Structural Equation Models And Mixture Models With Continuous Non-Normal Skewed Distributions

Tihomir Asparouhov and Bengt Muthén

Mplus Web Notes: No. 19
Version 2

July 3, 2014

Abstract

In this paper we describe a structural equation modeling framework that allows non-normal skewed distributions for the continuous observed and latent variables. This framework is based on the multivariate restricted skew t-distribution. We demonstrate the advantages of skewed structural equation modeling over standard SEM modeling and challenge the notion that structural equation models should be based only on sample means and covariances. The skewed continuous distributions are also very useful in finite mixture modeling as they prevent the formation of spurious classes formed purely to compensate for deviations in the distributions from the standard bell curve distribution. This framework is implemented in Mplus Version 7.2.

1 Introduction

Standard structural equation models reduce data modeling down to fitting means and covariances. All other information contained in the data is ignored. In this paper, we expand the standard structural equation model framework to take into account the skewness and kurtosis of the data in addition to the means and the covariances. This new framework looks deeper into the data to yield a more informative structural equation model.

There is a preconceived notion that standard structural equation models are sufficient as long as the standard errors of the parameter estimates are adjusted for failure to meet the normality assumption, but this is not correct. Even with robust estimation, the data are reduced to means and covariances. Only the standard errors of the parameter estimates extract additional information from the data. The parameter estimates themselves remain the same, i.e., the structural equation model is still concerned with fitting only the means and the covariances and ignoring higher-order information.

In this paper, we explore structural equation modeling based on the more flexible parametric family of distributions called the skew t-distribution. We will call these models skewed structural equation models (skew-SEM) as compared to the standard structural equation models which we will refer to simply as SEM. Fitting the skew t-distribution to the data allows us to extract more information from the data, namely, not just the means and the covariances but also to some extent the skewness and the kurtosis. Modeling these higher level moments is more intricate than modeling the means and the covariances. For example, modeling the skewness of the data is necessarily entangled with modeling the covariance.

In addition, fitting the skew t-distribution is not the same as fitting the skewness and kurtosis. The skewness and kurtosis are also limited characteristics of the data. By fitting the data to a flexible parametric family of distributions, we fit the means, the covariances, the skewness, the kurtosis, as well as the entire distribution.

All of the models described in this article are linear models. Unique properties of the skew t-distribution allow us to write structural equation models the same way they are written when the variables have Gaussian distributions. All observed variables, latent variables and residual variables in the structural equation models are allowed to have skew t-distributions.

Despite the fact that all models are linear, in certain skew-SEM settings some conditional expectations might not be linear. Thus skew-SEM can also be viewed as non-linear models despite the fact that we only specify linear models. The advantage of this approach to non-linear models is that we don't need to specify a particular non-linear model such as quadratic, logarithmic or exponential. The non-linearity of the skew-SEM model is determined by the skewness and the kurtosis of the data.

In standard structural equation models the relationships between the variables are perfectly linear. In real data this assumption may be unrealistic and violations of the assumption may not be benign. The skew t-distribution can instead accommodate approximate linearity and thus skew-SEM will be more accommodating to imperfections of the data that can be found in the real world. It is natural to assume that the relationship between the variables may not be exactly the same for observations in the center of the distribution and for observations in the tails of the distributions. Often skewness of the data is a sign of some kind of

non-linearity.

The skew t-distribution contains three different distributions as special cases: the skew normal distribution, the t-distribution, and the normal distribution. The fact that the normal distribution is a special case of the skew t-distribution allows us to easily compare skew-SEM with standard SEM using the likelihood ratio test (LRT) because the models are nested. In addition, if skew-SEM is not needed and is not appropriate for particular data set, the extra parameters in the skew t-distribution will not become statistically significant and therefore regular SEM would not be rejected in favor of skew-SEM. If the data does not support the need for skew and kurtosis modeling then the SEM model will arise naturally as the more parsimonious model.

Modeling with skew t-distribution is intended for those situations where the observed distribution is truly continuous and non-normal. Using the skew t-distribution is not suitable for modeling categorical data. Structural equation models based on the probit or logit link functions will still be preferable for categorical data.

Mixture modeling with continuous non-normal distributions is also very valuable. It is well known, see e.g. Schork and Schork (1988) and Bauer and Curran (2003), that mixture models of normal distributions rely heavily on the within-class normality assumption. If the normality assumption is not correct spurious classes can be found, i.e., latent subgroups can appear to exist only to accommodate the heavy tails of non-normal distributions rather than substantively meaningful latent subpopulations. However, if we use the more flexible skew t-distributions, we can resolve this problem. Latent classes found through mixtures of skew t-distributions would represent more meaningful

subpopulations. By allowing the within-class distributions to be skewed and to have heavy tails, we can focus on the true structural differences that are found in the latent classes. Spurious class formation due to non-normality and skewness will be eliminated.

Modeling with the skew t-distribution in general requires larger sample sizes than modeling with the normal distribution. The estimation of the skew-t distribution is based on being able to estimate well how heavy and how skewed the tails are and how the observed distribution curve deviates from the normal bell curve. To be able to extract this level of information from the data, a sufficient sample size is required. If the sample size is not sufficient the additional skewness parameters in the skew-t distribution will not be statistically significant and in that case they should be eliminated from the model to preserve model parsimony and minimize the standard errors for the remaining model parameters.

A number of articles have recently appeared that utilize the skew t-distribution for factor analysis models and mixture models, see for example, Lin et al. (2013) and Lee and McLachlan (2014). In this article we describe a general framework that includes general structural equation models based on the skew t-distribution as well as finite mixtures of such structural equation models. All models described in this paper can be estimated with Mplus Version 7.2.

2 Multivariate continuous skewed distributions

First we will define the skew t-distribution as the most general distribution considered in this paper. As a special case we derive the skew normal and the t-distributions. The normal distribution is also a special case of the

skew t-distribution. In this paper we adopt the parameterization for the skew t-distribution given in Lee and McLachlan (2014). Two different skew t-distributions are described in that article, the restricted and the unrestricted. These two distributions are not nested within each other and are equivalent only in the univariate case. We use only the restricted skew t-distribution because it allows explicit maximum-likelihood estimation for structural equation models. Suppose that a multivariate variable Y has a restricted skew t-distribution

$$Y \sim rMST(\mu, \Sigma, \delta, \nu), \tag{1}$$

where μ is a vector of intercepts, Σ is a variance covariance matrix, δ is a vector of skew parameters and ν is a positive parameter referred to as the degrees of freedom parameter. If Y is P dimensional variable, the size of the vectors μ and δ are also P and the variance covariance matrix Σ is of size $P \times P$. The density function of Y is given by

$$2t_{p,\nu}(y, \mu, \Omega)T_{1,\nu+p}(y_1/\lambda, 0, 1), \tag{2}$$

where

$$\Omega = \Sigma + \delta\delta^T, \tag{3}$$

$$d(y) = (y - \mu)^T\Omega^{-1}(y - \mu), \tag{4}$$

$$q = \delta^T\Omega^{-1}(y - \mu), \tag{5}$$

$$y_1 = q\sqrt{\frac{\nu + p}{\nu + d(y)}}, \tag{6}$$

$$\lambda^2 = 1 - \delta^T\Omega^{-1}\delta, \tag{7}$$

and $t_{p,\nu}(y, \mu, \Omega)$ is the multivariate t-distribution density function given by

$$t_{p,\nu}(y, \mu, \Omega) = \frac{\Gamma(\frac{\nu+p}{2})|\Omega|^{-1}}{(\pi\nu)^{p/2}\Gamma(\frac{\nu}{2})[1 + d(y)/\nu]^{(\nu+p)/2}} \tag{8}$$

and $T_{1,n}(z, 0, 1)$ is the standard univariate t-distribution function with n degrees of freedom.

