student-t [8] and generalized t [9] distributions, and elliptical scale mixtures [10, 11]. Indeed, the maximum likelihood estimators of logistic and probit regression models are not robust to outliers [7]. Heavy tailed links are not sensitive to outliers and thus allow outlier-robust inference. In particular, the links functions based on the Student t distribution incorporate observation-specific stochastic weights which can be used for outlier detection [7, 12]. Similarly, skew-probit [13], skew generalized t [9], asymmetric logistic [14], loglog and complementary loglog, power symmetric and reciprocal power symmetric [15] links were used among others to handle situations where the probability of a given binary response approaches zero at a different rate than it approaches one. Skew logistic distributions have also been developed (see *e.g.* [16]) and may be used with the same aim in mind. For example [9], discussed a prostate cancer study where the outcome variable Y represents the presence or the penetration of cancer in or near the prostate capsule of patients. The rate at which the probability of "$Y = 1$" approaches one is expected to be very different (slower) from the rate at which it approaches zero [9]. For this study, the skew generalized t-link fits best the data [9], indicating that the simultaneous use of skewed and heavy tailed link functions can lead to more effective modelling of binary data.

Furthermore, although random effects are traditionally assumed to be normally distributed in GLMMs, this may not be realistic [17, 18]. Therefore, huge efforts have been devoted to making the random effects distribution in GLMMs flexible, replacing the normal distribution

with, for instance Semi-Nonparametric [19], probability integral transformation of normal [20], skew normal [21], log-normal [22, 23], Student-t [24] and scale mixtures of normal [25] distributions.

The above background demonstrates the extent to which the number of possible approaches for fitting a flexible GLMM to correlated binary outcomes goes, although none of these approaches attempts to explicitly account for skewness and tail behavior of the link function as well as the random effects distribution simultaneously. However, the misspecification of the link function or the random effects distribution can introduce substantial bias and reduce the accuracy of the mean response as well as heterogeneity estimates [6, 18]. Standing in a fully parametric frame, we propose a unifying approach based on skew generalized t (SGT) distributions [26], that is the class of models including among others the normal, the skew normal and the Student t models. The use of a skew generalized t family instead of the Student t family allows to rescale fixed effects so that they have the same interpretation as in the mixed probit model in Eq (1).

Our contributions include *i*) an extension of the flexible generalized t-link model built for independent binary samples proposed by [9] to deal with dependent binary samples (mixed model); *ii*) a parameter-expanded EM algorithm [27] for computing the maximum likelihood of skew generalized t-link models for correlated binary data, extending the EM algorithm of [24] for t-link models; and *iii*) empirical Bayes estimators of skew t distributed random effects in mixed models for binary data.

The organization of the paper is as follows. Section 2 presents preliminary results on the SGT distributions and the truncated SGT distributions and their first two moments. Section 3 specifies the SGT-link model (SGTLM) and describes maximum likelihood estimation and cluster-specific inference based on random effects and weights. A simulation study assessing the relative performance of the SGTLM relative to existing methods and the application of the modelling approach to a real respiratory infection data are presented in Section 4. Concluding remarks are given in section 5.

Preliminary results

In this section, we present some useful properties of the skew generalized t distributions.

Multivariate skew generalized t distributions

Multivariate skew generalized t (SGT) distributions are special cases of multivariate skew scale mixture of normal (SSMN) distributions [28] (pages 102-103) which we first introduce. A random variable $\mathbf{Z}$ is said to follow a *p−variate* SSMN distribution with location $\boldsymbol{\mu}$, scale $\boldsymbol{\Omega}$, and shape $\boldsymbol{\lambda}$, if it can be represented as [29] (page 20, Eq 3.12):

$$\mathbf{Z} = \boldsymbol{\mu} + U^{-1/2}(\boldsymbol{\delta} Z_0 + \mathbf{X}), \quad Z_0 \sim \mathcal{HN}(\mathbf{0}, 1), \quad \mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \overline{\boldsymbol{\Omega}}) \tag{2}$$

where U, called scale mixing variable, is a positive random variable with cdf $\boldsymbol{F}_U(\cdot|\boldsymbol{\nu})$ indexed by a parameter vector $\boldsymbol{\nu}$, $\mathcal{HN}(0, 1)$ is the standard half normal distribution; Z_0, $\mathbf{X}$ and U are independent; $\overline{\boldsymbol{\Omega}} = \boldsymbol{\Omega} - \boldsymbol{\delta}\boldsymbol{\delta}^\top$ and $\boldsymbol{\delta} = (1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})^{-1/2} \boldsymbol{\Omega}^{1/2} \boldsymbol{\lambda}$. Different choices of the scale mixing distribution $F_U(\cdot|\boldsymbol{\nu})$ result in various sub-classes of skew elliptical distributions, for instance, the skew normal when $P(U = 1) = 1$ [28] (page 103); the skew contaminated normal when $\boldsymbol{\nu} = (\nu_1, \nu_2)^\top$, $0 < \nu_1 < 1$, $0 < \nu_1 \leq 1$ and U is discrete and takes the values $U = 1$ with probability $1 - \nu_1$ and $U = \nu_2$ with probability ν_1 [30] (page 308); the skew slash when $U \sim \mathcal{B}eta(\nu, 1)$, $\nu > 0$ [30] (page 307); and the skew generalized t when $\boldsymbol{\nu} = (\nu, \nu_0)^\top$, $\nu > 0$, $\nu_0 > 0$, $U \sim \mathcal{G}amma(\nu/2, \nu_0/2)$

[28] (page 105). The following result states conditions for the identifiability of SSMN distributions, a requirement for reliable inference using this class of distributions.

Lemma 1 (see S1 Appendix for a proof) *The free parameters* ($\boldsymbol{\mu}$, $\boldsymbol{\delta}$, Ω *and* $\boldsymbol{\nu}$) *of a SSMN distribution with the representation in Eq (2) are identifiable if and only if i) the scale mixing distribution* $F_U(\cdot|\nu)$ *is identifiable and ii)* $F_U(\cdot|\nu)$ *does not satisfy* $F_U(u|\nu) = H\left(\frac{u}{\nu_k}|\nu_{-k}\right)$ *for any element* ν_k *of* $\boldsymbol{\nu}$ *and any distribution function* $H(\cdot|\nu_{-k})$ *where* $\boldsymbol{\nu}_{-k}$ *is the vector* $\boldsymbol{\nu}$ *without the element* ν_k. *If U has a probability density function (pdf)* $f_U(u|\nu)$ *for all* $u > 0$, *then the condition ii) is equivalent to* $f_U(\cdot|\nu)$ *does not satisfy* $f_U(u|\nu) = \frac{1}{\nu_k} h_U\left(\frac{u}{\nu_k}|\nu_{-k}\right)$ *for any pdf* $h_U(\cdot|\nu_{-k})$.

On setting $c = \sqrt{2/\pi}$ and defining the expectations $\tilde{U}_t = \mathrm{E}_U\{U^{-t/2}\}$, and assuming that $\tilde{U}_t < \infty$ for the required expectations, the first two central moments of a SSMN vector $\mathbf{Z}$ are given by [28] (pages 109-110):

$$\mathrm{E}\{\mathbf{Z}\} = \boldsymbol{\mu} + c\tilde{U}_1\boldsymbol{\delta} \quad \text{and} \tag{3}$$

$$\mathrm{Var}\{\mathbf{Z}\} = \tilde{U}_2\boldsymbol{\Omega} - c^2\tilde{U}_1^2\boldsymbol{\delta}\boldsymbol{\delta}^\top. \tag{4}$$

The ability of the SSMN distributions to capture more data structure than the normal, the skew normal or the scale mixture of normals is reflected in the expressions for skewness (γ_{1_k}) and kurtosis (γ_{2_k}) indices given for the k^{th} marginal of $\mathbf{Z}$ as [31]:

$$\gamma_{1_k} = \frac{c\boldsymbol{\delta}_k}{\sigma_k^3}\left[3(\tilde{U}_3 - \tilde{U}_1\tilde{U}_2) - \boldsymbol{\delta}_k^2\left(\tilde{U}_3 - 4\frac{\tilde{U}_1^3}{\pi}\right)\right], \text{and} \tag{5}$$

$$\gamma_{2_k} = \frac{1}{\sigma_k^4}\left[3(\tilde{U}_4 - \tilde{U}_2^2) - 4c^2\boldsymbol{\delta}_k^2\tilde{U}_1\left[3(\tilde{U}_3 - \tilde{U}_1\tilde{U}_2) - \boldsymbol{\delta}_k^2\left(\tilde{U}_3 - 3\frac{\tilde{U}_1^3}{\pi}\right)\right]\right] \tag{6}$$

where δ_k is the k^{th} element of $\boldsymbol{\delta}$ and σ_k^2 is the k^{th} diagonal element of the covariance matrix given in Eq (4).

We notice from the expressions for skewness Eq (5) and kutosis Eq (6) indices that the parameter $\boldsymbol{\lambda}$ controls the shape of the distribution only through the working shape parameter $\boldsymbol{\delta} = (1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{-1/2}\,\Omega^{1/2}\,\boldsymbol{\lambda}$. This quantity is invariant under marginalization, *i.e.* by the stochastic representation in Eq (2), for any arbitrary subset of $\mathbf{Z}$, the working shape parameter is the corresponding subset of $\boldsymbol{\delta}$. It is worth noticing however that the quantity $\boldsymbol{\delta}$ cannot be specified unrestrictedly, independently of Ω. Indeed, we observe that $\boldsymbol{\delta} = (1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{-1/2}\,\Omega^{1/2}\,\boldsymbol{\lambda}$ implies that $\boldsymbol{\lambda} = (1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{1/2}\,\Omega^{-1/2}\,\boldsymbol{\delta}$. This in turn gives $\boldsymbol{\lambda}^\top\boldsymbol{\lambda} = (1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})\boldsymbol{\delta}^\top\,\Omega^{-1}\,\boldsymbol{\delta}$ which yields $\boldsymbol{\lambda}^\top\boldsymbol{\lambda}$ $(1-\boldsymbol{\delta}^\top\,\Omega^{-1}\,\boldsymbol{\delta}) = \boldsymbol{\delta}^\top\,\Omega^{-1}\,\boldsymbol{\delta}$. We then get $\boldsymbol{\lambda}^\top\boldsymbol{\lambda} = \frac{\boldsymbol{\delta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}}{1-\boldsymbol{\delta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}}$ so that $1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda} = \frac{1}{(1-\boldsymbol{\delta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\delta})}$, *i.e.* $1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda}$ $= (1 - \delta^\top\,\Omega^{-1}\,\boldsymbol{\delta})^{-1}$ provided $\boldsymbol{\delta}^\top\,\Omega^{-1}\,\boldsymbol{\delta} \neq 1$. Therefore, $\boldsymbol{\lambda}$ is recorvered from $\boldsymbol{\delta}$ and Ω under the constraint $\boldsymbol{\delta}^\top\,\Omega^{-1}\,\boldsymbol{\delta} < 1$ as:

$$\boldsymbol{\lambda} = (1 - \boldsymbol{\delta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\delta})^{-1/2}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\delta}. \tag{7}$$

It is nevertheless possible to unrestrictedly specify $\boldsymbol{\delta}$ and $\overline{\boldsymbol{\Omega}}$ (positive definite). In this case, Ω is recovered as $\boldsymbol{\Omega} = \overline{\boldsymbol{\Omega}} + \boldsymbol{\delta}\boldsymbol{\delta}^\top$. Using the Sherman-Morrison identity [32] (page 121, Eq 3.1), we have $\boldsymbol{\Omega}^{-1} = (\overline{\boldsymbol{\Omega}} + \boldsymbol{\delta}\boldsymbol{\delta}^\top)^{-1} = \overline{\boldsymbol{\Omega}}^{-1} - \frac{\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}}{1+\boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}}$ from which we get $\boldsymbol{\delta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\delta} = \boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta} -$ $\frac{\boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}}{1+\boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}}$ that simplifies as $\boldsymbol{\delta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\delta} = \frac{\boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}}{1+\boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}}$. We thus have $1 - \boldsymbol{\delta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\delta} = \frac{1}{1+\boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}}$ hence

$$\boldsymbol{\lambda} = (1 + \boldsymbol{\delta}^\top\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta})^{1/2}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\delta} \quad \text{with} \quad \boldsymbol{\Omega} = \overline{\boldsymbol{\Omega}} + \boldsymbol{\delta}\boldsymbol{\delta}^\top. \tag{8}$$

In the binary data modeling framework, we shall consider $\boldsymbol{\delta}$ and $\overline{\boldsymbol{\Omega}}$ as model parameters as they turn to be easier to estimate by the mean of the EM algorithm. For the multivariate Skew Generalized t (SGT) distribution, the mixing variable U is gamma distributed, *i.e.* $U \sim \mathcal{G}amma(v/2, v_0/2)$ with pdf [33] (page 1, Eq 1):

$$f_G(u|v/2, v_0/2) = \frac{\left(\frac{v_0}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)} u^{\frac{v}{2}-1} \exp\left(-\frac{v_0}{2}u\right) \text{ for } u > 0, v > 0, v_0 > 0. \tag{9}$$

The *p−variate* SGT distribution, denoted $\mathcal{SGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, v)$ with $\boldsymbol{\nu} = (v, v_0)^\top$ has pdf for $\mathbf{z} \in \mathbb{R}^p$ [28] (page 106):

$$SGt_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, v) \quad = \quad 2\ Gt_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\nu})\ T\left(\alpha\left(\frac{p+v}{v_0 + \mathbf{z}_0^\top \mathbf{z}_0}\right)^{1/2} |p+v\right), \tag{10}$$

where

$$Gt_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\nu}) \quad = \quad \frac{\Gamma\left(\frac{p+v}{2}\right)|\boldsymbol{\Omega}|^{-1/2} v_0^{v/2}}{\Gamma(v/2)(\pi)^{p/2}} (v_0 + \mathbf{z}_0^\top \mathbf{z}_0)^{-(p+v)/2} \tag{11}$$

is the pdf of the *p−variate* Generalized t (GT) distribution, $\mathbf{z}_0 = \boldsymbol{\Omega}^{-1/2}(\mathbf{z} - \boldsymbol{\mu})$, $\alpha = \boldsymbol{\lambda}^\top \mathbf{z}_0$, and $T(\cdot|v)$ is the cdf of the standard univariate t distribution with v degrees of freedom. For SGT distributions, the expectations $\tilde{U}_t$ required for computing moments given in Eqs (3)–(6) have for $t < v$ the form

$$\tilde{U}_t \quad = \quad \left(\frac{v_0}{2}\right)^{t/2} \frac{\Gamma\left(\frac{v-t}{2}\right)}{\Gamma(v/2)}. \tag{12}$$

It is worthwhile noticing that the gamma mixing pdf $f_G(\cdot|v/2, v_0/2)$ satisfies the condition *i*) of *Lemma 1* but not the condition *ii*). The SGT ditribution with v as a parameter is thus not identifiable. However, restricting v_0 to a fixed value (so that only v is considered as a parameter) is sufficient to ensure identifiability of the SGT family of distribution. When $v_0 = v$, the *p−variate* SGT distribution $\mathcal{SGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, v)$ reduces to the *p−variate* Skew t (ST) distribution [28] (page 106), denoted $\mathcal{ST}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, v)$ which is thus identifiable with pdf $St_p(\cdot|\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda}, v)$ and cdf $St_p(\cdot|\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda}, v)$. If $\boldsymbol{\lambda} = 0$, the SGT distribution reduces to the GT distribution which equals the Student t distribution for $v_0 = v$. The following lemma formalizes the relationship between skew generalized t and skew t distributions.

Lemma 2 (see S2 Appendix for a proof) *Let us consider the SGT distribution* $\mathcal{SGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ *with* $\boldsymbol{\nu} = (v, v_0)^\top$ *and pdf in* Eq (10). *Set* $\boldsymbol{\Omega}^* = \frac{v_0}{v}\boldsymbol{\Omega}$. *The following statements hold*:

1. $SGt_p(\mathbf{z}|\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda}, v) = St_p(\mathbf{z}|\boldsymbol{\mu}, \Omega^*, \boldsymbol{\lambda}, v)$;
2. If $\mathbf{Z} \sim \mathcal{SGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, v)$ then $\mathbf{Z} \sim \mathcal{ST}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}^*, \boldsymbol{\lambda}, v)$.

Lemma 2 indicates that any SGT vector is a rescaled version of a ST vector. However, in the frame of binary data models, the use of a SGT distribution instead of a simple ST distribution as link function allows to control the scale of the link function through the parameter v_0 [9]. Specifically, a skew generalized t-link allows to define a skewed and heavy-tailled binary mixed model where fixed effects have the same scale and interpretation as in the mixed probit model in Eq (1). Interestingly, the popular logit and probit links for binary data can be recast as

special cases of the cdf of the SGT class of distributions. Indeed, the normal distribution is a limiting case of SGT distributions when $\nu_0 = \nu \to \infty$ and $\boldsymbol{\lambda} = 0$. Moreover, the logistic distribution is well appoximated by a rescaled Student t distribution with appropriate degrees of freedom [8] (page 228). These constatations make the SGT distributions appropriate for extending the traditional probit and logistic GLMMs, accounting for skewness and heavy tail behaviors. To this end, we present in the next section some results on truncated multivariate SGT distributions since binary data can reflect truncation of latent continuous variables.

Truncated multivariate skew generalized t distributions

As seen for the mixed probit model in Eq (1), models for binary data can be defined by truncating latent variables following continuous distributions. We define in this section a class of truncated multivariate skew generalized t distributions which are useful for a latent variable representation of skew generalized t-link binary data models. We also give expressions to evaluate some joint moments of a truncated multivariate skew generalized t distribution and a gamma distribution, as they prove useful in the implementation of the EM algorithm for the skew generalized t-link model.

Let $\mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu, \mathbb{A})$ represent a *p−variate* skew generalized t (SGT) vector restricted to a *p*-dimensional hyperplane $\mathbb{A}$; with $\boldsymbol{\mu}$ a *p−vector* (location), Ω a $p \times p$ positive definite matrix (scale), $\boldsymbol{\lambda}$ a *p−vector* (shape) and $\boldsymbol{\nu} = (\nu, \nu_0)^\top$ a vector of positive scalars (degrees of freedom). The pdf of $\mathbf{Z} \sim \mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbb{A})$ is:

$$TSGt_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbb{A}) = \alpha_{st}^{-1}\ SGt_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu})\ I_{\mathbb{A}}(\mathbf{z})\ \text{ for } \mathbf{z} \in \mathbb{R}^p \tag{13}$$

where $SGt_p(\cdot|\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is the pdf in Eq (10) and $\alpha_{st} = \int_{\mathbb{A}} SGt_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu})d\mathbf{z}$ serves for normalization. The cdf of $\mathbf{Z}$ is denoted $TSGT_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbb{A})$. When $\boldsymbol{\lambda} = 0$, we obtain a truncated generalized t distribution denoted $\mathcal{TGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\nu}, \mathbb{A})$ with pdf $TGt_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\nu}, \mathbb{A})$ and cdf $TGT_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\nu}, \mathbb{A})$. When $\nu_0 = \nu$, we get a truncated ST distribution denoted $\mathcal{TST}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu, \mathbb{A})$ with pdf $TSt_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu, \mathbb{A})$ and cdf $TST_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu, \mathbb{A})$. If both $\boldsymbol{\lambda} = 0$ and $\nu_0 = \nu$, the truncated multivariate SGT distribution is reduced to a truncated multivariate t distribution [34] denoted $\mathcal{TT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu, \mathbb{A})$ with pdf $Tt_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\nu}, \mathbb{A})$ and cdf $TT_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\nu}, \mathbb{A})$.

In the frame of correlated binary data models, the truncation region $\mathbb{A}$ typically has the form $\mathbb{A} = \mathbb{A}_1 \times \mathbb{A}_2 \times \ldots \times \mathbb{A}_p$ where $\mathbb{A}_k$ are real intervals of the form $\mathbb{A}_k = (-\infty, a_k]$ or $\mathbb{A}_k = (a_k, \infty)$, for $a_k \in \mathbb{R}$ ($k = 1, 2, \cdots, p$). Let us consider for instance a vector $\mathbf{Y}$ of three binary outcomes obtained by truncating the elements of a 3*−variate* SGT vector $\mathbf{Z} \sim \mathcal{SGT}_3(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu)$: $Y_k = 0$ if $Z_k \leq 0$ and $Y_k = 1$ if $Z_k > 0$. In practice, however, only the binary outcomes ($\mathbf{Y}$) are observable whereas the latent outcome $\mathbf{Z}$ is unobservable. Suppose one observes the binary outcomes $\mathbf{y} = (1, 0, 1)^\top$. This implies that the corresponding value $\mathbf{z}$ of the latent vector $\mathbf{Z}$ satisfies $z_1 > 0$, $z_2 \leq 0$, and $z_3 > 0$. The conditional distribution of $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$ (required for maximum likelihood estimation using the EM algorithm) is thus $\mathcal{SGT}_3(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ truncated to the region $\mathbb{A}_{e.g.} = (0, \infty) \times (-\infty, 0] \times (0, \infty)$, *i.e.* $\mathbf{Z}|\mathbf{Y} = \mathbf{y} \sim \mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbb{A}_{e.g.})$ as defined in Eq (13).