With the above formulation the skew-t distribution reduces to the multivariate t-distribution if $\delta = 0$. The skew-t distribution reduces to the skew normal distribution if $\nu \to \infty$. The skew-t distribution reduces to the normal distribution if $\delta = 0$ and $\nu \to \infty$.

The multivariate skew t-distribution has the following stochastic representation.

$$Y = \mu + \delta|U_0| + U_1, \tag{9}$$

where U_1 is a vector of size P with a zero mean multivariate t-distribution and variance parameter Σ and degree of freedom parameter ν. The variable U_0 is a one-dimensional variable with standard t-distribution with mean 0, variance parameter 1 and degrees of freedom parameter ν. The variable U_0 is NOT independent of U_1, although the correlation between U_0 and U_1 is 0. This dependence between U_0 and U_1 is more of a technical issue rather than something that affects our development. For accuracy we provide the joint distribution for U_0 and U_1. The joint distribution is

$$(U_0, U_1) \sim t_{P+1}(0, \Sigma^*, \nu), \tag{10}$$

where

$$\Sigma^* = \begin{pmatrix} 1 & 0 \\ 0 & \Sigma \end{pmatrix}. \tag{11}$$

Another useful stochastic representation that can illuminate the dependence of U_0 and U_1 is a follows

$$Y = \mu + \delta \frac{|\bar{U}_0|}{\sqrt{W}} + \frac{\bar{U}_1}{\sqrt{W}}, \tag{12}$$

where

$$(\bar{U}_0, \bar{U}_1) \sim N(0, \Sigma^*), \tag{13}$$

$$W \sim Gamma(\nu/2, \nu/2), \tag{14}$$

$$U_0 = \frac{\bar{U}_0}{\sqrt{W}}, \tag{15}$$

$$U_1 = \frac{\bar{U}_1}{\sqrt{W}}. \tag{16}$$

The variables $\bar{U}_0$ and $\bar{U}_1$ are independent normal but U_0 and U_1 are connected through the Gamma distributed variable W.

Because of the absolute value around U_0 the distribution of the $|U_0|$ is skewed and it is essentially a half t-distribution. On the other hand, the distribution of U_1 is symmetric around 0 and thus the skewness of the distribution of Y is primarily due to the contribution of $|U_0|$ and if $\delta = 0$ the skewness of Y is 0. The variance parameter Σ is not exactly the variance of U_1. It is well known that the variance of the t-distribution is

$$Var(U_1) = \Sigma \frac{\nu}{\nu - 2} \tag{17}$$

when $\nu > 2$ and infinity otherwise. The parameter ν can be any positive number,

however the mean of Y is a finite number only if $\nu > 1$, the variance of Y is finite only if $\nu > 2$ and the skewness of Y is finite only when $\nu > 3$. Thus models with $\nu < 3$ should be used only for modeling data with substantial heavy tails and outliers. It is well known that the t-distribution with $\nu > 30$ closely approximates a normal distribution although a formal test for normality, i.e. $\nu = \infty$, should be conducted using the LRT test. Such a test should be used cautiously as we are testing boundary values. Testing the hypothesis $\nu = \infty$ with the T-test is formally not possible although testing the equivalent hypothesis $1/\nu = 0$ with the T-test is possible and would provide a good approximation in most cases. In many situations formal testing for normality should be conducted even when the estimated degrees of freedom parameter $\nu > 30$. The BIC criterion can also be used for model selection when formal testing is questionable.

Note that the skew t-distribution has exactly P+1 more parameters than the multivariate normal distribution. These are the P skew parameters δ and the degrees of freedom parameter ν. Note that ν is not variable specific. One parameter is used for the entire multivariate distribution. This will remain so even in structural equations models. While each variable in the structural equation model, observed or latent, will have its own skew parameter δ, the degrees of freedom parameter will be the same for all variables in the structural equation model. In mixture models or multiple group models the ν can be different across groups. The interpretation of the ν parameter is very simple. It is a general characteristic of how much deviation from normality there is in the population of variables, as measured by how much ticker the tails of the distributions can be as compared to the normal bell curve. The interpretation of the δ parameters is also very simple. The δ parameter is an indicator of how skewed the distribution

is to the left or the right. The δ parameter can be any real number, positive or negative, and a positive δ parameter yields a distribution skewed to the right while a negative δ parameter yields a distribution skewed to the left. Testing an individual variable for skewness is very simple as it is equivalent to $\delta = 0$ and can be performed with the standard T-test.[1] This allow us to also easily model skewed and non-skewed variables for example in the same model.

Modeling with the skew t-distribution can also be used for modeling with the skew normal distribution, the t-distribution, and the normal distribution. Fixing the ν parameter to a very large value such as 10000 yields the skew normal distribution. This essentially yields the same stochastic representation as (9) but now U_1 has a multivariate normal distribution, U_0 has a standard normal distribution, U_1 and U_0 are independent, and $|U_0|$ has a standard half normal distribution. Fixing all δ parameter to 0 we obtain the t-distribution. Fixing all δ parameter to 0 and the ν parameter to 10000 will yield the normal distribution.

2.1 Means, variance and skewness

The mean of Y for the skew t-distribution can be computed as follows

$$E(Y) = \mu + \delta \frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})} \sqrt{\frac{\nu}{\pi}}. \tag{18}$$

The variance of Y can be computed as follows

$$Var(Y) = \frac{\nu}{\nu - 2}(\Sigma + \delta\delta^T) - \left(\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})} \right)^2 \frac{\nu}{\pi} \delta\delta^T. \tag{19}$$

[1] In Mplus language the δ parameter for a variable Y is referred to as {Y} and the degrees of freedom parameter is referred to as {DF}.

The univariate skewness for a single Y variable can be computed as follows

$$Skew(Y) = v^{-3/2}\delta\sqrt{\frac{\nu}{\pi}}\left((2\delta^2+3\sigma)\frac{\nu}{\nu-2}\frac{\Gamma(\frac{\nu-3}{2})}{\Gamma(\frac{\nu-2}{2})} - \delta^2\frac{\nu}{\pi}\left(\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}\right)^3 - 3\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}v\right), \tag{20}$$

where $v = Var(Y)$ is given in the previous formula and the σ parameter is the diagonal element of Σ corresponding to the univariate variable. These formulas show that the ν and δ parameters affect all three quantities: the mean, the variance and the skew. The parameter μ affects only the mean and the σ parameters affect the variance covariance and the skew. With the skew t-distribution we do not have the simplicity of the normal distribution where μ is simply the mean and Σ is the variance covariance and they can be modeled independently. Here all three quantities are entangled and modeling one of them is not independent of the other.