We shall use the simplified notation $\mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{a})$ with $\mathbf{a} \in \mathbb{R}^p$ to denote a truncated SGT distribution $\mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu, \mathbb{A})$ when $\mathbb{A}$ is the right truncated hyperplane $\mathbb{A} = \{\mathbf{z} \in \mathbb{R}^p |\ z_1 \leq a_1, \ldots, z_p \leq a_p\}$. In this case, $\alpha_{st} = SGT_p(\mathbf{a}|\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda}, \boldsymbol{\nu})$ with $SGT_p(\cdot|\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda}, \boldsymbol{\nu})$ the cdf of the *p−variate* ST distribution. This corresponds for instance to the situation where all binary outcomes are zeros. When $\boldsymbol{\lambda} = 0$, the right truncated SGT distribution $\mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, 0, \boldsymbol{\nu}, \mathbf{a})$ is a right truncated GT distribution denoted $\mathcal{TGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\nu}, \mathbf{a})$ whose pdf

and cdf are respectively denoted $TGt_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu, \mathbf{a})$ and $TGT_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu, \mathbf{a})$. When $\nu_0 = \nu$, the right truncated SGT distribution is a right truncated ST distribution denoted $\mathcal{TST}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{a})$ with pdf $TSt_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu, \mathbf{a})$ and cdf $TST_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu, \mathbf{a})$. If both $\boldsymbol{\lambda} = 0$ and $\nu_0 = \nu$, the distribution is reduced to a right truncated t distribution denoted $\mathcal{TT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu, a)$ with pdf $Tt_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu, \mathbf{a})$ and cdf $TT_p(\cdot|\boldsymbol{\mu}, \Omega, \nu, a)$. In the above example, if $\mathbf{y} = (0, 0, 0)^\top$, then the truncation region becomes $\mathbb{A}_{e.g.} = (-\infty, 0] \times (-\infty, 0] \times (-\infty, 0]$. Since all truncation points are zeros, we shall write in this case $\mathbf{Z}|\mathbf{Y} = \mathbf{y} \sim \mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{a})$ with $\mathbf{a} = (0, 0, 0)^\top$ using the above simplified notation.

The implementation of an EM algorithm for a SGT distribution based binary data model requires joint moments of the form $\overline{u_r} = \mathrm{E}\{U^{r/2}\}$, $\overline{u_r \mathbf{z}_s} = \mathrm{E}\{U^{r/2}\mathbf{Z}^{(s)}\}$, $\overline{\tau_r} = \mathrm{E}\{U^{r/2}\zeta_1(U^{1/2}\alpha)\}$ and $\overline{\tau_r \mathbf{z}_s} = \mathrm{E}\{U^{r/2}\zeta_1(U^{1/2}\alpha)\mathbf{Z}^{(s)}\}$ for $s \in \{1, 2\}$, $\mathbf{Z}^{(1)} = \mathbf{Z}$ and $\mathbf{Z}^{(2)} = \mathbf{Z}\,\mathbf{Z}^\top$, $U \sim \mathcal{G}amma(\nu/2, \nu_0/2)$, $\mathbf{Z} \sim \mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu, \mathbb{A})$, $\alpha = \boldsymbol{\lambda}^\top \Omega^{-1/2}(\mathbf{Z} - \boldsymbol{\mu})$, $\zeta_1(x) = \phi(x)/\Phi(x)$ with $\phi(\cdot)$ the pdf of the standard normal distribution, and $\mathbb{A}$ is an hyperplane of the form $\mathbb{A} = \mathbb{A}_1 \times \ldots \times \mathbb{A}_p$ with $\mathbb{A}_k \in \{(-\infty, a_k], (a_k, \infty)\}, (a_k, \infty)\}$ for $\mathbf{a} = (a_1, \cdots, a_p)^\top$. *Proposition 1* hereafter will be useful for the derivation of $\overline{u_r}$, $\overline{u_r \mathbf{z}_s}$, $\overline{\tau_r}$ and $\overline{\tau_r \mathbf{z}_s}$.

Proposition 1 (see S3 Appendix for a proof) *Let* $\mathbf{Z} \sim \mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbb{A})$ *with* $\boldsymbol{\nu} = (\nu, \nu_0)^\top$, $U \sim \mathcal{G}amma(\nu/2, \nu_0/2)$ *and set* $\alpha = \boldsymbol{\lambda}^\top \Omega^{-1/2}(\mathbf{Z} - \boldsymbol{\mu})$. *Then, for any real* $r > -\boldsymbol{\nu}$ *and an integrable function* $g(\cdot)$ *of* $\mathbf{Z}$:

$$\mathrm{E}\{U^{r/2}g(\mathbf{Z})\} = C_r(\boldsymbol{\nu})\alpha_{st}^{-1}\alpha_{u,r}\ \mathrm{E}\{g(\mathbf{Z}_{u,r})\} \tag{14}$$

$$\mathrm{E}\{U^{r/2}\zeta_1(U^{1/2}\alpha)g(\mathbf{Z})\} = cMC_r(\boldsymbol{\nu})\alpha_{st}^{-1}\alpha_{\tau,r}\ \mathrm{E}\{g(\mathbf{Z}_{\tau,r})\} \tag{15}$$

where $C_r(\nu) = \left(\frac{2}{\nu_0}\right)^{r/2}\frac{\Gamma\left(\frac{\nu+r}{2}\right)}{\Gamma(\nu/2)}$, $\alpha_{st} = \int_{\mathbb{A}} St_p\left(\mathbf{z}|\boldsymbol{\mu}, \frac{\nu_0}{\nu}\boldsymbol{\Omega}, \boldsymbol{\lambda}, \nu\right)d\mathbf{z}$, $\alpha_{\tau,r} = \int_{\mathbb{A}} t_p(\mathbf{z}|\boldsymbol{\mu}, \overline{\boldsymbol{\Omega}}^*, \nu + r)d\mathbf{z}$, *and* $\alpha_{u,r} = \int_{\mathbb{A}} St_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Omega}^*, \boldsymbol{\lambda}, \nu + r)d\mathbf{z}$; $c = \sqrt{2/\pi}$, $\mathbf{Z}_{u,r} \sim \mathcal{TST}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}^*, \boldsymbol{\lambda}, \nu + r, \mathbb{A})$, $\mathbf{Z}_{\tau,r} \sim \mathcal{TT}_p(\boldsymbol{\mu}, \overline{\boldsymbol{\Omega}}^*, \nu + r, \mathbb{A})$, $M = (1 + \boldsymbol{\delta}\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta})^{-1/2}$ *with* $\boldsymbol{\delta} = \left(\frac{\nu_0}{\nu}\right)^{1/2}(1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{-1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{\lambda}$, $\overline{\boldsymbol{\Omega}} = \frac{\nu_0}{\nu}\boldsymbol{\Omega} - \boldsymbol{\delta}\boldsymbol{\delta}^\top$, $\overline{\boldsymbol{\Omega}}^* = \frac{\nu}{\nu+r}\overline{\boldsymbol{\Omega}}$ *and* $\boldsymbol{\Omega}^* = \frac{\nu_0}{\nu+r}\boldsymbol{\Omega}$.

By the mean of a simple linear transformation, we obtain from *Proposition 1* the joint expectations $\overline{u_r}$, $\overline{u_r \mathbf{z}_s}$, $\overline{\tau_r}$ and $\overline{\tau_r \mathbf{z}_s}$ in terms of moments of a truncated multivariate skew t distribution.

Corollary 1 (see S4 Appendix for a proof) *Let* $\mathbf{Z} \sim \mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbb{A})$ *with* $\boldsymbol{\nu} = (\nu, \nu_0)^\top$, $U \sim \mathcal{G}amma(\nu/2, \nu_0/2)$, $\mathbf{a} \in \mathbb{R}^p$ *and* $\mathbb{A} = \mathbb{A}_1 \times \ldots \times \mathbb{A}_p$ *with* $\mathbb{A}_k = (-\infty, a_k]$ *or* $\mathbb{A}_k = (a_k, \infty)$. *Then, on setting* $\boldsymbol{\delta} = \left(\frac{\nu_0}{\nu}\right)^{1/2}(1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{-1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{\lambda}$, $\overline{\boldsymbol{\Omega}} = \frac{\nu_0}{\nu}\boldsymbol{\Omega} - \boldsymbol{\delta}\boldsymbol{\delta}^\top$, $\mathbf{A} = diag(A_1, \cdots, A_p)$ *with* $A_k = 1$ *if* $\mathbb{A}_k = (-\infty, a_k]$ *and* $A_k = -1$ *if* $\mathbb{A}_k = (a_k, \infty)$, $\mathbf{a}^* = \mathbf{A}\mathbf{a}$, $\boldsymbol{\mu}^* = \mathbf{A}\,\boldsymbol{\mu}$, $\boldsymbol{\Omega}^* = \frac{\nu_0}{\nu}\mathbf{A}\boldsymbol{\Omega}\mathbf{A}$, $\overline{\boldsymbol{\Omega}}^* = \mathbf{A}\overline{\boldsymbol{\Omega}}\mathbf{A}$, $\boldsymbol{\lambda}^* = \mathbf{A}\boldsymbol{\lambda}$ *and* $\alpha_{st} = ST_p(\mathbf{a}^*|\boldsymbol{\mu}^*, \Omega^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu})$,

$$\overline{u_r} = C_r(\boldsymbol{\nu})\ \alpha_{st}^{-1}\ \alpha_{u,r} \tag{16}$$

$$\overline{u_r \mathbf{z}_1} = C_r(\boldsymbol{\nu})\ \alpha_{st}^{-1}\ \alpha_{u,r}\ \mathbf{A}\mathrm{E}\{\mathbf{X}_{u,r}\} \tag{17}$$

$$\overline{u_r \mathbf{z}_2} = C_r(\boldsymbol{\nu})\ \alpha_{st}^{-1}\ \alpha_{u,r}\ \mathbf{A}\mathrm{E}\{\mathbf{X}_{u,r}\mathbf{X}_{u,r}^\top\}\mathbf{A} \tag{18}$$

$$\overline{\tau_r} = cMC_r(\boldsymbol{\nu})\ \alpha_{st}^{-1}\ \alpha_{\tau,r} \tag{19}$$

$$\overline{\tau_r \mathbf{z}_1} = cMC_r(\boldsymbol{\nu})\ \alpha_{st}^{-1}\ \alpha_{\tau,r}\ \mathbf{A}\mathrm{E}\{\mathbf{X}_{\tau,r}\} \tag{20}$$

$$\overline{\tau_r \mathbf{z}_2} \quad = \quad cMC_r(\boldsymbol{\nu})\ \alpha_{st}^{-1}\ \alpha_{\tau,r}\ \mathbf{A}\mathrm{E}\{\mathbf{X}_{\tau,r}\mathbf{X}_{\tau,r}^{\top}\}\mathbf{A} \tag{21}$$

where $\mathbf{X}_{u,r} \sim \mathcal{TST}_p\left(\boldsymbol{\mu}^*, \frac{\nu}{\nu+r}\boldsymbol{\Omega}^*, \boldsymbol{\lambda}^*, \nu+r, \mathbf{a}^*\right)$, $\mathbf{X}_{\tau,r} \sim \mathcal{TT}_p\left(\boldsymbol{\mu}^*, \frac{\nu}{\nu+r}\overline{\boldsymbol{\Omega}}^*, \nu+r, \mathbf{a}^*\right)$, *and we have set* $\alpha_{u,r} = ST_p\left(\mathbf{a}^*|\boldsymbol{\mu}^*, \frac{\nu}{\nu+r}\boldsymbol{\Omega}^*, \boldsymbol{\lambda}^*, \nu+r\right)$ *and* $\alpha_{\tau,r} = T_p\left(\mathbf{a}^*|\boldsymbol{\mu}^*, \frac{\nu}{\nu+r}\overline{\boldsymbol{\Omega}}^*, \nu+r\right)$.

For a practical use of *Corollary 1*, the cumulative multivariate skew t distribution is required. To this end, the function *pmst* of the package *sn* [35] in R freeware [36] is an appropriate routine.

Moments of truncated multivariate skew generalized t distributions

The evaluation of expectations $\mathrm{E}\{\mathbf{X}_{u,r}^{(s)}\}$ involved in *Corollary 1* calls for general expressions for the first and second order moments of truncated multivariate SGT distributions. These moments are required in the implementation of an EM algorithm for a SGT distribution based binary data model. The moments have been derived for truncated multivariate t distributions by [34] and were used by [24] in their implementation of the EM algorithm for a t-link GLMM. We present in this section the expressions for the first two moments of the multivariate SGT distributions, relying on the Theorem 1 of [37] and the moments of truncated multivariate t distributions available from [34] (see also [38]).

Let $\mathbf{Z} \sim \mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{a})$ with $\nu = (\nu, \nu_0)^{\top}$ and $\mathbf{a} \in \mathbb{R}^p$, *i.e.* a *p−variate* SGT vector restricted to the right truncated hyperplane $\mathbb{A} = \{\mathbf{x} \in \mathbb{R}^p |\ x_1 \leq a_1, \ldots, x_p \leq a_p\}$. The pdf of $\mathbf{Z}$ is:

$$TSGt_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{a}) = \alpha_{st}^{-1}\ SGt_p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu})\ \mathrm{I}_{\mathbb{A}}(\mathbf{z}) \tag{22}$$

where $\alpha_{st} = SGT_p(\mathbf{a}|\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda}, \boldsymbol{\nu})$, $SGT_p(\cdot|\boldsymbol{\mu}, \Omega, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is the cdf of the *p−variate* SGT distribution. If $\boldsymbol{\mu} = 0$, Ω is a correlation matrix ($\Omega = \mathbf{R}$) and $\nu_0 = \nu$, then $\mathbf{Z} \sim \mathcal{TST}_p(0, \mathbf{R}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{a})$. In this case, the first two moments of $\mathbf{Z}$ can be evaluated using the following proposition which simply combines Theorem 3 in [34] with Theorem 1 in [37].

Proposition 2 (see S5 Appendix for a proof) *Let* $\mathbf{Z} \sim \mathcal{TST}_p(0, \mathbf{R}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{a})$ *with* $\mathbf{R}$ *a correlation matrix. Then,*

$$\mathrm{E}\{\mathbf{Z}\} \quad = \quad \frac{\nu\alpha_{st}^{-1}}{\nu-2}\left[C_0^*\boldsymbol{\delta} - \mathbf{R}q^*(\mathbf{a})\right] \text{ for } \quad \nu > 2, \text{ and for } \nu > 4 \tag{23}$$

$$\mathrm{E}\{\mathbf{Z}\mathbf{Z}^{\top}\} \quad = \quad \frac{\nu\alpha_{st}^{-1}}{\nu-2}\left[\alpha_{st}^*\mathbf{R} + \mathbf{R}(\mathbf{H}^* + \mathbf{D}^*)\mathbf{R} - \mathbf{R}\mathbf{H}_0^*\boldsymbol{\delta}^{\top} - \boldsymbol{\delta}\mathbf{H}_0^{*\top}\mathbf{R} + \mathbf{D}_0^*\boldsymbol{\delta}\boldsymbol{\delta}^{\top}\right] \tag{24}$$

where $C_0^* = \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu-2}{2}\right)(\nu\pi)^{1/2}} T_p\left(\mathbf{a}|0, \frac{\nu}{\nu-1}(\mathbf{R} - \boldsymbol{\delta}\boldsymbol{\delta}^{\top}), \nu-1\right)$ *with* $T_p(\cdot|\boldsymbol{\mu}, \Omega, \nu)$ *the cdf of the p−variate t distribution with location* $\boldsymbol{\mu}$, *scale* Ω *and degrees of freedom* $\boldsymbol{\nu}$; $q^*(\mathbf{a}) \in \mathbb{R}^p$ with i^{th} element $q_i^*(a_i) = \sqrt{\frac{\nu-2}{\nu}}\ t\left(\sqrt{\frac{\nu-2}{\nu}}a_i, \nu-2\right) T_p\left(\overline{\mathbf{a}}_2^{(i)}|a_i\overline{\mathbf{R}}_{12}^{(i)}, \frac{\nu+a_i^2}{\nu-1}\overline{\mathbf{R}}_{22.1}^{(i)}, \nu-1\right)$, $t(\cdot)$ *being the pdf of the standard Student t distribution;* $\alpha_{st}^* = T_{p+1}\left(\overline{\mathbf{a}}|0, \frac{\nu}{\nu-2}\mathbf{R}^*, \nu-2\right)$ *with* $\mathbf{R}^* = \begin{pmatrix} \sigma_0^2 & -\sigma_0\boldsymbol{\delta}^{\top} \\ -\sigma_0\boldsymbol{\delta} & \mathbf{R} \end{pmatrix}$, $\sigma_0 = \sqrt{1 + \boldsymbol{\lambda}^{\top}\boldsymbol{\lambda}}$ and $\boldsymbol{\delta} = \sigma_0^{-1}\mathbf{R}^{1/2}\boldsymbol{\lambda}$; $\mathbf{H}_0^* \in \mathbb{R}^p$ with i^{th} *element* $\mathbf{H}_{0i}^* = \frac{\nu}{\nu-4}\ t_2\left(a_1^{(0i)}|0, \frac{\nu}{\nu-4}\mathbf{R}_{11}^{(0i)}, \nu-4\right) T_{p-1}\left(\overline{\mathbf{a}}_2^{(0i)}|\overline{\boldsymbol{\mu}}_{2.1}^{(0i)}, \frac{\nu+\alpha_{0i}}{\nu-2}\overline{\mathbf{R}}_{22.1}^{(0i)}, \nu-2\right)$, $\alpha_{0i} = \frac{a_i^2}{(1-\boldsymbol{\delta}_i^2)}$; $\mathbf{H}^*$ *is the* $p \times p$

matrix with diagonal elements $\mathbf{H}^*_{ii} = 0$ *and off diagonal elements defined as*

$$\mathbf{H}^*_{ij} = \frac{v}{v-4}\ t_2\left(\mathbf{a}_1^{(ij)}|0, \frac{v}{v-4}\mathbf{R}_{11}^{(ij)}, v-4\right) T_{p-1}\left(\overline{\mathbf{a}}_2^{(ij)}|\overline{\boldsymbol{\mu}}_{2.1}^{(ij)}, \frac{v+\alpha_{ij}}{v-2}\overline{\mathbf{R}}_{22.1}^{(ij)}, v-2\right)$$

with $\alpha_{ij} = \frac{a_i^2 - 2\rho_{ij}a_ia_j + a_j^2}{(1-\rho_{ij}^2)}$; $\mathbf{D}^*_0 = \boldsymbol{\delta}^\top \mathrm{H}^*_0$; $\mathbf{D}^*$ *is the* $p \times p$ *diagonal matrix with diagonal elements* $\mathbf{D}^*_{ii} = \boldsymbol{\delta}_i \mathbf{H}^*_{0i} - a_i q^*_i(a_i) - \mathbf{R}_i \mathbf{H}^{*i}$, $\mathbf{H}^{*i}$ *denoting the* i^{th} *column of* $\mathbf{H}^*$; $\overline{\mathbf{a}} = (0, \mathbf{a})^\top$, $\mathbf{R}_{11}^{(0i)} = \begin{pmatrix} 1 & -\boldsymbol{\delta}_i \\ -\boldsymbol{\delta}_i & 1 \end{pmatrix}$, $\mathbf{R}_{11}^{(ij)} = \begin{pmatrix} 1 & \rho_{ij} \\ \rho_{ij} & 1 \end{pmatrix}$, $\overline{\mathbf{R}} = \begin{pmatrix} 1 & -\boldsymbol{\delta}^\top \\ -\boldsymbol{\delta} & \mathbf{R} \end{pmatrix}$, *with* δ_i *the* i^{th} *element of* $\boldsymbol{\delta}$, ρ_{ij} *the* $(ij)^{th}$ *element of* $\mathbf{R}$; $\mathbf{H}^i$ *the* i^{th} *column of* $\mathbf{H}$; $\overline{\mathbf{a}}_2^{(i)}$ *the vector* $\mathbf{a}$ *with its* $(i+1)^{th}$ *element (i.e.*a_i*) deleted;* $\overline{\mathbf{R}}_{12}^{(i)}$ *the* $(i+1)^{th}$ *column of* $\overline{\mathbf{R}}$ *with its* $(i+1)^{th}$ *element (i.e. 1) deleted;* $\overline{\mathbf{R}}_{22.1}^{(i)} = \overline{\mathbf{R}}_{22}^{(i)} - \overline{\mathbf{R}}_{12}^{(i)}\overline{\mathbf{R}}_{12}^{(i)T}$, $\overline{\mathbf{R}}_{22}^{(i)}$ *being* $\overline{\mathbf{R}}$ *with its* $(i+1)^{th}$ *row and column deleted;* $\mathbf{a}_1^{(ij)} = (a_i, a_j)^\top$; $\overline{\mathbf{a}}_2^{(ij)}$ *the vector* $\overline{\mathbf{a}}$ *with its* $(i+1)^{th}$ *and* $(j+1)^{th}$ *elements (i.e.*a_i and a_j*) deleted;* $\overline{\mathbf{R}}_{22.1}^{(ij)} = \overline{\mathbf{R}}_{22}^{(ij)} - \overline{\mathbf{R}}_{12}^{(ij)}[\mathbf{R}_{11}^{(ij)}]^{-1}\overline{\mathbf{R}}_{12}^{(ij)T}$, $\overline{\mathbf{R}}_{22}^{(ij)}$ *being* $\overline{\mathbf{R}}$ *with its* $(i+1)^{th}$ *and* $(j+1)^{th}$ *rows and columns deleted;* $\overline{\mathbf{R}}_{12}^{(ij)}$ *being the matrix* $\overline{\mathbf{R}}$ *with its* $(i+1)^{th}$ *and* $(j+1)^{th}$ *columns deleted, and only its* $(i+1)^{th}$ *and* $(j+1)^{th}$ *rows kept;* $\overline{\boldsymbol{\mu}}_{2.1}^{(ij)} = \overline{\mathbf{R}}_{12}^{(ij)}[\mathbf{R}_{11}^{(ij)}]^{-1}\mathbf{a}_1^{(ij)}$; $\mathbf{a}_1^{(0i)} = (0, a_i)^\top$; $\overline{\mathbf{a}}_2^{(0i)}$ *the vector* $\boldsymbol{a}$ *with its* i^{th} *element (i.e.*a_i*)* deleted; $\overline{\mathbf{R}}_{22.1}^{(0i)} = \overline{\mathbf{R}}_{22}^{(0i)} - \overline{\mathbf{R}}_{12}^{(0i)}[\mathbf{R}_{11}^{(0i)}]^{-1}\overline{\mathbf{R}}_{12}^{(0i)T}$, $\overline{\mathbf{R}}_{22}^{(0i)}$ *being* $\boldsymbol{R}$ *with its* i^{th} *row and column deleted;* $\overline{\mathbf{R}}_{12}^{(0i)}$ *being the matrix* $\overline{\mathbf{R}}$ *with its first and* $(i+1)^{th}$ *columns deleted, and only its first and* $(i+1)^{th}$ *rows kept; and* $\overline{\boldsymbol{\mu}}_{2.1}^{(0i)} = \overline{\mathbf{R}}_{12}^{(0i)}[\mathbf{R}_{11}^{(0i)}]^{-1}\mathbf{a}_1^{(0i)}$.

The following corollary gives the first two moments of a general right truncated SGT vector $\mathbf{Z} \sim \mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{a})$ with $\nu = (\nu, \nu_0)^\top$.

Corollary 2 *Let* $\mathbf{Z} \sim \mathcal{TSGT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mathbf{a})$ *with* $\boldsymbol{\nu} = (\nu, \nu_0)^\top$. *Then,*

$$\mathrm{E}\{\mathbf{Z}\} = \boldsymbol{\mu} + \Lambda \mathrm{E}\{\mathbf{X}\} \tag{25}$$

$$\mathrm{E}\{\mathbf{Z}\mathbf{Z}^\top\} = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \Lambda \mathrm{E}\{\mathbf{X}\}\boldsymbol{\mu}^\top + \boldsymbol{\mu}\mathrm{E}\{\mathbf{X}^\top\}\Lambda + \Lambda \mathrm{E}\{\mathbf{X}\mathbf{X}^\top\}\Lambda \tag{26}$$

where $\Lambda = \sqrt{\nu_0/\nu}\ \mathrm{diag}(\omega_1, \ldots, \omega_p)$, ω_i^2 *is the* i^{th} *diagonal element of* Ω, $\mathbf{X} \sim \mathcal{TST}_p(0, \mathbf{R}, \boldsymbol{\lambda}^*, \nu, \mathbf{a}^*)$, $\mathbf{R}$ *is the correlation matrix from* Ω, $\boldsymbol{\lambda}^* = \sqrt{\nu_0/\nu}\ \mathbf{R}^{-1/2}\Lambda^{-1}\boldsymbol{\Omega}^{1/2}\boldsymbol{\lambda}$, $\mathbf{a}^* = \Lambda^{-1}(\mathbf{a} - \boldsymbol{\mu})$ *and* $\mathrm{E}\{\mathbf{X}\}$ *and* $\mathrm{E}\{\mathbf{X}\mathbf{X}^\top\}$ *are available from Proposition 2.*

When $\nu \to \infty$, the truncated multivariate SGT family has the truncated multivariate skew normal family as a limiting case (see S5 Appendix for a definition and formulas for computing the first two moments).

Skew generalized t-link mixed binomial model

This section defines the skew generalized t-link model (SGTLM) and describes an Expectation-Maximization (EM) algorithm [39] accelerated using parameter expansion [27] for likelihood inference. Empirical Bayes estimators of random effects and weights are also obtained for cluster specific inference.

Model specification and marginal log-likelihood

The skew generalized t-link GLMM (SGTLM) considered in this work is defined as:

$$\begin{aligned} Y_{ij} = I_{(0,\infty)}(Z_{ij});\ \mathbf{Z}_i &\overset{ind}{\sim} \mathcal{SGT}_{n_i}(\boldsymbol{\eta}_i - c\tilde{U}_1 \upsilon_0 \delta_\varepsilon \mathbf{J}_{n_i}, \boldsymbol{\Omega}_{\varepsilon_i}, \boldsymbol{\delta}_\varepsilon \mathbf{J}_{n_i}, \boldsymbol{\nu}) \\ \mathbf{b}_i &\overset{ind}{\sim} \mathcal{ST}_q(-c\tilde{U}_1\boldsymbol{\delta}, \mathbf{D}, \boldsymbol{\lambda}, \nu) \end{aligned} \tag{27}$$

where Y_{ij} is the binary outcome of the j^{th} measurement ($j = 1, 2, \cdots, n_i$) on the i^{th} cluster ($i = 1, 2, \cdots, n$), $\mathbf{Z}_i$ is a latent continuous outcome which determines the observable $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$, and $\mathbf{b}_i$ is a vector of q random effects associated to the cluster i. In Eq (27), $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i$, $\boldsymbol{\beta}$, $\mathbf{X}_i$ and $\mathbf{W}_i$ are as in Eq (1); $c = \sqrt{2/\pi}$, $\tilde{U}_1 = \left(\frac{v}{2}\right)^{1/2}\frac{\Gamma\left(\frac{v-1}{2}\right)}{\Gamma(v/2)}$, $\boldsymbol{\delta}_\varepsilon \in \mathbb{R}$, $\mathbf{J}_{n_i}$ is the n_i–*vector* of all ones, $\boldsymbol{\Omega}_{\varepsilon_i} = \mathbf{I}_{n_i} + \boldsymbol{\delta}_\varepsilon^2\mathbf{J}_{n_i}\mathbf{J}_{n_i}^\top$, $v = (v, v\upsilon_0^2)^\top$ with $\upsilon_0 > 0$ and $v > 2$; $\mathbf{D} = \overline{\mathbf{D}} + \boldsymbol{\delta}\boldsymbol{\delta}^\top$ and $\boldsymbol{\lambda} = (1 + \boldsymbol{\delta}^\top\overline{\mathbf{D}}^{-1}\boldsymbol{\delta})^{1/2}\mathbf{D}^{-1/2}\boldsymbol{\delta}$ with $\boldsymbol{\delta} \in \mathbb{R}^p$ and $\overline{\mathbf{D}}$ a $q \times q$ positive define matrix.

In the SGTLM, the distribution of a single latent outcome Z_{ij} is $\mathcal{SGT}(\boldsymbol{\eta}_{ij} - c\tilde{U}_1\upsilon_0\boldsymbol{\delta}_\varepsilon, \omega_\varepsilon^2, \boldsymbol{\delta}_\varepsilon, v)$ where $\omega_\varepsilon^2 = 1 + \boldsymbol{\delta}_\varepsilon^2$ and $\mathcal{SGT}(\boldsymbol{\mu}, \omega^2, \boldsymbol{\lambda}, v)$ denotes a univariate SGT distribution with location μ, scale ω^2, shape $\boldsymbol{\lambda}$ and degrees of freedom v. Therefore, on denoting $SGT(\cdot|\boldsymbol{\mu}, \omega^2, \delta, \boldsymbol{\nu})$ the cdf of a scalar SGT distribution $\mathcal{SGT}(\boldsymbol{\mu}, \omega^2, \boldsymbol{\delta}, \boldsymbol{\nu})$, the success probability of an outcome Y_{ij} given the random effects $\mathbf{b}_i$ is $P(Y_{ij} = 1|\mathbf{b}_i) = SGT(0|\boldsymbol{\eta}_{ij} - c\tilde{U}_1\upsilon_0\boldsymbol{\delta}_\varepsilon, \omega^2, \boldsymbol{\delta}_\varepsilon, \boldsymbol{\nu})$. Unlike in the common probit model (PM) (see Eq (1)), for a given cluster i, the n_i latent outcomes Z_{ij} are not independent given the random effects $\mathbf{b}_i$. Indeed, on using Eq (4) and setting $\tilde{U}_2 = \frac{v}{v-2}$, the variance-covariance matrix of $\mathbf{Z}_i$ given $\mathbf{b}_i$ is

$$\boldsymbol{\Sigma}_{\varepsilon_i} = \upsilon_0^2[\tilde{U}_2\mathbf{I}_{n_i} + (\tilde{U}_2 - c\tilde{U}_1^2)\delta_\varepsilon^2\mathbf{J}_{n_i}\mathbf{J}_{n_i}^\top] \tag{28}$$

so that the correlation coefficient between two elements Z_{ij} and Z_{ik} of $\mathbf{Z}_i$ with $k \neq j$ is $\rho_0 = \frac{(\tilde{U}_2 - c\tilde{U}_1^2)\boldsymbol{\delta}_\varepsilon^2}{\tilde{U}_2 + (\tilde{U}_2 - c\tilde{U}_1^2)\boldsymbol{\delta}_\varepsilon^2}$. Thus, conditional on random effects, the n_i latent outcomes in $\mathbf{Z}_i$ are uncorrelated only when $\delta_\varepsilon = 0$, *i.e.* a skewed link function implies correlated latent outcomes within a cluster i.

The positive constant υ_0 in the SGTLM controls the scale of the latent variable Z_{ij} and thus the scale of the model link function. Indeed, from Eq (28), the conditional variance of Z_{ij} is $Var\{Z_{ij}|\mathbf{b}_i\} = \upsilon_0^2[\tilde{U}_2 + (\tilde{U}_2 - c^2\tilde{U}_1^2)\boldsymbol{\delta}_\varepsilon^2]$. Setting $\upsilon_0 = 1$ would yield a skew t-link model (*i.e.* $\boldsymbol{\nu} = (v, v)^\top$). However, to make fixed effects in the SGTLM comparable with fixed effects in the common probit model (PM) characterized by a link function with a unit scale (*i.e.* $Var\{Z_{ij}|\mathbf{b}_i\} = 1$), we have set

$$\upsilon_0 = [\tilde{U}_2 + (\tilde{U}_2 - c^2\tilde{U}_1^{\,2})\delta_\varepsilon^2]^{-1/2}. \tag{29}$$

The application of Eq (3) to $\mathbf{b}_i$ shows that, as in the PM, random effects in the SGTLM have null mean vector $\mathrm{E}\{\mathbf{b}_i\} = 0$. Using Eq (4), the variance-covariance matrix of random effects is given by

$$\boldsymbol{\Sigma}_b = \tilde{U}_2\overline{\mathbf{D}} + (\tilde{U}_2 - c^2\tilde{U}_1^2)\boldsymbol{\delta}\boldsymbol{\delta}^\top. \tag{30}$$

When $\delta_\varepsilon = 0$ and $\boldsymbol{\delta} = \mathbf{0}$, the SGTLM is reduced to the t-link model in [24] except $\upsilon_0 = 1$ therein. As $v \to \infty$ (so that $U_i = 1$), the STGLM has as limiting case the mixed skew-probit model (SPM) which reduces to the PM for $\delta_\varepsilon = 0$ and $\boldsymbol{\delta} = \mathbf{0}$.

By Eq (2), the SGTLM has the stochastic representation

$$\begin{aligned}
Y_{ij} &= I_{(0,\infty)}(Z_{ij}) \\
\mathbf{Z}_i|\mathbf{b}_i, U_i = u_i, V_i = v_i &\overset{ind}{\sim} \mathcal{N}_{n_i}(\boldsymbol{\eta}_i + [v_i u_i^{-1/2} - c\tilde{U}_1]\upsilon_0\delta_\varepsilon\mathbf{J}_{n_i}, \upsilon_0^2 u_i^{-1}\mathbf{I}_{n_i}) \\
\mathbf{b}_i|U_i = u_i, V_i = v_i &\overset{ind}{\sim} \mathcal{N}_q([v_i u_i^{-1/2} - c\tilde{U}_1]\boldsymbol{\delta}, u_i^{-1}\overline{\mathbf{D}}) \\
V_i &\overset{ind}{\sim} \mathcal{HN}(0,1) \quad \text{and} \\
U_i &\overset{ind}{\sim} \mathcal{G}amma(v/2, v/2)
\end{aligned} \tag{31}$$

where U_i and V_i are independent. In this representation of the SGTLM in terms of more common distributions, the n_i binary outcomes Y_{ij} of a cluster i are independent given the random effects $\mathbf{b}_i$, the scale mixing variable U_i and the half normal variable V_i. Note that given U_i and V_i, $\mathbf{Z}_i|\mathbf{b}_i$ and $\mathbf{b}_i$ are normally distributed and share the same U_i and V_i. As a result, the joint distribution of $\mathbf{Z}_i$ and $\mathbf{b}_i$ belongs to the class of SGT distributions. From the stochastic representation in Eq (31), we obtain the unconditional distributions of $\mathbf{Z}_i$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ as follows.

Proposition 3 (see S6 Appendix for a proof). *Let us consider the latent vector* $\mathbf{Z}_i$ *and the binary variable* Y_{ij} *in Eq (27) and define* $\boldsymbol{\mu}_i = \mathbf{X}_i\beta - c\tilde{U}_1\boldsymbol{\Delta}_i$ *with* $\boldsymbol{\Omega}_i = \overline{\boldsymbol{\Omega}}_i + \boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^\top$, $\overline{\boldsymbol{\Omega}}_i = \upsilon_0^2\mathbf{I}_{n_i} + \mathbf{W}_i\overline{\mathbf{D}}\mathbf{W}_i^\top$, $\boldsymbol{\Delta}_i = \upsilon_0\delta_\varepsilon\mathbf{J}_{n_i} + \mathbf{W}_i\boldsymbol{\delta}$, *and the related shape parameter* $\boldsymbol{\lambda}_i = \boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\Delta}_i(1 + \boldsymbol{\Delta}_i^\top\overline{\boldsymbol{\Omega}}_i^{-1}\boldsymbol{\Delta}_i)^{1/2}$. *Then* $\mathbf{Z}_i \sim \mathcal{ST}_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i, \nu)$. *Furthermore, the vector of binary outcomes* $\mathbf{Y}_i$ *has a multivariate Bernoulli distribution with joint probability mass function,*

$$f_{\mathbf{Y}}(\mathbf{y}_i|\boldsymbol{\beta}, \delta_\varepsilon, \boldsymbol{\delta}, \overline{\mathbf{D}}, \nu) = ST_{n_i}(\mathbf{0}|\mathbf{A}_i\boldsymbol{\mu}_i, \mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i, \mathbf{A}_i\boldsymbol{\lambda}_i, \nu) \quad (32)$$

and Y_{ij} *has a Bernoulli distribution with success probability* $P(Y_{ij} = 1) = ST(\boldsymbol{\mu}_{ij}|0, \omega_{ij}^2, -\boldsymbol{\lambda}_{ij}, \nu)$ *and probability mass function,*

$$f_Y(y_{ij}|\boldsymbol{\beta}, \delta_\varepsilon, \boldsymbol{\delta}, \overline{\mathbf{D}}, \nu) = ST(0|A_{ij}\mu_{ij}, \omega_{ij}^2, A_{ij}\lambda_{ij}, \nu) \quad (33)$$

where $\mathbf{A}_i = diag(A_{i1}, \ldots, A_{in_i})$, $A_{ij} = 1 - 2y_{ij}$, ω_{ij}^2 *is the* j^{th} *diagonal element of* $\boldsymbol{\Omega}_i$, $\boldsymbol{\lambda}_{ij} = \boldsymbol{\Delta}_{ij}/\overline{\omega}_{ij}$, *and* $\overline{\omega}_{ij}^2 = \omega_{ij}^2 - \boldsymbol{\Delta}_{ij}^2$ *with* Δ_{ij} *and* μ_{ij} *the* j^{th} *elements of* $\boldsymbol{\Delta}_i$ *and* $\boldsymbol{\mu}_i$ *respectively.*

Eq (32) conveniently expresses for a value $\mathbf{y}_i$ of $\mathbf{Y}_i$, the likelihood as a cumulative probability of a ST distribution whose location, scale and shape parameters depend on $\mathbf{y}_i$, using the identities $P(Z_{ij} > z_{ij}) = P(-Z_{ij} < -z_{ij})$ and $sign(Z_{ij}) = 2Y_{ij} - 1$ where $sign(\cdot)$ returns the sign of its argument. On using Eq (4) on the distribution of $\mathbf{Z}_i$ given in *Proposition 3*, the variance-covariance matrix of the outcomes at the latent scale is

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{\varepsilon_i} + \mathbf{W}_i\boldsymbol{\Sigma}_b\mathbf{W}_i^\top + \upsilon_0\delta_\varepsilon(\tilde{U}_2 - c\tilde{U}_1^2)[\mathbf{J}_{n_i}\boldsymbol{\delta}^\top\mathbf{W}_i^\top + \mathbf{W}_i\boldsymbol{\delta}\mathbf{J}_{n_i}^\top]. \quad (34)$$

Thus, in a model with a cluster-specific random intercept ($q = 1$) with $\boldsymbol{\delta} = 0$ and $\overline{\mathbf{D}} = \overline{\sigma}_b^2$, the latent intra-class correlation coefficient (the proportion of variance explained by clustering at latent scale) is given by

$$\rho = \frac{\tilde{U}_2\overline{\sigma}_b^2 + (\tilde{U}_2 - c\tilde{U}_1^2)\upsilon_0^2\delta_\varepsilon^2}{\tilde{U}_2(\upsilon_0^2 + \overline{\sigma}_b^2) + (\tilde{U}_2 - c\tilde{U}_1^2)\upsilon_0^2\delta_\varepsilon^2}. \quad (35)$$

The joint distribution of $\mathbf{Z}_i$ and $\mathbf{b}_i$ (*i.e.* $(\mathbf{Z}_i^\top, \mathbf{b}_i^\top)^\top$) is $\mathcal{ST}_{n_i+q}(\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Omega}}_i, \tilde{\boldsymbol{\lambda}}_i, \nu)$ where $\tilde{\boldsymbol{\mu}}_i = \begin{pmatrix} \mathbf{X}_i\beta - c\tilde{U}_1\boldsymbol{\Delta}_i \\ -c\tilde{U}_1\boldsymbol{\delta} \end{pmatrix}$, $\tilde{\boldsymbol{\Omega}}_i = \overline{\tilde{\boldsymbol{\Omega}}}_i + \tilde{\boldsymbol{\delta}}_i\tilde{\boldsymbol{\delta}}_i^\top$, $\overline{\tilde{\boldsymbol{\Omega}}}_i = \begin{pmatrix} \overline{\boldsymbol{\Omega}}_i & \mathbf{W}_i\overline{\mathbf{D}} \\ \overline{\mathbf{D}}\mathbf{W}_i^\top & \overline{\mathbf{D}} \end{pmatrix}$, $\tilde{\boldsymbol{\delta}}_i = \begin{pmatrix} \boldsymbol{\Delta}_i \\ \boldsymbol{\delta} \end{pmatrix}$ and $\tilde{\boldsymbol{\lambda}}_i = \tilde{\boldsymbol{\Omega}}_i^{-1/2}\tilde{\boldsymbol{\delta}}_i(1 + \tilde{\boldsymbol{\delta}}_i^\top\overline{\tilde{\boldsymbol{\Omega}}}_i^{-1}\tilde{\boldsymbol{\delta}}_i)^{1/2}$. Thus, for $j = 1, 2, \cdots, n_i$ and $k = 1, 2, \cdots, q$, the correlation between Z_{ij} and b_k is $\rho_{ijk} = \frac{\sigma_{ijk}}{\sigma_{ij}\sigma_k}$ with $\sigma_k^2 = \tilde{U}_2\overline{D}_k^2 + (\tilde{U}_2 - c^2\tilde{U}_1^2)\boldsymbol{\delta}_k^2$ the variance of b_k, $\sigma_{ij}^2 = \tilde{U}_2(\upsilon_0 + \mathbf{W}_{ij}\overline{\mathbf{D}}\mathbf{W}_{ij}^\top) + (\tilde{U}_2 - c^2\tilde{U}_1^2)(\upsilon_0\delta_\varepsilon + \mathbf{W}_{ij}\boldsymbol{\delta})^2$ the variance of Z_{ij} and $\sigma_{ijk} = \tilde{U}_2\mathbf{W}_{ij}\overline{\mathbf{D}}_k + (\tilde{U}_2 - c^2\tilde{U}_1^2)(\upsilon_0\delta_\varepsilon + \mathbf{W}_{ij}\boldsymbol{\delta})\delta_k$ is the covariance between Z_{ij} and b_k, $\mathbf{W}_{ij}$ is the j^{th} row of $\mathbf{W}_i$ and $\overline{\mathbf{D}}_k$ is the k^{th} column of $\overline{\mathbf{D}}$, $\overline{D}_k^2$ is the k^{th} diagonal element of $\overline{\mathbf{D}}$ and δ_k is the k^{th} element of $\boldsymbol{\delta}$.

The parameters of the SGTLM include $\boldsymbol{\beta}$, δ_ε, $\boldsymbol{\delta}$, $vech(\overline{\mathbf{D}})$ and ν where the $vech(\cdot)$ operator returns the lower triangle elements of its matrix argument. In order to avoid non-regular likelihood problems occurring in models based on the Student distribution and its extensions (in

particular when v is close to zero) [40], we follow some recent related works [24, 41] and first consider v as known, focusing on $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \delta_\varepsilon, \boldsymbol{\delta}^\top, vech(\overline{\mathbf{D}})^\top)^\top$. Classical inference on $\boldsymbol{\theta}$ is based on the marginal likelihood of the observed data $\mathbf{y}$. Using *Proposition 3*, the joint marginal log-likelihood of n independent clusters $\mathbf{y}_i$ ($i = 1, 2, \cdots, n$) is:

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \log f_i(\mathbf{y}_i|\boldsymbol{\theta}) \tag{36}$$

From Eq (36), an optimization routine like the R function *optim* can be used for inference on $\boldsymbol{\theta}$. We however develop an EM algorithm to circumvent the n_i-dimensional integral in Eq (32) when estimating $\boldsymbol{\theta}$.