For the skew-normal distribution the above formulas simplify to

$$E(Y) = \mu + \delta\sqrt{\frac{2}{\pi}}, \tag{21}$$

$$Var(Y) = \Sigma + \Big(1 - \frac{2}{\pi}\Big)\delta\delta^T, \tag{22}$$

$$Skew(Y) = v^{-3/2}\delta^3\sqrt{\frac{2}{\pi}}\Big(\frac{4}{\pi} - 1\Big). \tag{23}$$

From the last formula it is easy to see that the maximal skewness value for the skew normal distribution is obtained for $\sigma = 0$, which also implies that the Y variable is proportional to the half normal distribution. Thus within the family of skew normal distributions the maximum skewness that can be attained is the

skewness for the half normal distribution which is

$$\sqrt{\frac{2}{\pi-2}}\frac{4-\pi}{\pi-2}\approx 1. \tag{24}$$

More strictly speaking the skewness of a skew-normal variable is within the interval [-1,1]. This limit has very important modeling implications. The skew normal distribution can be used for modeling skewness but only if the skewness is moderately large. If the skewness observed in the data by absolute value exceeds 1 then probably the skew-normal distribution would not be a good fit and the skew t-distribution should be used instead. The skew-t distribution can attain any level of skewness.

For the t-distribution the above formulas simplify to

$$E(Y)=\mu, \tag{25}$$

$$Var(Y)=\frac{\nu}{\nu-2}\Sigma, \tag{26}$$

$$Skew(Y)=0. \tag{27}$$

We don't provide an explicit formula for the kurtosis for Y, however the kurtosis for the T-distribution alone is $6/(\nu-4)$ and therefore the skew t-distribution alone can be used to model any level of kurtosis.

2.2 Marginal and conditional and distributions

Obtaining the marginal distribution for the skew t-distribution is very simple. Suppose that Y has a skew t-distribution

$$Y \sim rMST(\mu, \Sigma, \delta, \nu). \tag{28}$$

Suppose that the vector Y is decomposed in two parts $Y = (Y_1, Y_2)$ where Y_1 is a vector of dimension P_1 and Y_2 is a vector of dimension P_2, where $P = P_1 + P_2$. Suppose also that the corresponding decomposition of the parameters is $\mu = (\mu_1, \mu_2)$, $\delta = (\delta_1, \delta_2)$ and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{29}$$

Then the marginal distribution of Y_1 is

$$Y_1 \sim rMST(\mu_1, \Sigma_{11}, \delta_1, \nu). \tag{30}$$

Thus the marginal properties of the skew t-distributions are very similar to those of the normal distribution. The same logic applies also for the skew normal and the t-distributions.

The conditional distribution of $[Y_1|Y_2]$, however, is somewhat more complicated. In fact it is shown in Arellano-Valle and Genton (2010) that this conditional distribution is no longer a skew t-distribution but is the so called extended skew t-distribution. Let's first focus on the conditional t-distribution, i.e., assuming

that $\delta = 0$. It is shown in Liu and Rubin (1995) that if

$$Y \sim t(\mu, \Sigma, \nu) \tag{31}$$

then

$$[Y_1|Y_2] \sim t(\mu_1^*, \Sigma_{11}^*, \nu + P_2), \tag{32}$$

where

$$\mu_1^* = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2), \tag{33}$$

$$\Sigma_{11}^* = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\frac{\nu + (Y_2 - \mu_2)^T\Sigma_{22}^{-1}(Y_2 - \mu_2)}{\nu + P_2}. \tag{34}$$

The implication of the above formulas is that the conditional expectation

$$E(Y_1|Y_2) = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2) \tag{35}$$

behaves and is computed exactly the same way as the normal conditional expectation. However, the conditional variance of $Var(Y_1|Y_2)$ is not computed the same way as for the normal distribution. In addition, note that the joint distribution of two independent t-distributions is not a t-distribution. That is, if $Y_1 \sim t(0, I, \nu)$ and $Y_2 \sim t(0, I, \nu)$, where I represents the identity matrix, and Y_1 and Y_2 are independent then Y is not $t(0, I, \nu)$. Note also that even if the covariance between Y_1 and Y_2 is 0, the variables Y_1 and Y_2 are not independent because, the conditional variance of $Var(Y_1|Y_2)$ depends of the value of Y_2 even though the conditional mean $E(Y_1|Y_2)$ does not depend on Y_2. The further away Y_2 is from its mean as measured by the Mahalanobis distance the bigger the conditional variance of Y_1 will be.

Next we focus on the conditional expectation for the skew t-distribution. For simplicity we will illustrate that only for the bivariate case. Assuming $Y = (Y_1, Y_2)$ and $Y \sim rMST(\mu, \Sigma, \delta, \nu)$ we want to compute $E(Y_1|Y_2)$. Let

$$\Omega = \Sigma + \delta\delta^T = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}, \tag{36}$$

and

$$\bar{\delta} = \frac{1}{\sqrt{1 - \delta\Omega^{-1}\delta^T}}\delta^T\Omega^{-1}, \tag{37}$$

$$\alpha = \frac{\nu + (Y_2 - \mu_2)^2/\omega_{22}}{\nu + 1}, \tag{38}$$

$$\omega^* = \omega_{11} - \omega_{21}\omega_{22}^{-1}\omega_{12}, \tag{39}$$

$$\tau^* = (\bar{\delta}_1\omega_{21}/\omega_{22} + \bar{\delta}_2)(Y_2 - \mu_2)\frac{1}{\sqrt{\alpha + \alpha\bar{\delta}_1^2\omega^*}}. \tag{40}$$

Then

$$E(Y_1|Y_2) = \mu_1 + \omega_{12}\omega_{22}^{-1}(Y_2 - \mu_2) + \frac{\bar{\delta}_1\omega^*\sqrt{\alpha}(\nu + 1 + \tau^{*2})}{\nu\sqrt{1 + \bar{\delta}_1^2\omega^*}}\frac{t_1(\tau^*, \nu + 1)}{T_1(\tau^*, \nu + 1)}, \tag{41}$$

where $t_1(*, \nu + 1)$ and $T_1(*, \nu + 1)$ are the density and the distribution function of the standard t-distribution with $\nu + 1$ degrees of freedom. Note that the first two terms in the above expression resemble the normal based conditional expectation treating Ω as the variance covariance matrix. The third term represents the non-linear dependence of this conditional expectation with respect to Y_2. Despite the complex expression there is a simple way to test statistical significance for a Y_2 effect on $E(Y_1|Y_2)$. If $\omega_{21} = 0$ and $\sigma_{21} = 0$ then $E(Y_1|Y_2)$ is independent of Y_2 and

thus a statistically significant effect exists if either ω_{21} or σ_{21} is significant. [2]

In the case of the skew normal distribution the above expression simplifies to

$$E(Y_1|Y_2) = \mu_1 + \omega_{12}\omega_{22}^{-1}(Y_2 - \mu_2) + \frac{\bar{\delta}_1\omega^*}{\sqrt{1+\bar{\delta}_1^2\omega^*}}\frac{\phi(\tau^*)}{\Phi(\tau^*)}, \tag{42}$$

where ϕ and Φ represent the standard normal density and distribution functions.

2.3 Factor model interpretation

Equation (9) has a special factor model interpretation. Consider first the case of a skew normal distribution. In that case both U_1 and U_0 are normally distributed variables. The model represented in equation (9) is simply a factor analysis model where the factor $|U_0|$ has a half-normal distribution. We will refer to this as the underlying skew factor of the distribution. The skew parameters δ are nothing more than the factor loadings for this factor. Note that if that factor had a normal distribution then the model would not be identified because U_1 has an unrestricted variance covariance matrix. The fact that $|U_0|$ has a skewed half-normal distribution is key to identifying the skew parameters δ. This model is identified entirely from the skewness in the data. When the model is fitted to the data, the skew parameters will be set so that the skewness of the data is represented by the component $\delta|U_0|$, while the remaining part of the observed variable Y

$$Y - \delta|U_0| = \mu + U_1 \tag{43}$$

[2]This joint test can be done in Mplus with the Model Test command.

is normal. If the variable Y has a skew t-distribution the same interpretation is given but now the variables U_0 and U_1 have a t-distribution instead of normal. Thus the goal of adding a skew factor to the modeling distribution is to take into account the skewness of the data. Note however, that the skew factor also contributes to the mean and the variance covariance matrix of Y as it can be clearly seen from the formulas in Section 2.1. The means and variance covariances however can be fitted further through the μ and Σ parameters.

3 The skewed structural equation model

Suppose that we have a vector of observed dependent variables Y of dimension P, a vector of observed dependent variables X of dimension Q, and a vector of latent variables η of dimension M. We are interested in constructing a structural equation model where all variables have a skew t-distribution. The structural equation model is given by the usual equations

$$Y = \nu + \Lambda\eta + \varepsilon, \tag{44}$$

$$\eta = \alpha + B\eta + \Gamma X + \xi, \tag{45}$$

where

$$(\varepsilon, \xi) \sim rMST(0, \Sigma_0, \delta, DF), \tag{46}$$

and

$$\Sigma_0 = \begin{pmatrix} \Theta & 0 \\ 0 & \Psi \end{pmatrix}. \tag{47}$$

The vector of parameters δ is of size $P+M$ and can be decomposed as $\delta = (\delta_Y, \delta_\eta)$. The vector δ_Y is a vector of skew parameters of dimension P which we will refer to as the skew parameters for the Y vector. The vector δ_η is the vector of skew parameters for the latent variables η.[3] From the above equations we obtain the conditional distributions

$$\eta|X \sim rMST((I-B)^{-1}(\alpha+\Gamma X), (I-B)^{-1}\Psi((I-B)^{-1})^T, (I-B)^{-1}\delta_\eta, DF), \quad (48)$$

$$Y|X \sim rMST(\mu, \Sigma, \delta_2, DF), \quad (49)$$

where as usual

$$\mu = \nu + \Lambda(I-B)^{-1}(\alpha + \Gamma X), \quad (50)$$

$$\Sigma = \Theta + \Lambda(I-B)^{-1}\Psi((I-B)^{-1})^T\Lambda^T, \quad (51)$$

$$\delta_2 = \delta_Y + \Lambda(I-B)^{-1}\delta_\eta. \quad (52)$$

In the above setup all variables, dependent observed variables Y, latent factors η, and residual variables ε and ξ, all have skewed distributions. As usual the distribution of the covariates X is not modeled. This actually is very important in the skew-SEM framework. In the standard SEM framework the model for the covariates can be optionally included. For example adding an unrestricted model for the covariates X, where the means are estimated at the sample means and the variance covariance matrix for X is estimated at the sample variance covariance matrix, does not affect the estimation of the structural equation model. This, however, is not the case for skew-SEM. The reason is that if we assume an

[3] In the Mplus language these parameters are referred to as {Y} and {η}.

unrestricted skew t-distribution for X then we will allow the DF parameter to be influenced by the distribution of X. The DF parameter is common for all variables and thus it will be affected by the covariates if they are included in the modeling. Thus we want to consider a true conditional distribution $[Y|X]$, where the X covariates are not modeled. To illustrate this further a simple path analysis regressing Y_1 on Y_2 will be a different model if Y_2 is a dependent variable that has an estimated skew t-distribution from a model where Y_2 is treated as a true covariate where no model is assumed for the distribution of Y_2 and only the conditional distribution of $[Y_1|Y_2]$ is modeled. Because of this treatment of the covariates X in the above model the linear dependence of $E(Y|X)$ and $E(\eta|X)$ is preserved but only if it is direct and not channeled through a non-linear dependence. For example, if η is regressed on a covariate X directly then $E(\eta|X)$ is linear in X. If, however, Y is regressed on η and the model implies a non-linear expression of $E(Y|\eta)$ in terms of η then $E(Y|X)$ will also be non-linear in terms of X. If $E(Y|\eta)$ is linear in terms of η then so will be $E(Y|X)$ in terms of X. When $E(Y|X)$ is linear in terms of X the slope in front of X is obtained exactly the same way it is obtained for the standard SEM.

In the above structural equation model the skew parameters δ_Y and δ_η are subject to identifiability just as the rest of the structural parameters. No more than P skew parameters can be identified in the above model. To understand this it is helpful to use the interpretation where the skew parameters are simply the loadings for the skew factor U_0. We can identify a maximum of P covariances between Y and U_0 and thus we can identify no more than P skew parameters. The skew parameters also behave the same way intercept parameters do. We can not identify more than P parameters among ν and α. Two special cases can be

mentioned. The first case is where $\delta_Y = 0$, i.e., δ_Y are fixed to 0. In that case the residual for Y is not skewed; it is either the symmetric t-distribution or the normal distribution if we are modeling with the skew normal distribution. Here we can also maintain the linearity in the conditional expectation $E(Y|\eta)$ with respect to η. The second special case is the case where $\delta_\eta = 0$. In that case the factor distribution is assumed to be symmetric and all the skewness in the data is assumed to come from the residuals of Y. In this case we also preserve the linearity in the conditional expectation $E(Y|\eta)$. In most common situations where a factor analysis model is concerned and a measurement instrument is modeled the factor is intended to extract the maximum amount of correlation among the measurement variables. If we use $\delta_\eta = 0$ and estimate δ_Y the correlation among the Y variables that is due to the skew factor will be taken away from the measured factor variable and thus this model is undesirable. In a common practical application we would want as much of the skewness in Y to be explained through the factor η. Thus as an optimal strategy for which skew parameters to estimate we would recommend estimating δ_η and estimating only those δ_Y that are statistically significant, i.e., assume that most of the skewness in the observed data can be explained through skewness of the factor and if some residual Y skewness is still left and significant it should only then be estimated. Naturally, models with minimal amount of δ_Y would be preferable. When for a particular measurement the δ_Y is estimated and is significant, the linearity property of $E(Y|\eta)$ will no longer hold for that measurement variable. The interpretation in that case is clear, for that particular measurement variable the linearity is insufficient. The kind of skewness observed in the data is due to more complex relationship between the latent factor and the measurement variable.