Model identifiability

Estimations in the skew generalized t-link model (SGTLM) may produce inconsistent results which would induce unreliable and misleading conclusions, if the model is not identifiable. It is thus of great importance to check whether different points in the parameter space can be distinguished from observations $\mathbf{y}_i$. We analyse in this section the identifiability of the SGTLM and indicate when it is necessary to restrict the parameter space to ensure reliable inference from observed data. We restrict attention to the case $v_0 = 1$ (ignoring Eq (29)) since v_0 is an artificial device only included to ensure a unit variance in the conditional link function as in the traditional probit mixed model.

Although not sufficient, the identifiability of the SGTLM requires both the marginal random effects distribution and the conditional model given random effects to be identifiable. The identifiability of the random effects distribution follows from the identifiability of multivariate skew t distributions. We survey the identifiability of the conditional model before turning to the marginal model.

• *Conditional identifiability*

Conditional on the random effects $\mathbf{b}_i$ and for fixed degrees of freedom parameter v, the identifiability of the SGTLM reduces to the identifiability of the fixed effects skew-probit model. This follows because the skew t inverse link function is an average of the skew-probit inverse link function with respect to the gamma mixing distribution. The identifiability of the skew-probit model with one covariate has been recently shown to depend on the nature (binary/continuous) of the covariate in the model [42]. Indeed, the fixed effect skew-probit model is not identifiable in the absence of any covariate (*i.e.* each $\mathbf{X}_i$ is a column of n_i ones) [42] (page 1624, Proposition 2.1) or in the presence of a binary covariate [42] (page 1626, Proposition 2.2). On the other hand, the fixed effect skew-probit model is identifiable when the covariate is continuous [42] (page 1627, Proposition 2.3). Extension to the case of multiple covariates is straightforwardly obtained by requiring the covariate matrix $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top, \ldots, \mathbf{X}_n^\top)^\top$ to be of full column rank as in the classical linear regression model context. Whenever binary covariates are considered or no covariate is considered, we advocate to set $\delta_\epsilon = 0$ so that the conditional model reduces to a classical probit model.

• *Marginal identifiability*

From *Proposition 3*, it appears that when $v_0 = 1$ the paramaters δ_ε and $\boldsymbol{\delta}$ enter the marginal distribution of $\mathbf{Y}_i$ only through the marginal working shape $\boldsymbol{\Delta}_i = \delta_\varepsilon \mathbf{J}_{n_i} + \mathbf{W}_i\boldsymbol{\delta}$ whose j^{th} element can be written $\boldsymbol{\Delta}_{ij} = \delta_\epsilon + \mathbf{W}_{ij}^\top\boldsymbol{\delta}$. As a result, caution is required when learning the model parameters from some realizations $\mathbf{y}_i$ of $\mathbf{Y}_i$. Indeed, if the model includes a random intercept term, *i.e.*$\mathbf{W}_{ij}$ has the form $\mathbf{W}_{ij} = (1, \mathbf{W}_{1ij}^\top)^\top$ where $\mathbf{W}_{1ij} \in \mathbb{R}^{q-1}$, we can partition the random effects working shape as $\boldsymbol{\delta} = (\delta_0, \boldsymbol{\delta}_1^\top)^\top$ with $\boldsymbol{\delta}_1 \in \mathbb{R}^{q-1}$ so that the j^{th} element of the marginal

working shape reads $\boldsymbol{\Delta}_{ij} = \delta_\epsilon + \delta_0 + \mathbf{W}_{1ij}^\top \boldsymbol{\delta}_1$. Therefore, only the sum $\delta_\epsilon + \delta_0$ could be estimated and it would not be possible to distinguish in the observed skewness the part due to the random intercept from the part due to the conditional link function. This confounding issues may actually be avoided by considering more complex models based for instance on the fundamental skew distributions [43]. Recall that skewness is introduced in the SGTLM through a hidden standard half normal variable, namely V_i. As opposed to the unique standard half normal variable V_i used for both $\mathbf{Z}_i$ and $\mathbf{b}_i$ in the SGTLM, the use of two different standard half normal hidden variables for $\mathbf{Z}_i$ and $\mathbf{b}_i$ [44] (page 667 eq 5-6) or two different standard half multivariate normal hidden vectors [45] (page 420 eq 2.2) remove the confounding problem.

Fortunately, the non identifiability of the skewness of conditional link function and random intercept does not affect the success probability of the response since this only depends on $\boldsymbol{\Delta}_i$, but not on the individual values of δ_ϵ and δ_0. However, since the conditional link scale depends on δ_ϵ through v_0, the confounding problem affects the scale and thus the interpretation of fixed effects. Moreover, inference on the random intercept is affected since the random intercept variance and skewness depend on δ_0. For example, $\delta_\epsilon + \delta_0 = 0$ only indicates null marginal skewness, and in no way absence of link function and random intercept skewness which could be equally strong but of opposite signs. Thus, only a lower bound can be given to the random intercept variance: $\tilde{U}_2\overline{D}_{11} + (\tilde{U}_2 - c^2\tilde{U}_1^2)\boldsymbol{\delta}_0^2 \geq \tilde{U}_2\overline{D}_{11}$ where $\overline{D}_{11}$ is the first diagonal element of $\overline{\mathbf{D}}$. To rule out this peculiar situation where the model is not marginally fully identifiable, some previous works on skew normal/skew t distributions have considered the restriction $\delta_\epsilon = 0$ (regardless of the presence of a random intercep term) for instance in the context of linear mixed effects models [30, 46, 47] (page 1492 eq 2, page 4100 eq 4, and page 309 eq 3.2 respectively), multivariate measurement error models [48] (page 35, Eq 4.11) and non linear mixed effect models [49] (page 7 eq 10), but no argument was given to support this choice.

In the very common situation where the mixed model includes a random intercept term, prior information on the shape of the link function and/or the random intercept is required to place a meaningful restriction on the parameter space by setting for instance $\delta_\epsilon = 0$ or $\delta_0 = 0$ or $\delta_\epsilon = \delta_0$. In the absence of such information, we advocate to consider the restriction $\delta_0 = 0$ because the success probability of a response may exibit skewness, irrespective of the presence of random effects. This restriction thus allows to recover a fixed effects skew generalized t-link model when no random effect is considered. Overall, the restriction $\delta_0 = 0$ simply expresses the unability of the SGTLM to capture any additional skewness structure from the data through the inclusion of only random intercept. For completeness, we develop in the next section an estimation procedure for the full model in Eq (27), since the introduction of any equality restriction on δ_ϵ and δ_0 can be straightforwardly reduced to $\delta_0 = 0$.

The two restrictions discussed in this section are related to the structure of the quantity $\boldsymbol{\Delta}_i$ and are required only for some specific data structures (models including a binary covariate or models with a random intercept only). However, even when a restriction is required on δ_ϵ or the first element of $\boldsymbol{\delta}$, the quantity $\boldsymbol{\Delta}_i$ itself can remain unrestricted. When a restriction is required, it forces the skewness from a data to be summarized either by δ_ϵ or elements of $\boldsymbol{\delta}$. Overall, the SGT-link function is always allowed to be skewed (unconditional link). But some designs do not allow to distinguish skewness in the conditional link function from skewness in the distribution of random effects.

Maximum likelihood inference

Estimation via the EM algorithm. The choice of the value of v_0 in Eq (29) is in line with one of our purposes: rescale fixed effects so that they have the same interpretation as in the mixed probit model. There is no need to define v_0 depending on situations, because in our

proposal, v_0 is fixed. However, since v_0 is simply a scaling factor, it may be given any positive value during estimation, as long as the estimates are rescaled after the convergence of the estimation procedure so that v_0 is finally given by Eq (29). Indeed, because routines are basically written for the skew t distributions, we used $v_0 = 1$ in the EM algorithm and rescaled the estimates at the end of the procedure. Let us consider the complete data $\mathbf{y}_{com_i} = \{\mathbf{y}_i, \mathbf{z}_i, \mathbf{b}_i, u_i, v_i\}$. Because $\mathbf{y}_i$ only retains the signs of elements of $\mathbf{z}_i$, the joint density of $\mathbf{y}_i$ and $\mathbf{z}_i$ is $f_{\mathbf{YZ}}(\mathbf{y}_i, \mathbf{z}_i) = f_{\mathbf{Z}}(\mathbf{z}_i) \times I_{\mathbb{A}_i}(\mathbf{z}_i)$ where $\mathbb{A}_i = \mathbb{A}_{i1} \times \ldots \times \mathbb{A}_{in_i}$ with $\mathbb{A}_{ij} = (-\infty, 0]$ if $y_{ij} = 0$ and $\mathbb{A}_{ij} = (0, \infty)$ if $y_{ij} = 1$. The density of $\mathbf{y}_{com_i}$ is thus $f_{\mathbf{y}_{com}}(\mathbf{y}_{com_i}) = f_{\mathbf{Z},\mathbf{b},U,V}(\mathbf{z}_i, \mathbf{b}_i, u_i, v_i) \times I_{\mathbb{A}_i}(\mathbf{z}_i)$. Hence by Bayes's rule and in light of Eq (27) with $v_0 = 1$, the density of $\mathbf{y}_{com_i}$ is:

$$\begin{aligned} f_{\mathbf{y}_{com}}(\mathbf{y}_{com_i}) &= f_{\mathbf{z}|\mathbf{b},U,V}(\mathbf{z}_i|\mathbf{b}_i, u_i, v_i) \times f_{\mathbf{b}|U,V}(\mathbf{b}_i|u_i, v_i) \times f_U(u_i) \times f_V(v_i) \times I_{\mathbb{A}_i}(\mathbf{z}_i) \\ &= \phi_{n_i}(\mathbf{z}_i|\boldsymbol{\eta}_i + (v_i u_i^{-1/2} - c\tilde{U}_1)\boldsymbol{\delta}_\varepsilon \mathbf{J}_{n_i}, u_i^{-1}\mathbf{I}_{n_i}) \\ &\quad \times \phi_q(\mathbf{b}_i|(v_i u_i^{-1/2} - c\tilde{U}_1)\boldsymbol{\delta}, u_i^{-1}\overline{\mathbf{D}}) \times f_G(u_i|v/2, v/2) \\ &\quad \times f_V(v_i) \times I_{\mathbb{A}_i}(\mathbf{z}_i) \end{aligned} \tag{37}$$

where $f_V(v_i) = 2\phi(v_i)\, I_{(0,\infty)}(v_i)$ and $f_G(\cdot|v/2, v/2)$ is given in Eq (9). By Eq (37) and on setting $N = \sum_{i=1}^n n_i$ and $C_\ell = -\frac{N+n(q+1)}{2}\log(2\pi) + n\log 2$, the complete data log-likelihood is:

$$\begin{aligned} \ell(\boldsymbol{\theta}|\mathbf{y}_{com}) &= C_\ell - \frac{1}{2}\sum_{i=1}^n \left[\boldsymbol{\beta}^\top \mathbf{X}_i^\top \mathbf{X}_i \boldsymbol{\beta} u_i - 2\beta^\top \mathbf{X}_i^\top \gamma_i u_i + \gamma_i^\top \gamma_i u_i\right] - \frac{n}{2}\log|\overline{\mathbf{D}}| \\ &\quad + \frac{1}{2}\sum_{i=1}^n n_i \log u_i + \frac{q}{2}\sum_{i=1}^n \log u_i - \frac{1}{2}\sum_{i=1}^n tr\left(\overline{\mathbf{D}}^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)(v_i - c\tilde{U}_1 u_i^{1/2})^2 \\ &\quad - \frac{1}{2}\sum_{i=1}^n tr\left(\overline{\mathbf{D}}^{-1}\left[u_i \mathbf{b}_i \mathbf{b}_i^\top - \left(v_i - c\tilde{U}_1 u_i^{1/2}\right)\left(u_i^{1/2}\mathbf{b}_i\boldsymbol{\delta}^\top + \boldsymbol{\delta} u_i^{1/2}\mathbf{b}_i^\top\right)\right]\right) \\ &\quad + \sum_{i=1}^n \log f_G(u_i|v/2, v/2) - \frac{1}{2}\sum_{i=1}^n v_i^2 + \sum_{i=1}^n \log I_{\mathbb{A}_i}(\mathbf{z}_i) \end{aligned} \tag{38}$$

where $\gamma_i = \mathbf{z}_i - \mathbf{W}_i\mathbf{b}_i - \boldsymbol{\delta}_\varepsilon(v_i u_i^{-1/2} - c\tilde{U}_1)\mathbf{J}_{n_i}$, and $tr(\cdot)$ is the trace operator. Let $\widehat{\boldsymbol{\theta}}^{(k)}$ the estimate of $\boldsymbol{\theta}$ at the k^{th} EM iteration. The E-step of the $(k+1)^{th}$ iteration finds the expectation $Q(\cdot|\widehat{\boldsymbol{\theta}}^{(k)})$ of $\ell(\cdot|\mathbf{y}_{com})$ given the observed data $\mathbf{y}$ and the current parameter estimate $\widehat{\boldsymbol{\theta}}^{(k)} = (\widehat{\boldsymbol{\beta}}^{(k)\top}, \widehat{\delta}_\varepsilon^{(k)}, \widehat{\boldsymbol{\delta}}^{(k)\top}, vech(\widehat{\overline{\mathbf{D}}}^{(k)})^\top)^\top$:

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= C_\ell - \frac{1}{2}\sum_{i=1}^n \left[\boldsymbol{\beta}^\top \mathbf{X}_i^\top \mathbf{X}_i \boldsymbol{\beta}\widehat{u_2}_i^{(k)} - 2\boldsymbol{\beta}^\top \mathbf{X}_i^\top \widehat{u_2\gamma}_i + tr\left(\widehat{u_2\gamma_2}_i^{(k)}\right)\right] \\ &\quad - \frac{n}{2}\log|\overline{\mathbf{D}}| - \frac{1}{2}\sum_{i=1}^n tr\left(\overline{\mathbf{D}}^{-1}\left[\widehat{u_2\mathbf{b}_2}_i^{(k)} + \left(c\tilde{U}_1\widehat{u_2\mathbf{b}}_i^{(k)} - \widehat{v u\mathbf{b}}_i^{(k)}\right)\boldsymbol{\delta}^\top\right.\right. \\ &\quad \left.\left. + \boldsymbol{\delta}(c\tilde{U}_1\widehat{u_2\mathbf{b}}_i^{(k)} - \widehat{v u\mathbf{b}}_i^{(k)})^\top\right]\right) + \frac{1}{2}\sum_{i=1}^n (n_i + q + v - 2)\widehat{\log u}_{2i}^{(k)} \\ &\quad - \frac{1}{2}\sum_{i=1}^n tr\left(\overline{\mathbf{D}}^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)\left(\widehat{v}_{2i}^{(k)} - 2c\tilde{U}_1\widehat{vu}_i^{(k)} + c^2\tilde{U}_1^2\widehat{u}_{2i}^{(k)}\right) \\ &\quad - \frac{v}{2}\sum_{i=1}^n \widehat{u}_{2i}^{(k)} - \frac{1}{2}\sum_{i=1}^n \widehat{v}_{2i}^{(k)} + \frac{nv}{2}\log\left(\frac{v}{2}\right) - n\log\Gamma\left(\frac{v}{2}\right) \end{aligned} \tag{39}$$

where $\widehat{u_2\gamma}_i^{(k)} = \widehat{u_2\mathbf{z}}_i^{(k)} - \mathbf{W}_i\widehat{u_2\mathbf{b}}_i^{(k)} - \delta_\varepsilon(\widehat{vu}_i^{(k)} - c\tilde{U}_1\widehat{u}_{2i}^{(k)})\mathbf{J}_{n_i}$ and

$$\begin{aligned}\widehat{u_2\gamma}_{2i}^{(k)} &= \widehat{u_2\mathbf{z}}_{2i}^{(k)} + \mathbf{W}_i\widehat{u_2\mathbf{b}}_{2i}^{(k)}\mathbf{W}_i^\top + \delta_\varepsilon^2(\widehat{v}_{2i}^{(k)} - 2c\tilde{U}_1\widehat{vu}_i^{(k)} + c^2\tilde{U}_1^2\widehat{u}_{2i}^{(k)})\mathbf{J}_{n_i}\mathbf{J}_{n_i}^\top \\ &\quad -\delta_\varepsilon[(\widehat{vu\mathbf{z}}_i^{(k)} - \mathbf{W}_i\widehat{vu\mathbf{b}}_i^{(k)}) - c\tilde{U}_1(\widehat{u_2\mathbf{z}}_i^{(k)} - \mathbf{W}_i\widehat{u_2\mathbf{b}}_i^{(k)})]\mathbf{J}_{n_i}^\top \\ &\quad -\delta_\varepsilon\mathbf{J}_{n_i}[(\widehat{vu\mathbf{z}}_i^{(k)} - \mathbf{W}_i\widehat{vu\mathbf{b}}_i^{(k)}) - c\tilde{U}_1(\widehat{u_2\mathbf{z}}_i^{(k)} - \mathbf{W}_i\widehat{u_2\mathbf{b}}_i^{(k)})]^\top \\ &\quad -[\mathbf{W}_i\ \widehat{u_2\mathbf{bz}}_i^{(k)} + (\widehat{u_2\mathbf{bz}}_i^{(k)})^\top\mathbf{W}_i^\top]\end{aligned}$$

so that we have

$$\begin{aligned}tr(\widehat{u_2\gamma}_{2i}^{(k)}) &= tr(\widehat{u_2\mathbf{z}}_{2i}^{(k)}) + tr(\mathbf{W}_i\widehat{u_2\mathbf{b}}_{2i}^{(k)}\mathbf{W}_i^\top) - 2tr(\mathbf{W}_i\widehat{u_2\mathbf{bz}}_i^{(k)}) \\ &\quad + n_i\delta_\varepsilon^2(\widehat{v}_{2i}^{(k)} - 2c\tilde{U}_1\widehat{vu}_i^{(k)} + c^2\tilde{U}_1^2\widehat{u}_{2i}^{(k)}) \\ &\quad - 2\boldsymbol{\delta}_\varepsilon\mathbf{J}_{n_i}^\top[(\widehat{vu\mathbf{z}}_i^{(k)} - \mathbf{W}_i\widehat{vu\mathbf{b}}_i^{(k)}) - c\tilde{U}_1(\widehat{u_2\mathbf{z}}_i^{(k)} - \mathbf{W}_i\widehat{u_2\mathbf{b}}_i^{(k)})].\end{aligned}$$

Note that the conditional expectation of $\log I_{\mathbb{A}_i}(\mathbf{z}_i)$ is 0 since given $\mathbf{y}_i$, $I_{\mathbb{A}_i}(\mathbf{z}_i|\mathbf{y}_i) = 1$. The E-step thus reduces to the computation of the conditional expectations $\widehat{u_2\mathbf{b}}_i^{(k)} = \mathrm{E}\{U_i\mathbf{b}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $\widehat{vu\mathbf{b}}_i^{(k)} = \mathrm{E}\{V_iU_i^{1/2}\mathbf{b}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $\widehat{u_2\mathbf{bz}}_i^{(k)} = \mathrm{E}\{U_i\mathbf{b}_i\mathbf{Z}_i^\top|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $\widehat{u_2\mathbf{b}}_{2i}^{(k)} = \mathrm{E}\{U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $\widehat{vu}_i^{(k)} = \mathrm{E}\{V_iU_i^{1/2}|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $\widehat{v}_{2i}^{(k)} = \mathrm{E}\{V_i^2|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $\widehat{u}_{2i}^{(k)} = \mathrm{E}\{U_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $\widehat{u_2\mathbf{z}}_i^{(k)} = \mathrm{E}\{U_i\mathbf{Z}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $\widehat{vu\mathbf{z}}_i^{(k)} = \mathrm{E}\{V_iU_i^{1/2}\mathbf{Z}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $\widehat{u_2\mathbf{z}}_{2i}^{(k)} = \mathrm{E}\{U_i\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$ and $\widehat{\log u}_{2i}^{(k)} = \mathrm{E}\{\log U_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(k)}\}$. The expressions for these expectations (except $\widehat{\log u}_{2i}^{(k)}$) are given in the following result where we have dropped the supraindex (k) for simplicity.

Proposition 4 (see S7 Appendix for a proof). *Consider the random variables Y_{ij}, $\mathbf{Z}_i$, $\mathbf{b}_i$, U_i, and V_i as defined in Eq (27) with $\upsilon_0 = 1$, and an update $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\delta}_\varepsilon, \widehat{\boldsymbol{\delta}}^\top, vech(\widehat{\mathbf{D}})^\top)^\top$ of the model parameter $\boldsymbol{\theta}$. Let $\widehat{\overline{\boldsymbol{\Omega}}}_i = \mathbf{I}_{n_i} + \mathbf{W}_i\widehat{\mathbf{D}}\mathbf{W}_i^\top$, $\widehat{\mathbf{r}}_i = \widehat{\mathbf{D}}\mathbf{W}_i^\top\widehat{\overline{\boldsymbol{\Omega}}}_i^{-1}$, $\widehat{\Lambda}_i = (\mathbf{I}_q - \widehat{\mathbf{r}}_i\mathbf{W}_i)\widehat{\mathbf{D}}$, $\widehat{\boldsymbol{\Delta}}_i = \widehat{\delta}_\varepsilon\mathbf{J}_{n_i} + \mathbf{W}_i\widehat{\boldsymbol{\delta}}$, $\widehat{\mathbf{s}}_i = \widehat{\boldsymbol{\delta}} - \widehat{\mathbf{r}}_i\widehat{\boldsymbol{\Delta}}_i$, $\widehat{\boldsymbol{\mu}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}} - c\tilde{U}_1\widehat{\boldsymbol{\Delta}}_i$, $\widehat{M}_i = (1 + \widehat{\boldsymbol{\Delta}}_i^\top\widehat{\overline{\boldsymbol{\Omega}}}_i^{-1}\widehat{\boldsymbol{\Delta}}_i)^{-1/2}$, $\widehat{\boldsymbol{\Omega}}_i = \widehat{\overline{\boldsymbol{\Omega}}}_i + \widehat{\boldsymbol{\Delta}}_i\widehat{\boldsymbol{\Delta}}_i^\top$, $\widehat{\boldsymbol{\lambda}}_i = \widehat{M}_i^{-1}\widehat{\boldsymbol{\Omega}}_i^{-1/2}\widehat{\boldsymbol{\Delta}}_i$, $\mathbf{Z}_{0i} = \widehat{\boldsymbol{\Omega}}_i^{-1/2}(\mathbf{Z}_i - \widehat{\boldsymbol{\mu}}_i)$. Then:*

$$\widehat{u_2\mathbf{b}}_i = \widehat{\mathbf{r}}_i(\widehat{u_2\mathbf{z}}_i - \widehat{u}_{2i}\mathbf{X}_i\widehat{\boldsymbol{\beta}}) + (\widehat{vu}_i - c\tilde{U}_1\widehat{u}_{2i})\widehat{\mathbf{s}}_i \tag{40}$$

$$\widehat{vu\mathbf{b}}_i = \widehat{\mathbf{r}}_i(\widehat{vu\mathbf{z}}_i - \widehat{vu}_i\mathbf{X}_i\widehat{\boldsymbol{\beta}}) + (\widehat{v}_{2i} - c\tilde{U}_1\widehat{vu}_i)\widehat{\mathbf{s}}_i \tag{41}$$

$$\widehat{u_2\mathbf{bz}}_i = \widehat{\mathbf{r}}_i(\widehat{u_2\mathbf{z}}_{2i} - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\ \widehat{u_2\mathbf{z}}_i^\top) + \widehat{\mathbf{s}}_i(\widehat{vu\mathbf{z}}_i - c\tilde{U}_1\widehat{u_2\mathbf{z}}_i)^\top \tag{42}$$

$$\begin{aligned}\widehat{u_2\mathbf{b}}_{2i} &= \widehat{\boldsymbol{\Lambda}}_i + (\widehat{v}_{2i} - 2c\tilde{U}_1\widehat{vu}_i + c^2\tilde{U}_1^2\widehat{u}_{2i})\widehat{\mathbf{s}}_i\widehat{\mathbf{s}}_i^\top \\ &\quad + \widehat{\mathbf{r}}_i[\widehat{u_2\mathbf{z}}_{2i} + \widehat{u}_{2i}\mathbf{X}_i\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top\mathbf{X}_i^\top - \widehat{u_2\mathbf{z}}_i\widehat{\boldsymbol{\beta}}^\top\mathbf{X}_i^\top - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\ \widehat{u_2\mathbf{z}}_i^\top]\widehat{\mathbf{r}}_i^\top \\ &\quad + \widehat{\mathbf{r}}_i[(\widehat{vu\mathbf{z}}_i - \widehat{vu}_i\mathbf{X}_i\widehat{\boldsymbol{\beta}}) - c\tilde{U}_1(\widehat{u_2\mathbf{z}}_i - \widehat{u}_{2i}\mathbf{X}_i\widehat{\boldsymbol{\beta}})]\widehat{\mathbf{s}}_i^\top \\ &\quad + \widehat{\mathbf{s}}_i[(\widehat{vu\mathbf{z}}_i - \widehat{vu}_i\mathbf{X}_i\widehat{\boldsymbol{\beta}}) - c\tilde{U}_1(\widehat{u_2\mathbf{z}}_i - \widehat{u}_{2i}\mathbf{X}_i\widehat{\boldsymbol{\beta}})]^\top\widehat{\mathbf{r}}_i^\top\end{aligned} \tag{43}$$

$$\widehat{vu}_i = \widehat{M}_i(\widehat{u_2\alpha}_i + \widehat{\tau}_i) \tag{44}$$

$$\widehat{vu\mathbf{z}}_i = \widehat{M}_i(\widehat{u_2\alpha\mathbf{z}}_i + \widehat{\tau\mathbf{z}}_i) \tag{45}$$

$$\widehat{v}_{2i} \quad = \quad \widehat{M}_i^2(1 + \widehat{\tau\alpha}_i + \widehat{u_2\alpha}_{2i}) \tag{46}$$

where $\widehat{u_2\alpha}_i = \widehat{M}_i\widehat{\boldsymbol{\Delta}}_i^\top\widehat{\boldsymbol{\Omega}}_i^{-1}(\widehat{u_2\mathbf{z}}_i - \widehat{u}_{2i}\widehat{\boldsymbol{\mu}}_i)$, $\widehat{\tau\alpha}_i = \widehat{M}_i\widehat{\boldsymbol{\Delta}}_i^\top\widehat{\boldsymbol{\Omega}}_i^{-1}(\widehat{\tau\mathbf{z}}_i - \widehat{\tau}_i\widehat{\boldsymbol{\mu}}_i)$, $\widehat{u_2\alpha\mathbf{z}}_i = \widehat{M}_i(\widehat{u_2\mathbf{z}_{2i}} - \widehat{u_2\mathbf{z}}_i\widehat{\boldsymbol{\mu}}_i^\top)\widehat{\boldsymbol{\Omega}}_i^{-1}\widehat{\boldsymbol{\Delta}}_i$, $\widehat{u_2\alpha}_{2i} = \widehat{M}_i^2[\widehat{u}_{2i}(\widehat{\boldsymbol{\Delta}}_i^\top\widehat{\boldsymbol{\Omega}}_i^{-1}\widehat{\boldsymbol{\mu}}_i)^2 + \widehat{\boldsymbol{\Delta}}_i^\top\widehat{\boldsymbol{\Omega}}_i^{-1}(\widehat{u_2\mathbf{z}_{2i}} - 2\widehat{u_2\mathbf{z}}_i\widehat{\boldsymbol{\mu}}_i^\top)\widehat{\boldsymbol{\Omega}}_i^{-1}\widehat{\boldsymbol{\Delta}}_i]$, *and the expectations* $\widehat{u_2\mathbf{z}}_i = \mathrm{E}\{U_i\mathbf{Z}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\}$, $\widehat{u_2\mathbf{z}_{2i}} = \mathrm{E}\{U_i\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\}$, $\widehat{\tau}_i = \mathrm{E}\{U_i^{1/2}\zeta_1(U_i^{1/2}\boldsymbol{\lambda}^\top\mathbf{Z}_{0i})|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\}$, $\widehat{\tau\mathbf{z}}_i = \mathrm{E}\{U_i^{1/2}\zeta_1(U_i^{1/2}\boldsymbol{\lambda}^\top\mathbf{Z}_{0i})\mathbf{Z}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\}$ *and* $\widehat{u}_{2i} = \mathrm{E}\{U_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\}$ *are to be evaluated directly using Corollary 1 applied to the conditional latent vector* $\mathbf{Z}_i|\mathbf{Y}_i = \mathbf{y}_i, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}} \sim \mathcal{TST}_{n_i}(\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Omega}}_i, \widehat{\boldsymbol{\lambda}}_i, \widehat{v}, \mathbb{A}_i)$, $\mathbb{A}_i = \mathbb{A}_{i1} \times \ldots \times \mathbb{A}_{in_i}$ *with* $\mathbb{A}_{ij} = (-\infty, 0]$ *if* $y_{ij} = 0$ *and* $\mathbb{A}_{ij} = (0, \infty)$ *if* $y_{ij} = 1$.

The M-step jointly maximizes $Q(\cdot|\widehat{\boldsymbol{\theta}}^{(k)})$ over $\boldsymbol{\theta}$. This yields the following updating expressions for $\boldsymbol{\theta}$.

Proposition 5 (see S8 Appendix for a proof) *Consider an identifiable SGTLM as defined in Eq (27) with* $v_0 = 1$; *and an estimate* $\widehat{\boldsymbol{\theta}}^{(k)}$ *of* $\boldsymbol{\theta}$. *Set* $\widehat{\mathbf{S}}_{1i}^{(k)} = \widehat{u_2\mathbf{z}}_i^{(k)} - \mathbf{W}_i\widehat{u_2\mathbf{b}}_i^{(k)}$, $\widehat{S}_{2i}^{(k)} = \widehat{vu}_i^{(k)} - c\tilde{U}_1\widehat{u}_{2i}^{(k)}$, $\widehat{S}_{3i}^{(k)} = \widehat{v}_{2i}^{(k)} - 2c\tilde{U}_1\widehat{vu}_i^{(k)} + c^2\tilde{U}_1^{2(k)}\widehat{u}_{2i}^{(k)}$, $\widehat{\mathbf{S}}_{4i}^{(k)} = \widehat{vu\mathbf{z}}_i^{(k)} - \mathbf{W}_i\widehat{vu\mathbf{b}}_i^{(k)}$, $\widehat{\mathbf{T}}_1^{(k)} = [\sum_{i=1}^n \widehat{u}_{2i}^{(k)}\mathbf{X}_i^\top\mathbf{X}_i]^{-1}$ *and* $\widehat{\mathbf{T}}_2^{(k)} = \sum_{i=1}^n \widehat{S}_{2i}^{(k)}\mathbf{J}_{n_i}^\top\mathbf{X}_i$. *At EM iteration* $(k+1)$, *the updates of* $\boldsymbol{\beta}$, δ_ε, $\boldsymbol{\delta}$ *and* $\overline{\mathbf{D}}$ *are given by*:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} \quad = \quad \widehat{\mathbf{T}}_1^{(k)}\sum_{i=1}^n \mathbf{X}_i^\top(\widehat{\mathbf{S}}_{1i}^{(k)} - \widehat{\delta}_\varepsilon^{(k+1)}\widehat{S}_{2i}^{(k)}\mathbf{J}_{n_i}) \tag{47}$$

$$\begin{aligned}\widehat{\delta}_\varepsilon^{(k+1)} \quad = \quad & \left[\widehat{\mathbf{T}}_2^{(k)}\widehat{\mathbf{T}}_1^{(k)}(\sum_{i=1}^n \mathbf{X}_i^\top\widehat{\mathbf{S}}_{1i}^{(k)}) - \sum_{i=1}^n \mathbf{J}_{n_i}^\top(\widehat{\mathbf{S}}_{4i}^{(k)} - c\tilde{U}_1\widehat{\mathbf{S}}_{1i}^{(k)})\right] \\ & \times \left[\widehat{\mathbf{T}}_2^{(k)}\widehat{\mathbf{T}}_1^{(k)}\widehat{\mathbf{T}}_2^{(k)} - \sum_{i=1}^n n_i\widehat{S}_{3i}^{(k)}\right]^{-1}\end{aligned} \tag{48}$$

$$\widehat{\boldsymbol{\delta}}^{(k+1)} \quad = \quad \left[\sum_{i=1}^n \widehat{S}_{3i}^{(k)}\right]^{-1}\sum_{i=1}^n(\widehat{vu\mathbf{b}}_i^{(k)} - c\tilde{U}_1\widehat{u_2\mathbf{b}}_i^{(k)}) \tag{49}$$

$$\widehat{\overline{\mathbf{D}}}^{(k+1)} = n^{-1}\left[\left(\sum_{i=1}^n \widehat{u_2\mathbf{b}}_{2i}^{(k)}\right) - \left(\sum_{i=1}^n \widehat{S}_{3i}^{(k)}\right)\widehat{\boldsymbol{\delta}}^{(k+1)}(\widehat{\boldsymbol{\delta}}^{(k+1)})^\top\right]. \tag{50}$$

At convergence of the EM algorithm, we obtain the estimate $\tilde{\boldsymbol{\theta}}(\boldsymbol{\nu})$ of $\boldsymbol{\theta}$. The corresponding estimate of the variance-covariance matrix of random effects is $\widehat{\Sigma}_b = \tilde{U}_2\widehat{\overline{\mathbf{D}}} + (\tilde{U}_2 - c^2\tilde{U}_1^2)\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}^\top$. More generally, when ν is not actually known, the M-step of the EM algorithm can be extended to include a profiled marginal log-likelihood maximization step. Indeed, at EM iteration k, we notice that the estimate $\tilde{\boldsymbol{\theta}}^{(k)}(\nu)$ of $\boldsymbol{\theta}$ depends on ν only through $\tilde{U}_1$ and the profiled marginal log-likelihood $L_\nu(\cdot)$ for ν can be obtained by simply substituting $\tilde{\boldsymbol{\theta}}^{(k)}(\nu)$ for $\boldsymbol{\theta}$ in Eq (36). We can thus find $\widehat{\nu}^{(k+1)}$ using a one dimensional optimization routine (*e.g. optimize* in R) to maximize $L_\nu(\cdot)$. Then, the update of the parameter $\boldsymbol{\theta}$ becomes $\widehat{\boldsymbol{\theta}}^{(k+1)} = \tilde{\boldsymbol{\theta}}^{(k)}(\widehat{\nu}^{(k+1)})$. The use of the profiled marginal log-likelihood instead of a profiled version of $Q(\cdot|\widehat{\boldsymbol{\theta}}^{(k)})$ can provide substantial gain of

efficiency [50] and mostly helps bypass the calculation of $\widehat{\log u_{2i}}^{(k)}$ which does not have any known closed form.

It is worthwhile noticing however that the inclusion of a profiled marginal log-likelihood maximization step would prevent the convergence of the whole estimation procedure if the marginal log-likelihood in Eq (36) as a function of ν is unbounded for a particular dataset. This issue especially when ν is close to zero. Another challenge associated to the estimation of ν is time. The use of this strategy requires a very fast routine to compute cumulative probabilities of the skew t distributions. As an alternative route to estimate ν, we point out the model selection approach of [51] (page 893). It consists in setting a grid of feasible values of ν and obtaining a sequence of estimates $\tilde{\boldsymbol{\theta}}(\nu)$ of $\boldsymbol{\theta}$. Then, the couple ν and $\tilde{\boldsymbol{\theta}}(\nu)$ maximizing the marginal log-like-lihood in Eq (36) is taken as the estimates of ν and $\boldsymbol{\theta}$.

Accelerating EM via parameter-expansion. Besides its attractiveness and stability for handling incomplete data models, the EM algorithm sometimes experiences slow convergence, which has motivated many methods to accelerate its linear convergence speed. Among popular EM accelerators, the so-called parameter-expanded (PX) EM algorithm was proposed by [27] to speed up convergence. Let us consider a complete data model $\boldsymbol{F}(\mathbf{y}_{com}|\boldsymbol{\theta})$. The PX-EM algorithm expands $\boldsymbol{F}(\mathbf{y}_{com}|\boldsymbol{\theta})$ to a larger model $\boldsymbol{F}_X(\mathbf{y}_{com}|\Theta)$ parameterized by $\Theta = (\boldsymbol{\theta}_\star^\top, \alpha^\top)^\top$ where $\boldsymbol{\theta}_\star$ plays in $\boldsymbol{F}_X(\mathbf{y}_{com}|\Theta)$ the role of $\boldsymbol{\theta}$ in $\boldsymbol{F}(\mathbf{y}_{com}|\boldsymbol{\theta})$ and $\boldsymbol{\alpha}$ is a working parameter. The use of the PX-EM algorithm requires that (1) $\boldsymbol{\alpha}$ admits a value $\boldsymbol{\alpha}_0$ that preserves the original complete data model and (ii) the observed-data model is preserved by a many-to-one reduction function $\boldsymbol{R}$: $\Theta \mapsto \boldsymbol{\theta} = \boldsymbol{R}(\Theta)$ which allows an unambigious recovering of $\boldsymbol{\theta}$ from Θ. We refer to [27] for more details. For the SGTLM, let us consider the following expanded complete data model obtained by including a working $q \times q$ scale matrix $\boldsymbol{\alpha}$ into the linear predictor as $\boldsymbol{\eta}_i = \mathbf{X}_i \beta_\star + \mathbf{W}_i \alpha \mathbf{b}_i$:

$$\begin{aligned}
Y_{ij} &= I_{(0,\infty)}(Z_{ij}) \\
\mathbf{Z}_i|\mathbf{b}_i, U_i = u_i, V_i = v_i &\overset{ind}{\sim} \mathcal{N}_{n_i}(\boldsymbol{\eta}_i + [v_i u_i^{-1/2} - c\tilde{U}_1]\boldsymbol{\delta}_{\varepsilon\star}\mathbf{J}_{n_i}, u_i^{-1}\mathbf{I}_{n_i}) \\
\mathbf{b}_i|U_i = u_i, V_i = v_i &\overset{ind}{\sim} \mathcal{N}_q([v_i u_i^{-1/2} - c\tilde{U}_1]\boldsymbol{\delta}_\star, u_i^{-1}\overline{\mathbf{D}}_\star) \\
U_i &\overset{ind}{\sim} \mathcal{G}amma(\nu_\star/2, \nu_\star/2) \\
V_i &\overset{ind}{\sim} \mathcal{HN}(0,1)
\end{aligned}$$

This expanded model equals the STGLM in Eq (27) when $\boldsymbol{\alpha}$ takes the value $\boldsymbol{\alpha}_0 = \mathbf{I}_q$, and has expanded parameter $\Theta = (\boldsymbol{\beta}_\star^\top, \delta_{\varepsilon\star}, \boldsymbol{\delta}_\star^\top, vech(\overline{\mathbf{D}}_\star)^\top, vec(\boldsymbol{\alpha})^\top)^\top$ where vec is the usual operator which stacks the columns of its matrix argument. Under this model, the marginal distribution of $\mathbf{Y}_i$ remains as given in Eq (32) with $\overline{\boldsymbol{\Omega}}_i = \mathbf{I}_{n_i} + \mathbf{W}_i \alpha \overline{\mathbf{D}}_\star \alpha^\top \mathbf{W}_i^\top$ and $\boldsymbol{\Delta}_i = \boldsymbol{\delta}_\varepsilon \mathbf{J}_{n_i} + \mathbf{W}_i \boldsymbol{\alpha}\boldsymbol{\delta}$ so that $\boldsymbol{\Theta}$ reduces as $\boldsymbol{\theta} = (\beta_\star^\top, \delta_{\varepsilon\star}, \boldsymbol{\delta}_\star^\top \boldsymbol{\alpha}^\top, vech(\boldsymbol{\alpha}\overline{\mathbf{D}}_\star\boldsymbol{\alpha}^\top)^\top)^\top$. As the observed-data model is preserved whatever the value of $\boldsymbol{\alpha}$, we fix $\boldsymbol{\alpha} = \mathbf{I}_q$ at each E-step of the EM procedure. Therefore, the E-step of the PX-EM algorithm uses *Proposition 4* to obtain conditional expectations required in Eq (39) as for the classical EM algorithm. At the M-step, the estimates of $\boldsymbol{\delta}_\star$ and $\overline{\mathbf{D}}_\star$ are still given by Eqs (49)–(50) respectively whereas the estimates of $\delta_{\varepsilon\star}$, $\boldsymbol{\beta}_\star$ and $vec(\boldsymbol{\alpha})$ are:

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_\star^{(k+1)} \\ vec(\widehat{\boldsymbol{\alpha}}^{(k+1)}) \end{pmatrix} = \Psi^{(k)}(\Upsilon^{(k)} - \widehat{\delta}_{\varepsilon\star}^{(k+1)}\boldsymbol{\xi}^{(k)}) \tag{51}$$

$$\begin{aligned}\widehat{\delta}_{\varepsilon\star}^{(k+1)} &= [\boldsymbol{\xi}^{(k)\top}\boldsymbol{\Psi}^{(k)}\boldsymbol{\Upsilon}^{(k)} - \sum_{i=1}^{n}\mathbf{J}_{n_i}^{\top}(\widehat{vu\mathbf{z}}_i^{(k)} - c\tilde{U}_1\widehat{u_2\mathbf{z}}_i^{(k)})] \\ &\quad \times[\boldsymbol{\xi}^{(k)\top}\boldsymbol{\Psi}^{(k)}\boldsymbol{\xi}^{(k)} - \sum_{i=1}^{n}n_i\widehat{S}_{3i}^{(k)}]^{-1}\end{aligned} \tag{52}$$

where $\boldsymbol{\xi}^{(k)} = \begin{pmatrix} \sum_{i=1}^{n}\widehat{S}_{2i}^{(k)}\mathbf{X}_i^{\top}\mathbf{J}_{n_i} \\ \sum_{i=1}^{n}(\mathbf{W}_i^{\top}\mathbf{J}_{n_i}) \otimes (\widehat{vu\mathbf{b}}_i^{(k)} - c\tilde{U}_1\widehat{u_2\mathbf{b}}_i^{(k)}) \end{pmatrix}$, $\boldsymbol{\Upsilon}^{(k)} = \begin{pmatrix} \sum_{i=1}^{n}\mathbf{X}_i^{\top}\widehat{u_2\mathbf{z}}_i^{(k)} \\ vec(\sum_{i=1}^{n}\mathbf{W}_i^{\top}\widehat{u_2\mathbf{b}\mathbf{z}}_i^{\top(k)}) \end{pmatrix}$, and $\boldsymbol{\Psi}^{(k)} = \begin{pmatrix} \sum_{i=1}^{n}\mathbf{X}_i^{\top}\mathbf{X}_i\widehat{u}_{2i}^{(k)} & \sum_{i=1}^{n}\mathbf{X}_i^{\top}(\widehat{u_2\mathbf{b}}_i^{(k)} \otimes \mathbf{W}_i^{\top})^{\top} \\ \sum_{i=1}^{n}(\widehat{u_2\mathbf{b}}_i^{(k)} \otimes \mathbf{W}_i^{\top})\mathbf{X}_i & \sum_{i=1}^{n}\widehat{u_2\mathbf{b}}_{2i}^{(k)} \otimes (\mathbf{W}_i^{\top}\mathbf{W}_i) \end{pmatrix}^{-1}$ with $\otimes$ the direct product operator. Using the reduction function, the original model parameter estimates can be recovered as $\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}_{\star}^{(k+1)}$, $\widehat{\delta}_{\varepsilon}^{(k+1)} = \widehat{\delta}_{\varepsilon\star}^{(k+1)}$, $\widehat{\boldsymbol{\delta}}^{(k+1)} = \widehat{\alpha}^{(k+1)}\widehat{\boldsymbol{\delta}}_{\star}^{(k+1)}$ and $\widehat{\mathbf{D}}^{(k+1)} = \widehat{\alpha}^{(k+1)}\widehat{\mathbf{D}}_{\star}^{(k+1)}\widehat{\boldsymbol{\alpha}}^{(k+1)\top}$. In the neighbourhood of the ML estimate of $\boldsymbol{\theta}$, the working scale estimate $\widehat{\boldsymbol{\alpha}}$ becomes close to $\boldsymbol{\alpha}_0 = \mathbf{I}_q$ [27] so that the advantage of the PX-EM algorithm over the classical EM algorithm disappears. We thus propose to stop the PX acceleration once $|\boldsymbol{\lambda}_{max}| < \epsilon$ with $\boldsymbol{\lambda}_{max}$ the dominant eigen value of $\widehat{\alpha}^{(k+1)} - \mathbf{I}_q$ and ϵ a pre-specified tolerance value (*e.g.* $\epsilon = 10^{-2}$).