3.1 Estimation

The models are estimated by maximum likelihood. Using equations (2) and (49) the log-likelihood can be written explicitly and maximized with a general maximization algorithm such as the Quasi-Newton optimization method as long as the derivatives of the log-likelihood can be computed. All of these derivatives, while intricate, are computable. The only derivative that is a matter of more advanced methods is the derivative of $T_1(x, \nu)$ with respect to ν, where T_1 is the standard t-distribution function. For this derivative we have utilized the method developed in Boik and Robison-Cox (1998). Most other published articles on similar models have used the EM-algorithm where the skew factor U_0 and the Gamma distributed variable W are treated as unobserved variables, see for example Lin et al. (2013) as well as Liu and Rubin (1995) for the t-distribution. We have found that somewhat unnecessary. Direct maximization appears to work well and is relatively fast. The standard error estimates are based on the inverse of the information matrix as usual with the ML estimator and robust standard errors can also be computed using the sandwich estimator.[4]

3.2 The dilemma of $\lambda = 0$

One of the underlying restrictions in the parameters in the skew t-distributions arises from equations (2) and (7). The parameter λ is defined only when

$$1 - \delta^T(\Sigma + \delta\delta^T)^{-1}\delta \geq 0 \tag{53}$$

[4]Complex survey features of stratification, weights, and clustering are also handled in Mplus.

In fact because in equation (2) we divide by λ the above inequality has to be strict. It turns out that when Σ is positive definite then the above inequality is always satisfied but when Σ is not positive λ can converge towards zero and the parameter estimates can land on this boundary condition

$$1 - \delta^T(\Sigma + \delta\delta^T)^{-1}\delta = 0 \tag{54}$$

at which point estimation can become very difficult as numerically we operate in a small band near the boundary condition. In addition the term $T_{1,\nu+p}(y_1/\lambda, 0, 1)$ becomes either 0 or 1. In some cases the term will be 1 for almost all observations and near 0 only for one or two observations. What this implies is that the log-likelihood value will be driven primarily by those one or two observations. With different starting values the small number of observations that drive the log-likelihood value may change and thus it may appear that when we choose different starting values we obtain different optimal estimates. It also appears that when λ converges to zero the log-likelihood that we are optimizing becomes quite rugged and indeed a number of different solutions can be found. Estimating a model where Σ is no longer positive definite in many way is similar to what is known in factor analysis models as a Heywood case. The maximum-likelihood estimation converges towards a singular Σ and possibly to a solution with negative variance. Due to the boundary condition however these solutions are ill defined. Consider the interpretation of the skew t-distribution as a factor analysis on a skewed factor with half t-distribution. A formal Heywood case is exactly that, namely, that the residual variance Σ can become not positive definite.

Another interpretation of the $\lambda = 0$ case is revealed when you consider the

univariate skew t-distribution. In that case $\lambda = 0$ implies that the stochastic decomposition (9) collapses down to

$$Y = \mu + \delta|U_0| \tag{55}$$

or equivalently $Var(U_1) = 0$ and the residual distribution of Y no longer consist of a linear combination of a t-distribution and half t-distribution but only of the half t-distribution. On the other hand, if $\delta = 0$ and $Var(U_1)$ is not zero we get the case where the residual is not a combination of both distributions but only of the t-distribution. This case of course is well behaved. In particular, if the DF is large this is essentially the normal distribution. Thus having $\lambda = 0$ is nothing more than another special situation of the skew t-distribution where the residual has a skew t-distribution. Unfortunately, however, numerically this special case is not easy to handle. In the multivariate case the failure of Σ to stay positive definite can occur in more complicated ways than just having a residual variance converge to 0. The non-positive definiteness can be due to a particular combination of the normal residuals having zero variance which will be hard to interpret and deal with. In such situations model modifications that convert a covariance relationship into regression can be useful, see for example the relationship between the variance covariance saturated model and the sequential saturated model described in the next two sections.

Another more critical interpretation of the $\lambda = 0$ case is that the skew t-distribution model has failed to extract the skewness of the data and the skew factor analysis is essentially not identified by the skewness of the data but is simply extracting the covariance and as such the Σ matrix is no longer identified

separately. The original idea of the skew factor is that the variance covariance matrix will be fully identified after the skewness of the data is taken into account by the skew factor. When Σ is no longer a valid variance covariance matrix it appears that conceptually the skew t-distribution has failed and possibly it is not an appropriate distribution. It is interesting, however, that the skew t-distribution model can usually be estimated for any smaller subset of variables, i.e., the skewness in the data can well identify the skewness parameters and factor. For models with a larger number of variables, however, it is increasingly likely that $\lambda = 0$ occurs.

The problem with $\lambda = 0$ appears to happen often enough that it becomes a critical issue for skew-SEM. The occurrence of $\lambda = 0$ needs to be monitored.[5]

3.3 The unrestricted model

Just like the standard SEM models, the skew-SEM are nested within a saturated model. Comparison between a structural equation model and the saturated model provides a test of fit for the structural equation model. The skew saturated model, which we also refer to as the H1 model is given by

$$Y = \nu + \Gamma X + \varepsilon \tag{56}$$

[5]In Mplus the final estimated λ is reported at the end of the technical 8 output section and it should be monitored. In most cases a value above 0.001 is evidence that the parameter estimates are away from the boundary condition. If however the value becomes less than 0.001 then Mplus will suggest that multiple random starting values are used to verify that the most optimal solution is reached. Even if the most optimal value is not reached however the model can still be interpreted and used. It maybe difficult to run a huge number of starting values to search for the best solution when $\lambda = 0$ and the log-likelihood has many local maxima.

The optimal estimation, understanding and handling of the case $\lambda = 0$ may still be out of reach with the current algorithm implemented in Mplus Version 7.2. What makes things even more complicated is that this case appears only for real data sets and not for simulated data, i.e., it is difficult to demonstrate the $\lambda = 0$ case with a simulation study.

where

$$\varepsilon \sim rMST(0, \Theta, \delta, DF). \tag{57}$$

The number of parameters in this model are as follows. The vector ν has P parameters, the matrix Γ has $P \times Q$ parameters, the matrix Θ has $P(P+1)/2$ parameters, the vector δ has P parameters and the DF parameter is just a single parameter. Thus the total number of parameters in the skew saturated model is

$$2P + P(P+1)/2 + PQ + 1 \tag{58}$$

which is $P + 1$ more parameters than the saturated normal model. Any skew structural equation model is a restriction of the above model. If we refer to the structural equation model as the H0 model the test of fit that can be constructed comparing the skew structural equation model H0 and the skew saturated model H1 is the LRT test based on

$$T = 2(LL_{H1} - LL_{H0}). \tag{59}$$

Under the null H0 hypothesis the distribution of T is a chi-square distribution with D degrees of freedom where D is the difference between the number of parameters in the H1 and H0 models. Other test of fit that can be of interest is comparing a structural normal model against a saturated skew model. Such a test will show a test of fit for the standard SEM that goes beyond a test of fit for the mean and variance. It would test if the standard SEM fits the data well, including the potential skewness of the data. Other tests that can be of interest are the test of the skew normal structural equation model against the saturated skew normal or

the structural t-distribution model against the saturated t-distribution model.[6]

For large number of variables the H1 model often will lead to a $\lambda = 0$ solution and possible convergence problems. Thus in the next section we suggest an alternative saturated model parameterization that yields better convergence rates.