Summary of the estimation procedure. The estimation procedure starts with a parameter $\widehat{\theta}^{(0)}$, $k = 0$ and iterates the following six steps until convergence.

1. E-step: compute conditional expectations defined by Eqs (40)–(46) with $\widehat{\boldsymbol{\theta}}^{(k)}$.
2. PX M-step: obtain $\widehat{\boldsymbol{\theta}}^{(k+1)}$ and $\widehat{\boldsymbol{\alpha}}^{(k+1)}$ using Eqs (49)–(52) and the reduction function: $\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}_{\star}^{(k+1)}$, $\widehat{\delta}_{\varepsilon}^{(k+1)} = \widehat{\delta}_{\varepsilon\star}^{(k+1)}$, $\widehat{\boldsymbol{\delta}}^{(k+1)} = \widehat{\boldsymbol{\alpha}}^{(k+1)}\widehat{\boldsymbol{\delta}}_{\star}^{(k+1)}$ and $\widehat{\mathbf{D}}^{(k+1)} = \widehat{\boldsymbol{\alpha}}^{(k+1)}\widehat{\mathbf{D}}_{\star}^{(k+1)}\widehat{\boldsymbol{\alpha}}^{(k+1)\top}$.
3. Test: compute $\boldsymbol{\lambda}_{max}$ the dominant eigen value of $\widehat{\boldsymbol{\alpha}}^{(k+1)} - \mathbf{I}_q$. If $|\boldsymbol{\lambda}_{max}| < 10^{-2}$ then compute the marginal likelihood $\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}|\mathbf{y})$ using Eq (36) and **go to 4)** with $k = k + 1$, otherwise **return to 1)** with $k = k + 1$.
4. E-step: compute conditional expectations defined by Eqs (40)–(46) with $\widehat{\boldsymbol{\theta}}^{(k)}$.
5. M-step: obtain $\widehat{\boldsymbol{\theta}}^{(k+1)}$ using Eqs (47)–(50).
6. Test: compute the marginal likelihood $\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}|\mathbf{y})$ using Eqs (36). If $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}|\mathbf{y}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)}|\mathbf{y})|/|\ell(\widehat{\boldsymbol{\theta}}^{(k)}|\mathbf{y})| < 10^{-6}$ then **go to 7)**, otherwise **return to 4)** with $k = k + 1$.
7. Rescaling: compute υ_0 using (29) and rescale the estimates as $\widehat{\boldsymbol{\beta}} \leftarrow \upsilon_0\widehat{\boldsymbol{\beta}}^{(k)}$, $\widehat{\boldsymbol{\delta}} \leftarrow \upsilon_0\widehat{\boldsymbol{\delta}}^{(k)}$ and $\widehat{\mathbf{D}} \leftarrow \upsilon_0^2\widehat{\mathbf{D}}^{(k)}$. Return $\widehat{\boldsymbol{\theta}} = (\widehat{\beta}^{\top}, \widehat{\boldsymbol{\delta}}_{\epsilon}, \widehat{\boldsymbol{\delta}}^{\top}, vech(\widehat{\mathbf{D}})^{\top})^{\top}$.

Approximate standard errors. With a view to allow asymptotic inference in SGTLM, we follow the empirical information-based method of [52] (pages 132-133) to compute the asymptotic variance-covariance matrix of the ML estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ under some general regularity conditions. The observed information matrix is defined to be $I_o(\widehat{\boldsymbol{\theta}}|\mathbf{y}) = \sum_{i=1}^{n}\widehat{\mathbf{g}}_i\widehat{\mathbf{g}}_i^{\top}$ where $\widehat{\mathbf{g}}_i = \mathbf{g}_i(\widehat{\boldsymbol{\theta}})$, $\mathbf{g}_i(\boldsymbol{\theta}) = \frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}$, $Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ being the contribution of the single observation $\mathbf{y}_i$ to the expected complete data log-likelihood in Eq (39). On setting $\widehat{\widehat{vu\mathbf{b}}}_i = \widehat{vu\mathbf{b}}_i - c\tilde{U}_1\widehat{u_2\mathbf{b}}_i$, the

elements $\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \beta}$, $\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\delta}_\varepsilon}$, $\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\delta}}$, $vech\left(\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \overline{\mathbf{D}}}\right)$ of the score $\mathbf{g}_i$ can be explicitly evaluated using:

$$\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \beta} = -[\widehat{u}_{2i}\mathbf{X}_i^\top \mathbf{X}_i\boldsymbol{\beta} - \mathbf{X}_i^\top(\widehat{\mathbf{S}}_{1i} - \boldsymbol{\delta}_\varepsilon \widehat{S}_{2i}\mathbf{J}_{n_i})], \tag{53}$$

$$\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\delta}_\varepsilon} = -[\widehat{S}_{2i}\mathbf{J}_{n_i}^\top \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\delta}_\varepsilon n_i \widehat{S}_{3i} - \mathbf{J}_{n_i}^\top(\widehat{\mathbf{S}}_{4i} - c\tilde{U}_1\widehat{\mathbf{S}}_{1i})], \tag{54}$$

$$\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\delta}} = -\overline{\mathbf{D}}^{-1}[\widehat{S}_{3i}\boldsymbol{\delta} - \widehat{\overline{vu}\mathbf{b}}_i], \tag{55}$$

$$\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \overline{\mathbf{D}}} = -\frac{1}{2}\overline{\mathbf{D}}^{-1} + \frac{1}{2}\overline{\mathbf{D}}^{-1}\left[\widehat{u_2\mathbf{b}_{2i}} - \boldsymbol{\delta}\widehat{\overline{vu}\mathbf{b}}_i^\top - \widehat{\overline{vu}\mathbf{b}}_i\boldsymbol{\delta}^\top + \widehat{S}_{3i}\boldsymbol{\delta}\boldsymbol{\delta}^\top\right]\overline{\mathbf{D}}^{-1}. \tag{56}$$

Afterwards, the standard errors of estimated model parameters are approximated by square roots of diagonal elements of $[I_o(\widehat{\boldsymbol{\theta}}|\mathbf{y})]^{-1}$ and confidence intervals can be built assuming asymptotic normality.

Empirical Bayes estimators of random effects and weights. In this section, we provide the empirical Bayes estimators of cluster specific random effects and weights that are useful for evaluating individual intercepts and slopes as well as detecting outlying individuals. From Eq (27), the distribution of $\mathbf{b}_i$ conditional on $\mathbf{Z}_i = \mathbf{z}_i$, $U_i = u_i$ and $V_i = v_i$ is multivariate normal with mean $\mathbf{r}_i(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta}) + (v_i u_i^{-1/2} - c\tilde{U}_1)\mathbf{s}_i$ and covariance matrix $u_i^{-1}\Lambda_i$ where $\mathbf{r}_i = \overline{\mathbf{D}}\mathbf{W}_i^\top\overline{\boldsymbol{\Omega}}_i^{-1}$, $\mathbf{s}_i = \delta - \mathbf{r}_i\Delta_i$, $\boldsymbol{\Delta}_i = \upsilon_0\boldsymbol{\delta}_\varepsilon\mathbf{J}_{n_i} + \mathbf{W}_i\boldsymbol{\delta}$, $\overline{\boldsymbol{\Omega}}_i = \upsilon_0^2\mathbf{I}_{n_i} + \mathbf{W}_i\overline{\mathbf{D}}\mathbf{W}_i^\top$, and $\Lambda_i = (\mathbf{I}_q - \mathbf{r}_i\mathbf{W}_i)\overline{\mathbf{D}}$. The conditional mean of $\mathbf{b}_i$ given $\mathbf{Y}_i = \mathbf{y}_i$ is thus:

$$\begin{aligned}\overline{\mathbf{b}}_i(\boldsymbol{\theta}) &= E\{\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i, \boldsymbol{\theta}\} = E\{E\{\mathbf{b}_i|\mathbf{Z}_i = \mathbf{z}_i, U_i = u_i, V_i = v_i, \boldsymbol{\theta}\}|\mathbf{Y}_i = \mathbf{y}_i, \boldsymbol{\theta}\} \\ &= \mathbf{r}_i(\overline{\mathbf{z}}_i - \mathbf{X}_i\boldsymbol{\beta}) + (\overline{vu_{-1}}_i - c\tilde{U}_1)\mathbf{s}_i\end{aligned} \tag{57}$$

where $\overline{vu_{-1}}_i = M_i(\overline{u\alpha}_i + \overline{\tau_{-1}}_i)$, $\overline{u\alpha}_i = M_i\Delta_i^\top\overline{\boldsymbol{\Omega}}_i^{-1}(\overline{u\mathbf{z}}_i - \overline{u}_i\boldsymbol{\mu}_i)$, $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} - c\tilde{U}_1\boldsymbol{\Delta}_i$ and the quantities $\overline{u}_i = \mathrm{E}\{U_i^{1/2}|\mathbf{y}_i, \boldsymbol{\theta}\}$, $\overline{u\mathbf{z}}_i = \mathrm{E}\{U_i^{1/2}\mathbf{Z}_i|\mathbf{y}_i, \boldsymbol{\theta}\}$, $\overline{\tau_{-1}}_i = \mathrm{E}\{U_i^{-1/2}\zeta_1(U_i^{1/2}\boldsymbol{\lambda}^\top\mathbf{Z}_{0i})|\mathbf{y}_i\}$ and $\overline{\mathbf{z}}_i = \mathrm{E}\{\mathbf{Z}_i|\mathbf{y}_i, \boldsymbol{\theta}\}$ are to be evaluated using *Corollary 1* applied to $\mathbf{Z}_i|\mathbf{Y}_i = \mathbf{y}_i \sim \mathcal{TST}_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i, \nu, \mathbb{A}_i)$ with $\boldsymbol{\Omega}_i = \overline{\boldsymbol{\Omega}}_i + \Delta_i\Delta_i^\top$, $\boldsymbol{\lambda}_i = (1 + \boldsymbol{\Delta}_i^\top\overline{\boldsymbol{\Omega}}_i^{-1}\boldsymbol{\Delta}_i)^{1/2}\boldsymbol{\Omega}_i^{-1/2}\boldsymbol{\Delta}_i$, $\mathbb{A}_i = \mathbb{A}_{i1} \times \ldots \times \mathbb{A}_{in_i}$, $\mathbb{A}_{ij} = (-\infty, 0]$ if $y_{ij} = 0$ and $\mathbb{A}_{ij} = (0, \infty)$ if $y_{ij} = 1$. The empirical Bayes estimators of $\mathbf{b}_i$ can then be obtained as $\widehat{\mathbf{b}}_i = \overline{\mathbf{b}}_i(\widehat{\boldsymbol{\theta}})$.

For outlying individuals detection, individual weights U_i are predicted by $\overline{u_2}_i = E\{U_i|\mathbf{y}_i, \boldsymbol{\theta}\}$ [53] which is given by Eq (16) in *Corollary 1* applied to $\mathbf{Z}_i|\mathbf{Y}_i = \mathbf{y}_i$. The empirical Bayes estimators of U_i are thus given by $\widehat{u}_{2i} = E\{U_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}\}$. Relatively low weights (< 1) are indicative of outlying individuals.

Applications

This section presents a simulation study for assessing performance of SGTLM, and an application of the modeling approach to a real dataset.

Simulation study

We conducted a simulation to evaluate the proposed approach to the analysis of correlated binary data. The simulation experiment targeted four specific objectives. First, it assessed for

different sample sizes, the abilities of the probit (PM), the skew-probit (SPM), the generalized t-link (GTLM) and the skew generalized t-link (SGTLM) models to recover population parameters when the common normality assumption for the link function is either violated or not. The widely used logistic model was not investigated as the logistic distribution can be considered as a special case of the Student t distribution [8] hence the logistic model is a special case of GTLM. Second, the experiment evaluated the extent to which asymptotic 95% confidence intervals ($CI_{95\%}$) can detect the presence of spurious skewness. Third, the experiment evaluated the ability of empirical Bayes estimators of random effects to predict true random effects. Finally, the simulation study assessed the ability of Akaike's information criterion (*AIC*), Schwarz's Bayesian information criterion (*BIC*) and Hannan-Quinn criterion (*HQ*) to select the correct model fit. All computations were performed in R.

Simulation design. Mimicking the structure of the simulation model studied in [24] (page 1116), we considered the following GLMM:

$$\begin{aligned} Y_{ij} &= I_{(0,\infty)}(Z_{ij}) \\ \mathbf{Z}_i|\mathbf{b}_i, U_i = u_i, V_i = v_i &\overset{ind}{\sim} \mathcal{N}_{n_i}(\boldsymbol{\eta}_i + [v_i u_i^{-1/2} - c\tilde{U}_1]\upsilon_0\delta_\varepsilon \mathbf{J}_{n_i}, \upsilon_0^2 u_i^{-1}\mathbf{I}_{n_i}) \\ \mathbf{b}_i|U_i = u_i, V_i = v_i &\overset{ind}{\sim} \mathcal{N}_2([v_i u_i^{-1/2} - c\tilde{U}_1]\boldsymbol{\delta}, u_i^{-1}\overline{\mathbf{D}}) \\ U_i &\overset{ind}{\sim} \mathcal{F}_U(v) \text{ and } V_i \overset{ind}{\sim} \mathcal{HN}(0,1) \end{aligned}$$

where $\boldsymbol{\eta}_i = (\eta_{i1}, \cdots, \eta_{i6})^\top$, $\eta_{ij} = \beta_0 + \beta_1 X_{1i} + b_{0i} + b_{1i} W_{1ij}$, $\mathbf{b}_i = (b_{0i}, b_{1i})^\top$; X_1 is a dichotomous covariate (Bernoulli distribution with sucess probability 0.5) and W_1 is a continuous occasion-varying random covariate (standard normal distribution); β_0 is an intercept, *i.e.* the general mean of the linear predictor η_{ij} and β_1 is the fixed-effects associated to the covariate X_1 with values arbitrarily fixed to $\beta_0 = 1$ and $\beta_1 = -1$; b_{0i} is a random intercept associated to the cluster i, b_{1i} is the random slope associated to W_{1i}; $\overline{\mathbf{D}}$ is a 2×2 scale matrix with diagonal elements 0.5 and 1, and off diagonal element 0.25; $\mathcal{F}_U(\boldsymbol{\nu})$ is a positive distribution with finite first two negative moments, *i.e.* $\tilde{U}_t = \mathrm{E}\{U_i^{-t}\} < \infty$ for $t = 1, 2$; and $\upsilon_0 = [\tilde{U}_2 + (\tilde{U}_2 - c^2\tilde{U}_1^2)\boldsymbol{\delta}_\varepsilon^2]^{-1/2}$. We considered $\boldsymbol{\delta} = (0, \delta_1)^\top$, *i.e.* a null random intercept skewness to ensure the identifiability of the model.

Under this general class of SSMN latent models, we considered two data models. The first is the probit data model where $U = 1$, $\delta_1 = 0$ and $\delta_\varepsilon = 0$ (probit link, $\upsilon_0 = 1$). The second is the skew generalized t-link data model with $\delta_\varepsilon = -2$ and $\delta_1 = 2$ and $U \sim \mathcal{G}amma(\nu/2, \nu/2)$ with $\nu = 5$ ($\upsilon_0 = 0.4598$). We considered for each data model, sample sizes (n) of 100, 500 and 1000 and thus generated three sets of covariates which were used for all simulations involving each of the two data models. Under each of the six resulting simulation settings, we generated 250 datasets to which we fitted the four fitting models under evaluation (PM, SPM, GTLM, SGTLM), considering the model degrees of freedom as known and equals to $\nu = 5$ for the SGTLM. Fixed effects (β_0 and β_1) and skewness parameters (δ_ϵ and δ_1) were initialized to zero whereas the scale matrix $\overline{\mathbf{D}}$ was initialized to the 2×2 identity matrix.

Performance measures. In addition to estimates of fixed effects ($\widehat{\beta}_k$, $k = 0, 1$) and skewness ($\widehat{\boldsymbol{\delta}}_l$ for SPM and SGTLM, $l = \epsilon, 1$) and related empirical standard errors and $CI_{95\%}$, we recorded random effects variances (σ_1^2 of b_{0i} and σ_2^2 of b_{1i}) and covariance (σ_{12}) and their approximate standard errors derived using the delta method [54] as implemented in the R package *car* [55], empirical Bayes estimates of individual random effetcs ($\widehat{\mathbf{b}}_i$), and the *AIC*, *BIC* and *HQ* criteria defined as: $AIC = -2\widehat{\ell} + 2N_p$, $BIC = -2\widehat{\ell} + N_p \log N$, and $HQ = -2\widehat{\ell} + 2N_p \log(\log N)$

Table 1. Measures of the performance of binomial fitting models.

Measure	Formula	Role
$\%Bias(\widehat{\sigma}_l)$	$100 \times (\overline{\widehat{\sigma}}_{b_l} - \sigma_l)/\lvert\sigma_l\rvert$	Unbiasedness of $\widehat{\sigma}_l$
$RMSE(\widehat{\sigma}_l)$	$\sqrt{\sum_{r=1}^{250}(\widehat{\sigma}_l^{(r)} - \sigma_l)^2/250}$	Accuracy of $\widehat{\sigma}_l$
$\%Bias(\widehat{\beta}_k)$	$100 \times (\overline{\widehat{\beta}}_k - \beta_k)/\lvert\beta_k\rvert$	Unbiasedness of $\widehat{\beta}_k$
$RMSE(\widehat{\beta}_k)$	$\sqrt{\sum_{r=1}^{250}(\widehat{\beta}_k^{(r)} - \beta_k)^2/250}$	Accuracy of $\widehat{\beta}_k$
$SD(\widehat{\beta}_k)$	$\sqrt{\sum_{r=1}^{250}(\widehat{\beta}_k^{(r)} - \overline{\widehat{\beta}}_k))^2/249}$	Precision of $\widehat{\beta}_k$
$\overline{SE}(\widehat{\beta}_k)$	$\sqrt{\sum_{r=1}^{250} se(\widehat{\beta}_k^{(r)})^2/250}$	Reliability of estimated SE
$\overline{CP}(\widehat{\beta}_k)$	$\sum_{r=1}^{250} I_{CI^{(k,r)}}(\beta_k)/250$	Coverage probability of $\widehat{\beta}_k$
$\overline{CP}(\widehat{\boldsymbol{\delta}}_l)$	$\sum_{r=1}^{250} I_{CI^{(k,r)}}(\boldsymbol{\delta}_l)/250$	Coverage probability of $\widehat{\boldsymbol{\delta}}_l$
$\overline{R^2}(b_l, \widehat{b}_l)$	$\sum_{r=1}^{250} R^2(\widehat{b}_l^{(r)}, b_l^{(r)})/250$	Predictive power
$\overline{AIC}$	$\sum_{r=1}^{250} AIC^{(r)}/250$	Model selection
$\overline{BIC}$	$\sum_{r=1}^{250} BIC^{(r)}/250$	Model selection
$\overline{HQ}$	$\sum_{r=1}^{250} HQ^{(r)}/250$	Model selection

Notes: SE = standard error, $\overline{\widehat{\beta}}_k = \frac{\sum_{r=1}^{250}\widehat{\beta}_k^{(r)}}{250}$; $b_i^{(r)} = (b_{i0}^{(r)}, b_{i1}^{(r)})^\top$ is the r^{th} simulated vector of random effects for the subject i and $\widehat{b}_i^{(r)}$ is the empirical Bayes estimates for $b_i^{(r)}$; $R^2(\widehat{b}_l^{(r)}, b_l^{(r)})$ is the square of Pearson's coefficient of correlation between $\widehat{b}_l^{(r)}$ and $b_l^{(r)}$; $se(\widehat{\beta}_k^{(r)})$ is the empirical information-based standard error for $\widehat{\beta}_k^{(r)}$ and $CI^{(k,\,r)}$ is the 95% confidence interval for β_k from the r^{th} generated dataset. The same applies for δ_l and σ_l.

https://doi.org/10.1371/journal.pone.0249604.t001

where $\widehat{\ell}$ is the maximized log-likelihood value, N is the total number of observations and N_p is the number of estimated model parameters. These data were used to compute various performance measures (Table 1) including the relative bias (%*Bias*) and the root mean square error (*RMSE*) in $\widehat{\beta}_k$ and $\widehat{\sigma}_l$; the standard deviations (SD) of $\widehat{\beta}_k$; the quadratic mean ($\overline{SE}$) of standard errors of $\widehat{\beta}_k$; the coverage probabilities ($\overline{CP}$) of $\widehat{\beta}_k$ and $\widehat{\boldsymbol{\delta}}_l$, *i.e.* the proportion of times the $CI_{95\%}$ for β_k or δ_l included the true value; the arithmetic mean ($\overline{R^2}(b_l, \widehat{b}_l)$) of the square of Pearson's correlation (coefficient of determination) between simulated and Bayes estimates of subject random effects; and the arithmetic means of information criteria AIC ($\overline{AIC}$), BIC ($\overline{BIC}$) and HQ ($\overline{HQ}$) across the 250 simulated datasets per simulation setting.