3.4 The alternative sequential unrestricted model

The sequential unrestricted model is given by the following equations

$$Y = \nu + BY + \Gamma X + \varepsilon \tag{60}$$

where

$$\varepsilon \sim rMST(0, \Theta, \delta, DF). \tag{61}$$

and Θ is a diagonal matrix while the matrix B has all entries on and below the diagonal fixed to 0, i.e., instead of estimating a full variance covariance matrix Θ we estimate a diagonal Θ and all Y variables are regressed on the following Y variables. That is, variable Y_1 is regressed on Y_2,..., Y_P. Variable Y_2 is regressed on Y_3,..., Y_P, etc. This model has the same number of parameters as the variance covariance unrestricted model described in the previous section and is equivalent to that model. Under normal circumstances the two models should yield the same log-likelihood value and a test of fit. This parameterization has the advantage that the model does not have a parametric variance covariance matrix that has

[6]The test of model fit within the same family of distributions can be obtained automatically in Mplus with the H1MODEL option of the OUTPUT command. The test of fit is not computed by default, as with standard SEM, because the estimation of the H1 model may be more difficult than the estimation of the H0 model and may take longer to estimate especially if multiple random starting values are used. Thus the H1 model will be estimated only if it is requested.

to stay positive definite. In the optimization algorithm it is much easier to keep individual residual variance parameters to be positive than to keep a multivariate matrix to be positive definite. Thus the sequential unrestricted model yields better convergence rates, however, it is slower to estimate and so it should be used only when the variance covariance unrestricted model described in the previous section does not converge.[7]

3.5 Restricted v.s. unrestricted skew t-distribution

The restricted skew t-distribution has one major assumption about the skewness in the data. The assumption is that the skewness is due to one single skew factor, U_0. An alternative model that allows each residual to have a univariate independent skew distribution is in principle a possibility, for example the unrestricted skew t-distribution has this capability. Currently however such distributions do not generalize easily to structural equation models in the following sense. In the unrestricted skew t-distribution each residual has its own skew factor and the number of skew factors has to be exactly the same as the number of variables. If the latent variable and the residuals have their own skew factors the number of skew factors in the model will be larger than the number of observed variables, i.e., the model implied distribution for the observed variables is not an unrestricted skew t-distribution. Thus the observed likelihood does not have a closed form expression. In such a situation a direct maximum-likelihood estimation is not possible but alternative estimation methods using numerical integration or Bayesian estimation is possible. The unrestricted and the restricted skew-t

[7] In the Mplus language to obtain the sequential unrestricted model estimation one has to use the H1MODEL(sequential) option of the OUTPUT command.

distributions are not nested within each other. In the unrestricted version the correlation between the skew factors is 0, while in the restricted version the correlation is 1, leading to a unique skew factor.

In the restricted skew t-distribution, the assumption that the skewness of the data has a common source is not unrealistic for the kind of data that is typically used with structural equation models. For example, when multiple characteristics of an individual are observed and skewness is present in these characteristics, it is not an unreasonable assumption to hypothesize that one underlying individual characteristic exists that makes the individual more or less extreme and out of the norm on all of the manifest variables. The restricted skew t-distribution is also a parsimonious model. It allows us to model skewness of the data without sacrificing simplicity and interpretation.

3.6 Estimation of the factor scores

We estimate the factor score using the conditional expectation $E(\eta|Y, X)$. The joint distribution of η and Y is given by

$$\eta, Y|X \sim rMST(\mu, \Sigma, \delta_2, DF), \tag{62}$$

where

$$\mu = ((I - B)^{-1}(\alpha + \Gamma X), \nu + \Lambda(I - B)^{-1}(\alpha + \Gamma X)), \tag{63}$$

$$\delta_2 = ((I - B)^{-1}\delta_\eta, \delta_Y + \Lambda(I - B)^{-1}\delta_\eta), \tag{64}$$

$$\Sigma = \begin{pmatrix} (I - B)^{-1}\Psi((I - B)^{-1})^T & (I - B)^{-1}\Psi((I - B)^{-1})^T\Lambda^T \\ \Lambda(I - B)^{-1}\Psi((I - B)^{-1})^T & \Theta + \Lambda(I - B)^{-1}\Psi((I - B)^{-1})^T\Lambda^T \end{pmatrix}. \tag{65}$$

Give that the joint distribution is a multivariate skew t-distribution we can use the method of Arellano-Valle and Genton (2010), which was also illustrated for the bivariate case in (41), to estimate $E(\eta|Y, X)$.

3.7 Estimation of direct and indirect effects in mediation models

Consider the following mediation path analysis model

$$Y_1 = \alpha_1 + \beta_1 Y_2 + \beta_2 X + \varepsilon_1, \tag{66}$$

$$Y_2 = \alpha_2 + \beta_3 X + \varepsilon_2. \tag{67}$$

The usual definitions of a direct and indirect effect are β_2 and $\beta_1 \times \beta_3$ (see, e.g., MacKinnon, 2008). Using the skew-t distribution for ε_1 and ε_2, these effects remain valid even though $E(Y_1|Y_2)$ is not linear in terms of Y_2 . Let the variance and the skew parameters for ε_i be σ_i and δ_i. Let the degrees of freedom parameter be ν. Consider the definitions based on counterfactuals (see, e.g. VanderWeele & Vansteelandt, 2009; Muthén & Asparouhov, 2014a), also referred to as causal effects. Letting $M = Y_2$ and $Y = Y_1$, the key component of the causal effect definitions, $E[Y(x, M(x^*)]$, can be expressed as follows integrating over the mediator M

$$E[Y(x, M(x^*))] = \int_{-\infty}^{+\infty} E[Y|X = x, M = m] \times f(m|X = x^*)\, \partial m. \tag{68}$$

The above integral is the marginal mean of the skew-t distribution $rMST(\mu, \sigma, \delta, \nu)$ where

$$\mu = \alpha_1 + \beta_1(\alpha_2 + \beta_3 x^*) + \beta_2 x, \tag{69}$$

$$\sigma = \sigma_1 + \beta_1^2 \sigma_2, \tag{70}$$

$$\delta = \delta_1 + \beta_1 \delta_2. \tag{71}$$

Using formula (18) we obtain

$$E[Y(x, M(x^*))] = \alpha_1 + \beta_1(\alpha_2 + \beta_3 x^*) + \beta_2 x + (\delta_1 + \beta_1 \delta_2) \frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})} \sqrt{\frac{\nu}{\pi}}. \tag{72}$$

The causal direct and indirect effects are computed from

$$E[Y(x_1, M(x_1^*))] - E[Y(x_2, M(x_2^*))] = \beta_1 \beta_3 (x_1^* - x_2^*) + \beta_2 (x_1 - x_2). \tag{73}$$

In line with VanderWeele and Vansteelandt (2009), special cases of this formula give the direct effect as

$$E[Y(x, M(x^*))] - E[Y(x^*, M(x^*))] = \beta_2 (x - x^*) \tag{74}$$

and the indirect effect as

$$E[Y(x, M(x))] - E[Y(x, M(x^*))] = \beta_1 \beta_3 (x - x^*). \tag{75}$$

The above formulas are identical to the normal distribution case and thus the direct and indirect effects are not affected by the skewness of the residuals.

3.8 Missing data

Given that the marginal distribution is easy to derive for the skew t-distribution, see Section 2.2, we can still compute and optimize the observed data log-likelihood directly. The ML estimator can guarantee unbiased parameter estimates under the general missing at random assumption.

3.9 Mixture modeling

The general mixture of skew structural equation models is similar to the mixture of normal structural equation models given in Muthén and Shedden (1999) and Muthén and Asparouhov (2009). Within each class, however, we now have a skew structural equation model as in (44-47) where all the coefficients are now class specific including the skew parameters and the degree of freedom parameters. The estimation also follows the estimation method used in Muthén and Shedden (1999) and Muthén and Asparouhov (2009). To use that EM-based algorithm all we need is the ability to compute the log-likelihood for Y conditional on C. This is simply accomplished by using formula (2) with the class specific parameters.

4 Examples

In this section we discuss some of the basic concepts of modeling with the skew t-distribution using several real data examples and simulated examples. Both path analysis and factor analysis examples are considered.

4.1 National Longitudinal Survey of Youth BMI example

Data from the 1997 National Longitudinal Survey of Youth (NLSY97) are used to illustrate the new methods. NLSY97 is a nationally representative, longitudinal survey of people born between 1980 and 1984 who were living in the United States in 1997. A more detailed description of the data can be found in Nonnemaker et al. (2009).[8] For our illustration we use the subsample of females and we use the body mass index (BMI) variable at age 12 and age 17. These two variables are strongly non-normal with sample skewness/kurtosis values 1.34/2.77 and 1.86/5.29, respectively. The sample consist of 3839 individuals. We estimate the path analysis regression

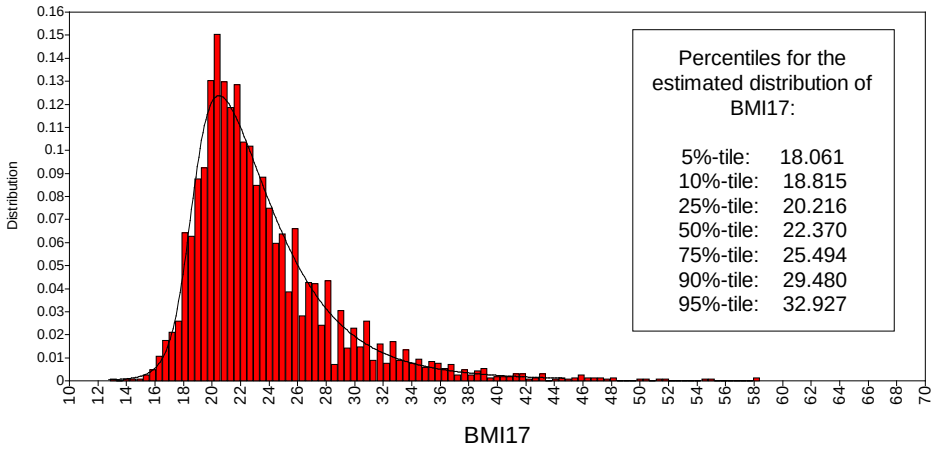
$$BMI17 = \alpha + \beta BMI12 + \varepsilon \tag{76}$$

assuming a skew t-distribution for both $BMI12$ and $BMI17$. Figure 1 shows the observed distribution in the sample for BMI17 as well as the skew-t estimated distribution.

The parameter estimates for this analysis and their SE are given in Table 1. There are a total of 8 parameters, α and β from the above regression model, the intercept parameter μ_{BMI12} for BMI12, the variance parameter θ_{BMI12} for BMI12, the skew parameter δ_{BMI12} for BMI12, the residual variance parameter θ_ε for the variance of ε, the skew parameter δ_ε for ε and the degrees of freedom parameter DF. All the parameters are significant. The log-likelihood for this model is -11769.657 and

[8]We thank James Nonnemaker for providing the data to us.

Figure 1: Observed and estimated distribution for BMI17



the log-likelihood for the same model assuming normality is -12664.533. The chi-square test for the skew-t model against the normally distributed regression model has a value of 1790 with 3 degrees of freedom which clearly rejects the normally distributed model in favor of the more flexible skew-t distribution. Next we use formula (41) to compute the conditional expectation of $E(BMI17|BMI12)$. Because both $BMI12$ and ε have non-zero skew parameters this conditional expectation will be non-linear in terms of $BMI12$. For comparison purposes we also compute the conditional expectation for the normally distributed model. The two expectation functions are plotted in Figure 2. The non-linearity of the conditional expectation function for the skew t-distribution is clearly visible. In addition, the difference between the function estimated by the skew t-distribution and the function estimated by the normally distribution becomes substantial in the tail of the distribution of the BMI variables, i.e., the normally distributed model fails to provide sufficiently accurate results for the tails of the distributions.

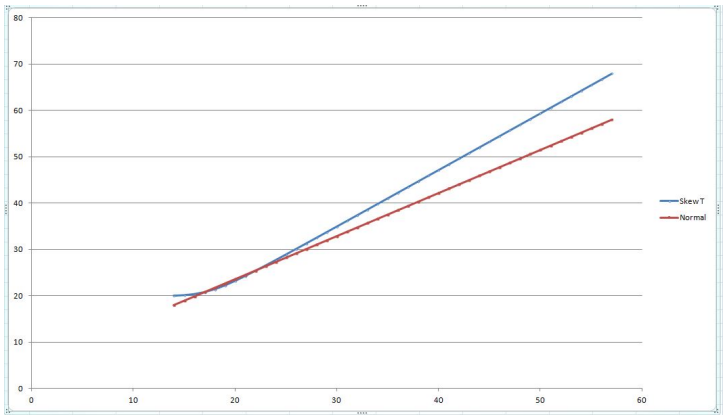
Note also that the β coefficient is 0.179. The corresponding coefficient for the normal regression is 0.931. These two coefficients are not comparable. In the

Table 1: BMI example parameters

parameter	estimate(SE)
α	15.565(1.436)
β	0.179(0.085)
μ_{BMI12}	16.770(0.217)
θ_{BMI12}	3.824(0.562)
θ_{ε}	1.431(0.179)
δ_{BMI12}	2.990(0.253)
δ_{ε}	4.409(0.281)
DF	3.870(0.268)

skew-t model much of the effect of $BMI12$ on $BMI17$ is channeled through the skewness factor. Note also that in the path analysis model above the residual variable ε and $BMI12$ are not independent of each other. This independence is a standard assumption of the linear regression model with normally distributed variables. Robustness of normal-theory maximum-likelihood estimation against non-normality relies on this independence assumption (see, e.g. Satorra, 2002). In the skew-t model the independence will hold if either of the two skew parameters is not present and the degrees of freedom parameter is large. If the degrees of freedom parameter is not large and only one of the two skewness parameters is present then $BMI12$ and ε would be uncorrelated but not independent. When only one of the skewness parameters is present in the model the non-linearity depicted in Figure 2 will also disappear. In this example, the skewness of BMI12 does not fully account for the skewness of BMI17 and therefore a skewness parameter is needed for both variables.