Simulation results. Simulation results presented in Tables 2 and 3 show that under the probit data generation mechanism, the probit, the skew-probit, the generalized t (GT)-link and the skew generalized t (SGT)-link models recovered the population parameter values. Indeed, the percentage of bias was below 5% at all levels for fixed effects, whereas for variance components, the percentage of bias was below 20%. We particularly noticed a high relative bias in the variance component (%*Bias* = 17.33) under probit fit to probit data (assuming the true model) with small sample size (n = 100). This may be explained by the maximum likelihood estimation method which is known for providing biased variance components [56]. Nevertheless, this can be improved by opting for residual maximum likelihood estimation procedure [56]. However, it is worth noticing that the estimation improves as the sample size increases and the empirical standard error estimates agree with the standard deviations from the simulations. The results for n = 100 and n = 500 are consistent with findings in [24] where empirical information based standard errors approached Monte Carlo standard errors. Moreover, the 95% confidence interval for the skewness parameters allows to detect spurious

Table 2. Results based on 250 replications of probit samples: Probit and skew-probit fits.

Parameters	Measures	n = 100		n = 500		n = 1000	
		Probit	Skew probit	Probit	Skew probit	Probit	Skew probit
$\beta_0(-1)$	Mean	-1.04 (0.25)	-1.03 (0.30)	-1.01 (0.12)	-1.04 (0.31)	-1.00 (0.08)	-1.00 (0.11)
	%Bias	-3.61	-3.01	-0.64	-4.03	-0.32	-0.23
	$\overline{SE}$	0.25 [0.23]	0.27 [0.26]	0.11 [0.12]	0.29 [0.24]	0.08 [0.08]	0.09 [0.11]
	CP	0.96	0.97	0.95	0.95	0.97	0.97
$\beta_1(1)$	Mean	1.03 (0.24)	1.02 (0.41)	1.01 (0.10)	1.00 (0.41)	1.00 (0.07)	1.00 (0.42)
	%Bias	3.21	1.96	0.59	0.46	0.13	0.16
	$\overline{SE}$	0.23 [0.22]	0.43 [0.43]	0.1 [0.10]	0.48 [0.47]	0.07 [0.07]	0.08 [0.07]
	CP	0.95	0.96	0.97	0.97	0.97	0.97
$\sigma_1^2(0.5)$	Mean	0.54 (0.24)	0.51 (0.29)	0.51 (0.09)	0.53 (0.29)	0.50 (0.06)	0.51 (0.07)
	%Bias	7.98	2.24	1.39	2.35	0.12	1.06
$\sigma_{12}(0.25)$	Mean	0.30 (0.34)	0.23 (0.48)	0.25 (0.13)	0.22 (0.48)	0.25 (0.09)	0.24 (0.49)
	%Bias	18.48	-6.52	1.84	-12.02	1.55	-2.02
$\sigma_2^2(1)$	Mean	1.17 (0.91)	1.10 (0.95)	1.02 (0.34)	1.11 (0.94)	1.02 (0.22)	1.00 (0.47)
	%Bias	17.33	9.94	2.42	10.98	1.96	0.99
$\delta_\varepsilon(0)$	Mean	-	0.02 (0.65)	-	0.02 (0.46)	-	0.00 (0.46)
	$\overline{SE}$	-	0.61 [0.62]	-	0.71 [0.51]	-	0.61 [0.46]
	CP	-	1.00	-	1.00	-	1.00
$\delta_1(0)$	Mean	-	-0.02 (0.37)	-	-0.02 (0.38)	-	-0.00 (0.25)
	$\overline{SE}$	-	2.00 [1.84]	-	2.00 [1.72]	-	2.00 [0.22]
	CP	-	1.00	-	1.00	-	1.00
$\overline{R^2}(b_0, \widehat{b}_0)$	Mean	0.45	0.46	0.46	0.46	0.46	0.46
$\overline{R^2}(b_1, \widehat{b}_1)$	Mean	0.23	0.24	0.26	0.26	0.27	0.28
AIC	Mean	548.29	552.14	3732.11	3733.80	7522.50	7524.04
BIC	Mean	570.27	582.92	3762.14	3775.85	7556.00	7570.94
HQ	Mean	556.85	564.12	3742.91	3748.92	7534.13	7540.33

Mean value and root mean square error (in parentheses), percentage of bias (%Bias), mean standard error (SE) and sample standard deviation (in square brackets), coverage probability (CP) of estimates, coefficient of determination (R2) and averages of AIC, BIC and HQ criteria for sample sizes (n) of 100, 500 and 1000.
Notes: In the first column (Parameters), model parameters are followed in parentheses by their respective true values;—means "not applicable"; the coverage probability (CP) is the probability that an approximate confidence interval (assuming asymptotic normality for the model parameter) contains the true parameter value.

https://doi.org/10.1371/journal.pone.0249604.t002

skewness in the skew-probit and the SGT-link models with coverage probabilities of 100%. This result can be explained by the underlined high accuracy of information based standard errors in this type of model. The power in predicting random effects varied from $R^2 = 0.45$ to $R^2 = 0.47$ for random intercepts and from $R^2 = 0.23$ to $R^2 = 0.28$ for random slopes, but was comparable for the three fitting models. Finally, it appears that on average all model selection criteria correctly considered probit fitting model as the parsimonious model.

Under a SGT-link data generation mechanism, the probit model performed poorly, showing large relative fixed effects bias values which decreased from 46% for samples of size $n = 100$ to 18% for samples of size $n = 1000$ (Table 4). The SGT-link model estimates were the less biased ($\%Bias < 12$) as well as the most accurate with the lowest root mean square errors across all levels (Table 5). The same observations apply to variance components which were highly biased downward for probit, skew-probit and GT-link models ($\%Bias$ up to 90) relative to the SGT-link model ($\%Bias < 7$). Regarding estimates of skewness parameters, the coverage

Table 3. Results based on 250 replications of probit samples: Generalized t (GT)-link and skew generalized t (SGT)-link fits.

Parameters	Measures	n = 100		n = 500		n = 1000	
		GT-link	SGT-link	GT-link	SGT-link	GT-link	SGT-link
$\beta_0(-1)$	Mean	-1.05 (0.26)	-1.05 (0.24)	-1.01 (0.11)	-1.03 (0.11)	-.1.00 (0.10)	-1.00 (0.08)
	%Bias	-4.95	-4.61	-1.06	-2.94	-0.17	-0.17
	$\overline{SE}$	0.25 [0.24]	0.27 [0.25]	0.15 [0.12]	0.11 [0.12]	0.08 [0.08]	0.06 [0.05]
	CP	0.98	0.97	0.97	0.98	0.97	0.97
$\beta_1(1)$	Mean	1.03 (0.32)	1.03 (0.23)	1.01 (0.14)	1.03 (0.10)	1.00 (0.12)	1.00 (0.07)
	%Bias	3.16	3.28	1.62	2.90	0.08	0.13
	$\overline{SE}$	0.23 [0.20]	0.25 [0.22]	0.13 [0.13]	0.1 [0.10]	0.07 [0.07]	0.07 [0.07]
	CP	0.97	0.96	0.98	0.97	0.98	0.98
$\sigma_1^2(0.5)$	Mean	0.47 (0.12)	0.47 (0.10)	0.49 (0.11)	0.50 (0.09)	0.50 (0.05)	0.50 (0.03)
	%Bias	-5.36	-5.57	-1.76	0.01	0.17	0.12
$\sigma_{12}(0.25)$	Mean	0.26 (0.53)	0.24 (0.40)	0.24 (0.22)	0.23 (0.02)	0.25 (0.19)	0.25 (0.02)
	%Bias	4.00	-4.82	-3.96	-6.26	-1.68	1.55
$\sigma_2^2(1)$	Mean	1.09 (.62)	0.99 (0.49)	1.00 (0.18)	0.99 (0.12)	0.99 (0.11)	0.99 (0.12)
	%Bias	8.98	-1.12	-0.78	-1.13	-0.87	-0.91
$\delta_\varepsilon(0)$	Mean	-	0.00 (0.56)	-	0.01 (0.54)	-	0.00 (0.41)
	$\overline{SE}$	-	0.44 [0.56]	-	0.46 [0.56]	-	0.48 [0.44]
	CP	-	1.00	-	1.00	-	1.00
$\delta_1(0)$	Mean	-	0.00 (0.18)	-	0.00 [0.14]	-	0.00 [0.14]
	$\overline{SE}$	-	1.22 [0.18]	-	0.99 [0.16]	-	0.62 [0.17]
	CP	-	1.00	-	1.00	-	1.00
$\overline{R^2}(b_0, \widehat{b}_0)$	Mean	0.45	0.46	0.47	0.46	0.47	0.47
$\overline{R^2}(b_1, \widehat{b}_1)$	Mean	0.26	0.27	0.27	0.27	0.27	0.28
AIC	Mean	552.09	554.37	3733.10	3736.21	7523.12	7525.21
BIC	Mean	581.43	589.55	3762.43	3784.26	7568.95	7573.26
HQ	Mean	561.97	568.06	3744.11	3764.90	7538.91	7542.49

Mean value and root mean square error (in parentheses), percentage of bias (%Bias), mean standard error (SE) and sample standard deviation (in square brackets), coverage probability (CP) of estimates, coefficient of determination (R2) and averages of AIC, BIC and HQ criteria for sample sizes (n) of 100, 500 and 1000.
Notes: In the first column (Parameters), model parameters are followed in parentheses by their respective true values;—means "not applicable"; the coverage probability (CP) is the probability that an approximate confidence interval (assuming asymptotic normality for the model parameter) contains the true parameter value.

https://doi.org/10.1371/journal.pone.0249604.t003

probability was low (53% to 94%) for small sample size (n = 100) and approached nominal (95%) value for larger sample sizes (n = 500, 1000). The skew-probit model estimates (coverage probability down to 53%) were less reliable than estimates from the SGT-link model (coverage probability above 90%).

Clearly, the SGT-link model adjusted better with non normal data and accordingly, random effects prediction is better with SGT-link model ($R^2 \geq 0.49$) than with the probit or the skew-probit model. Moreover, all the considered model selection criteria namely AIC, BIC and HQ on average correctly selected the SGT-link model as the preferred model.

Application to the respiratory infection data

To demonstrate the usefulness of the proposed approach to correlated binary data modeling, we revisited the respiratory illness data (available in *geepack* package [57] in R) which was used by [24] to illustrate their t-link GLMM. The respiratory illness data was obtained from a

Table 4. Results based on 250 replications of skew generalized t-link samples (probit and skew-probit fits).

Parameters	Measures	$n = 100$		$n = 500$		$n = 1000$	
		Probit	Skew probit	Probit	Skew probit	Probit	Skew probit
$\beta_0(-1)$	Mean	-1.46 (1.53)	-1.33 (1.30)	-1.28 (1.29)	-1.28 (1.31)	-1.18 (1.27)	-1.22 (0.92)
	%Bias	-46.18	-33.41	-28.30	-28.13	-18.27	-22.08
	$\overline{SE}$	0.95 [0.99]	0.89 [0.86]	0.96 [1.02]	0.89 [0.94]	0.97 [0.99]	0.92 [0.89]
	CP	0.93	0.95	0.95	0.96	0.96	0.96
$\beta_1(1)$	Mean	0.95 (0.46)	0.97 (0.47)	0.95 (0.41)	0.98 (0.42)	0.95 (0.31)	0.98 (0.37)
	%Bias	-5.11	-3.31	-5.18	-2.11	-5.21	-2.11
	$\overline{SE}$	0.51 [0.48]	0.44 [0.48]	0.26 [0.39]	0.41 [0.43]	0.31 [0.29]	0.40 [0.41]
	CP	0.95	0.96	0.97	0.97	0.97	0.97
$\sigma_1^2(0.83)$	Mean	0.34 (0.94)	0.61 (0.39)	0.41 (0.91)	0.67 (0.38)	0.56 (0.76)	0.73 (0.34)
	%Bias	-59.04	-26.51	-50.60	-19.28	-32.53	-12.05
$\sigma_{12}(0.42)$	Mean	0.80 (0.64)	0.53 (0.58)	0.81 (0.53)	0.42 (0.38)	0.75 (0.50)	0.40 (0.39)
	%Bias	90.48	26.19	92.86	0.09	78.57	-4.76
$\sigma_2^2(4.73)$	Mean	2.07 (1.19)	2.85 (1.08)	2.22 (1.44)	2.97 (1.99)	2.20 (1.32)	2.88 (1.70)
	%Bias	-56.24	-39.75	-53.07	-37.21	-53.49	-39.11
$\delta_\varepsilon(-2)$	Mean	-	-0.48 (1.75)	-	-0.91 (1.25)	-	-0.90 (1.04)
	%Bias	-	76.01	-	54.50	-	54.99
	$\overline{SE}$	-	0.60 [0.72]	-	1.23 [1.22]	-	1.03 [1.04]
	CP	-	0.87	-	0.91	-	0.95
$\delta_1(2)$	Mean	-	1.03 (0.97)	-	1.13 (0.97)	-	1.16 (0.92)
	%Bias	-	-48.51	-	-43.49	-	-42.00
	$\overline{SE}$	-	1.01 [1.00]	-	1.04 [1.01]	-	0.99 [0.96]
	CP	-	0.53	-	0.61	-	0.66
$\overline{R^2}(b_0, \widehat{b}_0)$	Mean	0.35	0.38	0.38	0.42	0.38	0.46
$\overline{R^2}(b_1, \widehat{b}_1)$	Mean	0.14	0.26	0.16	0.31	0.26	0.33
AIC	Mean	578.22	576.34	3824.31	3814.72	7662.07	7713.55
BIC	Mean	600.20	607.12	3854.34	3856.77	7695.56	7681.25
HQ	Mean	586.78	588.32	3835.11	3829.84	7673.70	7678.562

Mean value and root mean square error (in parentheses), percentage of bias (%Bias), mean standard error (SE) and sample standard deviation (in square brackets), coverage probability (CP) of estimates, coefficient of determination (R2) and averages of AIC, BIC and HQ criteria for sample sizes (n) of 100, 500 and 1000.
Notes: In the first column (Parameters), model parameters are followed by their respective true values in parentheses;—means "not applicable"; the coverage probability (CP) is the probability that an approximate confidence interval (assuming asymptotic normality for the model parameter) contains the true parameter value.

https://doi.org/10.1371/journal.pone.0249604.t004

clinical study of the effect of a treatment on 111 patients with respiratory illness, recruited from two different clinical centers. The patients were examined and their respiratory state (categorized as 1 = good, 0 = poor) determined (baseline). They were then randomized to receive either placebo or an active treatment. The goal of the study was to determine whether the treatment induced a better respiratory state in treated patients. The outcome is the respiratory state measured at four visits for each patient as good ($y = 1$) or poor ($y = 0$). In addition to the treatment (treat = 0 for placebo group (P) and treat = 1 for treated group (A)), the following fixed covariates were included: the clinical center (center = 0 for the first center and center = 1 for the second center), the baseline (respiratory state at the first visit), gender (sex = 0 for female (F) and sex = 1 for male (M)) and the interaction of treatment and gender. Following [24], we assumed that the age effect is patient-specific (random slope) and thus considered the patient

Table 5. Results based on 250 replications of skew generalized t-link samples (generalized t-link and skew generalized t-link fits).

Parameters	Measures	$n = 100$		$n = 500$		$n = 1000$	
		GT-link	SGT-link	GT-link	SGT-link	GT-link	SGT-link
$\beta_0(-1)$	Mean	-1.40 (1.54)	-1.17 (1.31)	-1.25 (1.22)	-1.11 (1.09)	-1.11	-1.02 (0.09)
	%Bias	-39.17	-17.33	-24.56	-11.4	11.41	-2.09
	$\overline{SE}$	0.96 [0.89]	0.64 [0.55]	0.83 [0.92]	0.60 [0.51]	0.89 [0.86]	0.40[0.37]
	CP	0.93	0.96	0.95	0.98	0.96	0.98
$\beta_1(1)$	Mean	0.98 (0.37)	1.01 (0.33)	0.99 (0.36)	1.01 (0.16)	0.98 (0.22)	1.00 (0.09)
	%Bias	-2.19	1.10	-1.14	1.23	-1.98	0.37
	$\overline{SE}$	0.32 [0.29]	0.28 [0.32]	0.23 [0.28]	0.11 [0.16]	0.28 [0.21]	0.09 [0.10]
	CP	0.95	0.96	0.97	0.97	0.97	0.98
$\sigma_1^2(0.83)$	Mean	0.37 (0.55)	0.86 (0.16)	0.40 (0.56)	0.84 (0.09)	0.53 (0.63)	0.84 (0.07)
	%Bias	-55.45	3.61	-51.80	1.20	-36.14	1.20
$\sigma_{12}(0.42)$	Mean	0.68 (0.61)	0.36 (0.31)	0.71 (0.56)	0.38 (0.23)	0.66 (0.56)	0.38 (0.09)
	%Bias	61.91	-14.29	69.05	-9.52	57.14	-9.52
$\sigma_2^2(4.73)$	Mean	2.29 (1.23)	4.56 (1.51)	2.33 (1.17)	4.55 (1.29)	2.34 (1.11)	4.58 (1.31)
	%Bias	51.59	-3.59	-50.74	-3.81	-50.53	-3.17
$\delta_\varepsilon(-2)$	Mean	-	-2.18 (1.06)	-	-2.12 (1.04)	-	-2.06 (0.99)
	%Bias	-	-9.00	-	-6.02	-	-3.04
	$\overline{SE}$	-	1.44 [1.04]	-	1.04 [1.05]	-	0.88 [0.98]
	CP	-	0.94	-	0.96	-	0.97
$\delta_1(2)$	Mean	-	1.64 (1.18)	-	1.69 [1.16]	-	1.84 [1.04]
	%Bias	-	-18.13	-	-15.50	-	-8.10
	$\overline{SE}$	-	1.13 [1.18]	-	1.16 [1.17]	-	1.02 [1.03]
	CP	-	.90	-	0.94	-	0.96
$\overline{R^2}(b_0, \widehat{b}_0)$	Mean	0.34	0.53	0.38	0.56	0.40	0.58
$\overline{R^2}(b_1, \widehat{b}_1)$	Mean	0.22	0.49	0.26	0.52	0.26	0.52
AIC	Mean	578.04	524.06	3820.10	3804.12	7657.13	7652.56
BIC	Mean	601.18	597.24	3853.79	3852.17	7689.12	7671.17
HQ	Mean	588.01	575.75	3834.30	3821.40	7667.66	7542.49

Mean value and root mean square error (in parentheses), percentage of bias (%Bias), mean standard error (SE) and sample standard deviation (in square brackets), coverage probability (CP) of estimates, coefficient of determination (R2) and averages of AIC, BIC and HQ criteria for sample sizes (n) of 100, 500 and 1000.
Notes: In the first column (Parameters), model parameters are followed by their respective true values in parentheses;—means "not applicable"; the coverage probability (CP) is the probability that an approximate confidence interval (assuming asymptotic normality for the model parameter) contains the true parameter value.

https://doi.org/10.1371/journal.pone.0249604.t005

age centered around its median (31 years) as a random covariate. Since the fixed covariates included binary variables (treat, gender, center and baseline), a conditional skew-probit model is not identifiable given random effects and we thus set $\delta_\epsilon = 0$ to ensure identifiability.

For the purpose of comparison, we fitted the probit, skew-probit, GT-link and SGT-link models. We initialized fixed effects $\boldsymbol{\beta}$ and the random slope skewness parameter δ to zero whereas the random slope scale was initialized to one. For the GT-link and the SGT-link models, we considered the model selection approach of [51] with degrees of freedom $\nu = 2.5, 2.6, \ldots, 15$.

As depicted in Fig 1, the profiled marginal log-likelihood for the GT-link model is unbounded, with smaller ν corresponding to better fit in accordance with the t-link model fits in [24] ($\nu \leq 4$). We thus set $\nu = 2.5$ for the t-link model. For the SGT-link fit, Fig 1 indicates

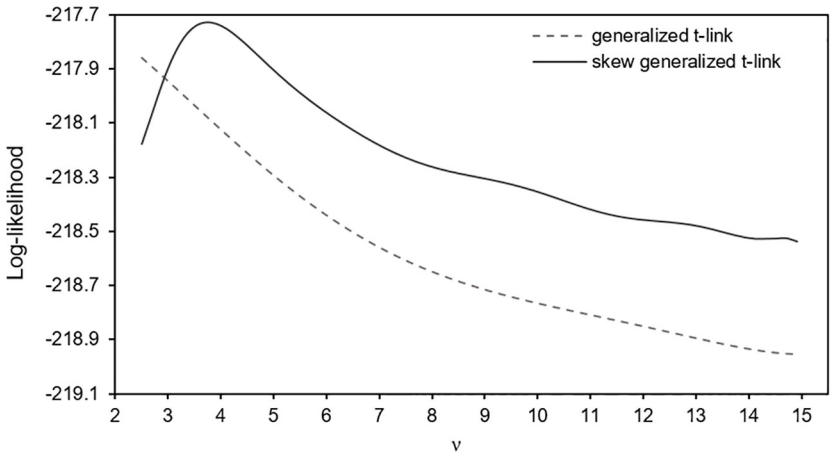

Fig 1. Fitting the generalized t-link and the skew generalized t models to the respiratory infection data: Plot of the marginal log-likelihood profiled for the degrees of freedom ν.

https://doi.org/10.1371/journal.pone.0249604.g001

that the log-likelihood is bounded with a maximum at $\nu = 3.7$, suggesting heavy tail link function and random slope distributions. The difference in the behaviours of the GT-link and SGT-link models may be explained by the implication of ν in the location of the skew t-link model through $\tilde{U}_1$ (see Eq (32)).

The maximum likelihood (ML) estimates under the probit, skew-probit, GT-link and SGT-link models (Table 6) are somewhat close for the four fitted models which all show that respiratory illness is associated to clinical center, baseline state and treatment, with the treatment effect varying with gender. The SGT-link fit additionally indicates that, irrespective of the treatment, the respiratory state is poorer for male patients (Table 6, $\widehat{\beta}_3 < 0$) than females. We notice for this dataset, that the intercept coefficient estimate increases with model complexity and estimates of fixed effects and their respective standard errors are shrunk toward zero for the skew-probit model relative to the probit one, and for the skew-probit model relative to the

Table 6. Maximum likelihood fits of probit, skew-probit, Generalized T (GT)-link and Skew Generalized T (SGT) -link models to the respiratory infection data.

Parameter (variable)	Probit			Skew probit			GT-link ($\nu = 2.5$)			SGT-link ($\nu = 3.7$)		
	estimate	se	z value	estimate	se	z value	estimate	se	z value	estimate	se	z value
β_0 (Intercept)	0.3429	0.4994	0.6866	0.4527	0.5012	0.9031	0.3570	0.5491	0.6502	0.5383	0.4659	1.1555
β_1 (center = 2)	0.6292	0.2639	2.3846	0.6890	0.2611	2.6391	0.6767	0.2101	2.9410	0.5993	0.2395	2.5026
β_2 (baseline)	1.6689	0.2900	5.7544	1.6312	0.2837	5.7499	1.8100	0.4118	4.3952	1.5152	0.2584	5.8627
β_3 (sex = M)	-0.8228	0.5218	-1.5768	-1.0170	0.5293	-1.9216	-0.8833	0.5274	-1.6748	-1.0129	0.4944	-2.0489
β_4 (treat = P)	-2.1198	0.6018	-3.5224	-2.0980	0.5958	-3.5215	-2.2634	0.7910	-2.8615	-2.0398	0.5366	-3.8010
β_5 (M×P)	1.3840	0.6461	2.1419	1.4055	0.6403	2.1950	1.3251	0.6300	2.1033	1.3425	0.5867	2.2884
δ (age)	-	-	-	0.0411	0.0216	1.9011	-	-	-	0.0262	0.0139	1.8906
σ^2 (age)	0.0117	0.0050	-	0.0113	0.0049	-	0.0143	0.0061	-	0.0118	0.0059	-
AIC	455.6305	-	-	453.7766	-	-	453.2140	-	-	451.4658	-	-
BIC	484.3013	-	-	486.5432	-	-	485.9399	-	-	484.2324	-	-
HQ	466.9370	-	-	466.6983	-	-	465.7514	-	-	464.3875	-	-

Notes: M = male patient; P = placebo; M×P is a shortcut for sex = M×treat = P; σ^2 and δ are the variance and the skewness parameter respectively for the patient-specific slope of median centered age (year); se = standard error; z value = estimate/se (a z value $\geq$1.96 roughly indicates 5% significance assuming asymptotic normality of z values);—means "not applicable"

https://doi.org/10.1371/journal.pone.0249604.t006

SGT-link one. The skew-probit model fit also gave a higher skewness ($\widehat{\boldsymbol{\delta}} = 0.0411$) as compared with the SGT-link model fit ($\widehat{\boldsymbol{\delta}} = 0.0262$). Although the estimated skewness is relatively low for both skew-probit and SGT-link models, the use of a skewed and heavy tail link clearly improved, not only the precision of estimates but also the adequacy between data and model. Indeed, the asymptotic 95% confidence interval for δ under the SGT-link includes zero ($CI_{95\%}$ = [−0.0010, 0.0534]), but we noticed from the simulation results that asymptotic $CI_{95\%}$ for skewness parameters becomes reliable only in large samples ($n \geq 500$), whereas information criteria are reliable for all tested sample sizes. Thus, based on the AIC, BIC and HQ criteria in Table 6, the SGT-link fit is the best for the respiratory illness data. The estimate of the variance of the random slope of age is $\widehat{\sigma}^2 = 0.0118$ for the SGT-link fit, with close values under probit and skew-probit models. From the SGT-link fit, it appears that the treatment induced an overall better respiratory state for treated patients (with a negative coefficient $\beta_4 = -2.0398$ for the placebo group). Moreover, the treatment has on average a better effect on female patients than on male patients (with a positive coefficient, $\beta_5 = 1.3425$ for male patients in the placebo group). However, as noted by [24], new studies are required to check this latter trend because of the highly unbalanced proportion of males (79%) and females (21%) in the data.

Conclusion

This work has considered the skew generalized t class of distributions for both link and random effects distributions in mixed models for binary data. The objective was to improve the exploitation of binary data bearing oddities such as skewness and tails thicker/thinner than the normal distribution. To allow inference in such models, we developped a maximum likelihood estimation procedure based on the EM algorithm. We combined results from [34] and [37] to obtained expressions for computing moments of truncated multivariate skew t distributions. The computation used existing R functions for the multivariate skew t cumulative distribution function. Our simulation experiment showed that, irrespective of sample size, the SGT-link model outperforms the probit GLMM when the underlying data generation mechanism is not normal. We also demonstrated that the skew generalized-link model performed better than the skew-probit and the generalized t-link GLMMs, when the underlying data is both skewed and heavy tailed.

An important finding is that when the model degrees of freedom ν is small and very large values are assumed (fitting probit and skew-probit models), the estimates of fixed effects are biased, whereas when ν is large but small values are assumed, the estimates of fixed effects are not biased. Moreover, asymptotic inference using information based standard errors proved highest ability accuracy in detecting spurious skewness in large samples ($n \geq 500$) and information criteria on average selected the correct model fit for all tested sample sizes (n = 100, 500, 1000). These findings extend results in [24] on t-link GLMM to SGT-link GLMM, asserting that information criteria are reliable for selecting the best model for a particular dataset.

However, the simulation experiments revealed that the EM algorithm has a high computational cost. For instance, in a model with q = 2 random effects, n = 100 clusters and n_i = 6 observations per cluster, the mean running time for the SGT-link model fit was 4.76 minutes which is almost 135 times the time required by the probit model fit (2.12 seconds). Our implementation relies on the *pmst* function of the R package *sn* [35] to compute the cumulative probabilities of skew t distributions. This function uses the one dimensional routine *integral* of R on the multivariate normal cumulative distribution function. The use of the EM algorithm for large q values (*e.g.* q = 10, 15) requires the prior development of a faster routine for the computation of cumulative probabilities of skew t distributions. This will make the EM algorithm scalable for large $q + n_i$. On multicore plateformes, parallel computing can

also substantially speed computations up. The expressions provided for computing moments of truncated multivariate skew t distributions is limited to work for models with $\nu > 2$. The use of formulae given in [38] will extend our EM algorithm to very small degrees of freedom ($1 < \nu \leq 2$).

Binary data related to very rare events often require special treatment and are generally analysed using zero inflated models [58]. The development of a skew generalized t-link model with zero inflation can significantly improve the exploitation of such data. In addition to binary data, GLMMs handle other data types like count, proportional and ordinal outcomes. From the good performance demonstrated in this work and in previous related ones [9, 24], we believe that the simultaneous introduction of flexible links and random effects distributions in GLMM would benefit knowledge extraction from observed data in applied research fields where advances rely on modeling capacity.

Supporting information

S1 Appendix. Proof of Lemma 1. This supporting information gives a proof of *Lemma 1*.
(PDF)

S2 Appendix. Proof of Lemma 2. This supporting information gives a proof of *Lemma 2*.
(PDF)

S3 Appendix. Proof of Proposition 1. This supporting information gives a proof of *Proposition 1*.
(PDF)

S4 Appendix. Proof of Corollary 1. This supporting information gives a proof of *Corollary 1*.
(PDF)

S5 Appendix. S5 Proof and limiting case of *Proposition 2*. This supporting information gives a proof of *Proposition 2*. The first two moments of truncated multivariate skew normal distributions (limiting case as $\nu \to \infty$) are also given (required for fitting skew-probit link models).
(PDF)

S6 Appendix. Proof of Proposition 3. This supporting information gives a proof of *Proposition 3*.
(PDF)

S7 Appendix. Proof of Proposition 4. This supporting information gives a proof of *Proposition 4*.
(PDF)

S8 Appendix. Proof of Proposition 5. This supporting information gives a proof of *Proposition 5*.
(PDF)

Acknowledgments

The authors wish to thank the editor and two referees for their relevant comments and suggestions. They are also grateful to Matthews Lazaro (Kamuzu College of Nursing, Lilongwe, Malawi) for the time he devoted to edit the manuscript for language usage, spelling, and grammar.

Author Contributions

Conceptualization: Chénangnon Frédéric Tovissodé, Romain Glèlè Kakaï.

Methodology: Chénangnon Frédéric Tovissodé.

Supervision: Aliou Diop, Romain Glèlè Kakaï.

Writing – original draft: Chénangnon Frédéric Tovissodé.

Writing – review & editing: Chénangnon Frédéric Tovissodé, Aliou Diop, Romain Glèlè Kakaï.

References

1. El-Saeiti IN. Performance of mixed effects for clustered binary data models. In: AIP Conference Proceedings. vol. 1643. AIP; 2015. p. 80–85.
2. Nelder JA, Wedderburn RW. Generalized linear models. Journal of the Royal Statistical Society: Series A (General). 1972; 135(3):370–384. https://doi.org/10.2307/2344614
3. McCulloch CE. Maximum likelihood variance components estimation for binary data. Journal of the American Statistical Association. 1994; 89(425):330–335. https://doi.org/10.1080/01621459.1994.10476474
4. Chen MH. Skewed link models for categorical response data. In: Skew-Elliptical Distributions and Their Applications. Chapman and Hall/CRC; 2004. p. 151–172.
5. McCulloch CE, Neuhaus JM. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. Statistical science. 2011; 28(3):388–402.
6. Czado C, Santner TJ. The effect of link misspecification on binary regression inference. Journal of statistical planning and inference. 1992; 33(2):213–231. https://doi.org/10.1016/0378-3758(92)90069-5
7. Stewart MB. Semi-nonparametric estimation of extended ordered probit models. Stata Journal. 2004; 4 (1):27–39. https://doi.org/10.1177/1536867X0100400102
8. Liu C. Robit regression: a simple robust alternative to logistic and probit regression. In: Gelman A, Meng XL, editors. Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives. England: Wiley London; 2004. p. 227–238.
9. Kim S, Chen MH, Dey DK. Flexible generalized t-link models for binary response data. Biometrika. 2008; 95(1):93–106. https://doi.org/10.1093/biomet/asm079
10. Abanto-Valle CA, Dey DK. State space mixed models for binary responses with scale mixture of normal distributions links. Computational Statistics & Data Analysis. 2014; 71:274–287. https://doi.org/10.1016/j.csda.2013.01.009
11. Basu S, Mukhopadhyay S. Binary response regression with normal scale mixture links. BIOSTATISTICS-BASEL-. 2000; 5:231–242.
12. Pinheiro JC, Liu C, Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. Journal of Computational and Graphical Statistics. 2001; 10(2):249–276. https://doi.org/10.1198/10618600152628059
13. Chen MH, Dey DK, Shao QM. A new skewed link model for dichotomous quantal response data. Journal of the American Statistical Association. 1999; 94(448):1172–1186. https://doi.org/10.1080/01621459.1999.10473872
14. Komori O, Eguchi S, Ikeda S, Okamura H, Ichinokawa M, Nakayama S. An asymmetric logistic regression model for ecological data. Methods in Ecology and Evolution. 2016; 7(2):249–260. https://doi.org/10.1111/2041-210X.12473
15. Lemonte AJ, Bazán JL. New links for binary regression: an application to coca cultivation in Peru. Test. 2018; 27(3):597–617. https://doi.org/10.1007/s11749-017-0563-1
16. Asgharzadeh A, Esmaeili L, Nadarajah S, Shih S. A generalized skew logistic distribution. REVSTAT–Statistical Journal. 2013; 11(3):317–338.
17. Carlin JB, Wolfe R, Brown CH, Gelman A. A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. Biostatistics. 2001; 2(4):397–416. https://doi.org/10.1093/biostatistics/2.4.397 PMID: 12933632
18. Agresti A, Caffo B, Ohman-Strickland P. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. Computational Statistics & Data Analysis. 2004; 47 (3):639–653. https://doi.org/10.1016/j.csda.2003.12.009

19. Chen J, Zhang D, Davidian M. A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. Biostatistics. 2002; 3(3):347–360. https://doi.org/10.1093/biostatistics/3.3.347 PMID: 12933602

20. Nelson KP, Lipsitz SR, Fitzmaurice GM, Ibrahim J, Parzen M, Strawderman R. Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. Journal of Computational and Graphical Statistics. 2006; 15(1):39–57. https://doi.org/10.1198/106186006X96854

21. Hosseini F, Eidsvik J, Mohammadzadeh M. Approximate Bayesian inference in spatial GLMM with skew normal latent variables. Computational Statistics & Data Analysis. 2011; 55(4):1791–1806. https://doi.org/10.1016/j.csda.2010.11.011

22. Broström G, Holmberg H. Generalized linear models with clustered data: Fixed and random effects models. Computational Statistics & Data Analysis. 2011; 55(12):3123–3134. https://doi.org/10.1016/j.csda.2011.06.011

23. Gad AM, El Kholy RB. Generalized linear mixed models for longitudinal data. International Journal of Probability and Statistics. 2012; 1(3):41–47. https://doi.org/10.5923/j.ijps.20120103.03

24. Prates MO, Costa DR, Lachos VH. Generalized linear mixed models for correlated binary data with t-link. Statistics and Computing. 2014; 24(6):1111–1123. https://doi.org/10.1007/s11222-013-9423-3

25. Santos CC, Loschi RH. EM-Type algorithms for heavy-tailed logistic mixed models. Journal of Statistical Computation and Simulation. 2017; 87(15):2940–2961. https://doi.org/10.1080/00949655.2017.1350678

26. Azzalini A, Capitanio A. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003; 65(2):367–389. https://doi.org/10.1111/1467-9868.00391

27. Liu C, Rubin DB, Wu YN. Parameter expansion to accelerate EM: the PX-EM algorithm. Biometrika. 1998; 85(4):755–770. https://doi.org/10.1093/biomet/85.4.755

28. Branco MD, Dey DK. A general class of multivariate skew-elliptical distributions. Journal of Multivariate Analysis. 2001; 79(1):99–113. https://doi.org/10.1006/jmva.2000.1960

29. Hugo LDV, Cabral CRB. Scale Mixtures of Skew-Normal Distributions. In: Hugo LDV, Cabral CRB, Zeller CB, editors. Finite Mixture of Skewed Distributions. Switzerland: Springer International Publishing; 2018. p. 15–36.

30. Lachos VH, Ghosh P, Arellano-Valle RB. Likelihood based inference for skew-normal independent linear mixed models. Statistica Sinica. 2010; 20:303–322.

31. Capitanio A. On the canonical form of scale mixtures of skew-normal distributions; 2012. Available from: https://arxiv.org/abs/1207.0797.

32. Kéri G. The Sherman-Morrison formula for the determinant and its application for optimizing quadratic functions on condition sets given by extreme generators. In: Giannessi F, Pardalos P, T R, editors. Optimization Theory. Boston: Springer; 2001. p. 119–138.

33. Ahmed A, Reshi J, Mir K. Structural properties of size biased Gamma distribution. IOSR J Mathem. 2013; 5:55–61. https://doi.org/10.9790/5728-0525561

34. Ho HJ, Lin TI, Chen HY, Wang WL. Some results on the truncated multivariate t distribution. Journal of Statistical Planning and Inference. 2012; 142(1):25–40. https://doi.org/10.1016/j.jspi.2011.06.006

35. Azzalini A. The R package `sn`: The Skew-Normal and Related Distributions such as the Skew-*t* (version 1.5-2).; 2018. Available from: http://azzalini.stat.unipd.it/SN.

36. R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: https://www.R-project.org/.

37. Galarza CE, Matos LA, Lachos VH. Moments of the doubly truncated selection elliptical distributions with emphasis on the unified multivariate skew-*t* distribution. arXiv preprint arXiv:200714980. 2020.

38. Galarza CE, Lin TI, Wang WL, Lachos VH. On moments of folded and truncated multivariate Student-t distributions based on recurrence relations. Metrika. 2021; p. 1–26.

39. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological). 1977; 39(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

40. Fernandez C, Steel MF. Multivariate Student-t regression models: Pitfalls and inference. Biometrika. 1999; 86(1):153–167. https://doi.org/10.1093/biomet/86.1.153

41. da Silva Braga A, Cordeiro GM, Ortega EM, Silva GO. The Odd Log-Logistic Student t Distribution: Theory and Applications. Journal of Agricultural, Biological and Environmental Statistics. 2017; 22(4):615–639. https://doi.org/10.1007/s13253-017-0301-x

42. Lee D, Sinha S. Identifiability and bias reduction in the skew-probit model for a binary response. Journal of Statistical Computation and Simulation. 2019; 89(9):1621–1648. https://doi.org/10.1080/00949655.2019.1590579

43. Arellano-Valle RB, Genton MG. Fundamental skew distributions. Journal of Multivariate Analysis. 2005; 96:93–116. https://doi.org/10.1016/j.jmva.2004.10.002

44. Arellano-Valle R, Bolfarine H, Lachos V. Bayesian inference for skew-normal linear mixed models. Journal of Applied Statistics. 2007; 34(6):663–682. https://doi.org/10.1080/02664760701236905

45. Arellano-Valle R, Bolfarine H, Lachos V. Skew-normal linear mixed models. Journal of data science. 2005; 3(4):415–438.

46. Lin TI, Lee JC. Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. Statistics in medicine. 2008; 27(9):1490–1507. https://doi.org/10.1002/sim.3026 PMID: 17708515

47. Lachos VH, Dey DK, Cancho VG. Robust linear mixed models with skew-normal independent distributions from a Bayesian perspective. Journal of Statistical Planning and Inference. 2009; 139(12):4098–4110. https://doi.org/10.1016/j.jspi.2009.05.040

48. Lachos VH, Labra FV, Ghosh P. Multivariate skew-normal/independent distributions: properties and inference. Pro Mathematica. 2014; 28(56):11–53.

49. Pereira MAA, Russo CM. Nonlinear mixed-effects models with scale mixture of skew-normal distributions. Journal of Applied Statistics. 2019; 46(9):1602–1620. https://doi.org/10.1080/02664763.2018.1557122

50. Liu C, Rubin DB. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika. 1994; 81(4):633–648. https://doi.org/10.1093/biomet/81.4.633

51. Lange KL, Little RJ, Taylor JM. Robust statistical modeling using the t distribution. Journal of the American Statistical Association. 1989; 84(408):881–896. https://doi.org/10.2307/2290063

52. Meilijson I. A fast improvement to the EM algorithm on its own terms. Journal of the Royal Statistical Society Series B (Methodological). 1989; 51(1):127–138. https://doi.org/10.1111/j.2517-6161.1989.tb01754.x

53. Meza C, Osorio F, De la Cruz R. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. Statistics and Computing. 2012; 22(1):121–139. https://doi.org/10.1007/s11222-010-9212-1

54. Cox C. Delta method. Encyclopedia of biostatistics. 2005; 2. https://doi.org/10.1002/0470011815.b2a15029

55. Fox J, Weisberg S. An R Companion to Applied Regression. 3rd ed. Thousand Oaks CA: Sage; 2019. Available from: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

56. Meza C, Jaffrézic F, Foulley JL. Estimation in the probit normal model for binary outcomes using the SAEM algorithm. Computational Statistics & Data Analysis. 2009; 53(4):1350–1360. https://doi.org/10.1016/j.csda.2008.11.024

57. Yan J. geepack: Yet Another Package for Generalized Estimating Equations. R-News. 2002; 2/3:12–14.

58. Hall DB. Zero-Inflated Poisson and Binomial Regression with random effects: A Case Study. Biometrics. 2000; 56(4):1030–1039. https://doi.org/10.1111/j.0006-341X.2000.01030.x PMID: 11129458