Figure 2: $E(BMI17|BMI12)$ as a function of BMI12



4.2 The ATLAS mediation example

The ATLAS mediation model was considered in MacKinnon et al. (2004). The intervention program ATLAS (Adolescent Training and Learning to Avoid Steroids) was administered to high school football players to prevent the use of anabolic steroids. The data consists of 404 individuals in the treatment group and 457 individuals in the control group. The two variables that will be analyzed are the SEVERITY and the NUTRITION variables. The SEVERITY variable represents the perceived severity of using steroids. The NUTRITION variable represents good nutrition behavior. It is hypothesized that the treatment variable TX increases the NUTRITION variable indirectly by increasing the SEVERITY variable which in turn positively affects the NUTRITION variable. Typically this hypothesis is tested with the following mediation model

$$SEVERITY = \alpha_1 + \beta_1 TX + \varepsilon_1, \tag{77}$$

$$NUTRITION = \alpha + \beta_2 SEVERITY + \beta_3 TX + \varepsilon_2, \tag{78}$$

where ε_1 and ε_2 are assumed normal and the ML estimation yields standard regression results. The main parameter of interest is the indirect effect parameter $\beta_1\beta_2$.

The problem with the standard mediation model is that it does not provide a good representation for these data. While the variable NUTRITION is approximately normally distributed the SEVERITY variable is highly skewed. For nearly half of all individuals in the sample the value of the SEVERITY variable is 7 and that is the maximum value that can be obtained. Figures 3 and 4 contain the histogram of the SEVERITY variable for the treatment and the control groups. The linear model (77) is not a good representation of these data because it implies that all individuals benefit equally from the treatment effect which is clearly not true because individuals that are at the maximum value will benefit 0. Despite that, model (77) can be used to estimate the means of SEVERITY in the control and the treatment groups correctly. It can not however be used to make inference for a particular individual and it can not be used to predict the treatment effect for a particular individual or even for a particular school. Only a model that truly represents the data can be used for this kind of detailed inference. From the histograms of SEVERITY it is clear that the effect of the treatment is to thin out the left skewed tail rather than to shift the distribution as the linear regression model implies. In normal distributions if the variance changes the mean of the distribution is not affected. In skewed distributions this is not the case. By changing the variance of the skew component the mean/average of the distribution is affected. This can be seen in formula (18) and it can be seen by comparing the histograms in Figures 3 and 4.

The skew-normal distribution can be used in this example to provide a more

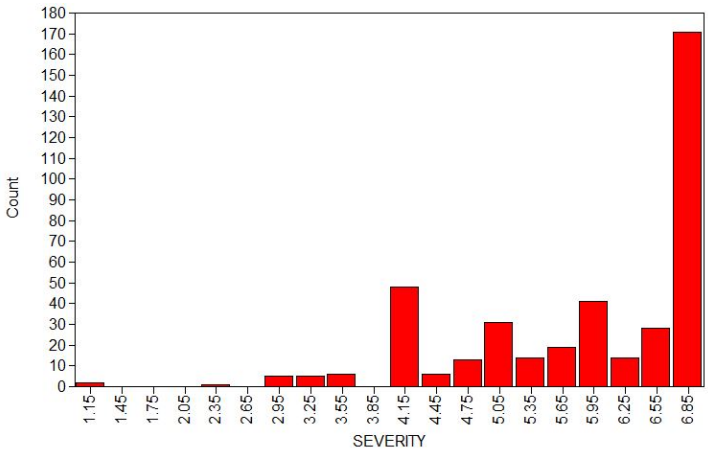
realistic model for the ATLAS data. To fully allow the TX variable to affect the distribution of SEVERITY we replace equation (77) with

$$SEVERITY|TX = j \sim rMSN(\mu_j, \sigma_j, \delta_j), \tag{79}$$

where $j = 0$ or 1, i.e., we model the distribution as a skew-normal distribution with group specific parameters. The second equation (78) for NUTRITION remains unchanged. If we use the skew-t distribution the degrees of freedom parameter is estimated to a large value and thus it is not needed. Modeling the NUTRITION variable as a skew-normal instead of normal is also not needed. The parameter estimates indicate that $\mu_0 = \mu_1$ which shows as expected that the treatment effect does not provide a shift in the distribution. The variance parameters $\sigma_0 = \sigma_1 = 0$ which means that the best approximate distribution for the SEVERITY variable in the skew-normal family is the half-normal distribution. The skew parameters δ_0 and δ_1 are significantly different and equation (18) can be used to obtain the effect of the intervention on the average SEVERITY value. The BIC value for SEVERITY|TX for the skew-normal model is 6199 and for the standard normal model is 6836 which indicates that the skew-normal model is a much better fit for these data. The direct effect estimate and its standard error are 0.016(0.008) for the skew-normal model and 0.020(0.011) for the normal model both indicating marginal statistical significance.

There are three advantages of the skew-normal model in this example. First we obtain a model that is a better fit for the data and provides a better representation for the processes and variables. Second, the model can be used for better predictions. For example, the skew-normal model implies that different schools can

Figure 3: Histogram of SEVERITY in the treatment group

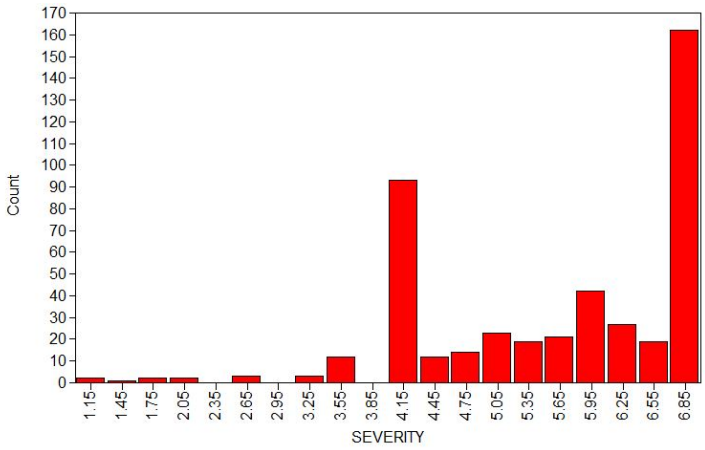


benefit differently from the intervention depending on what level of SEVERITY is observed before the treatment. The normal model implies that all schools benefit equally which clearly is not the case because a school with high level of SEVERITY is expected to benefit less than a school with low level of SEVERITY. The third advantage is that the MSE of the parameter estimates will be smaller due to more accurate model specification, i.e., the parameter estimates are more accurate with the skew-normal model.

4.3 Failure of robust ML estimation in linear regression models

In this section we illustrate with simulated data that the most basic linear regression model can yield inaccurate results simply from misspecifying the

Figure 4: Histogram of SEVERITY in the control group



distribution of the residual. Suppose that Y is a dependent variable and X is a covariate. We are interested in comparing the estimates for the simple linear regression model

$$Y = \alpha + \beta X + \varepsilon, \tag{80}$$

where ε is independent of X. This example is simpler than the example considered in the previous section because X is a true covariate, i.e., we are not concerned with modeling the distribution of X, but simply want to use X to provide the best prediction model for Y. Under the assumption of perfect linearity the ML estimates yield asymptotically unbiased results. The question is what happens when the perfect linearity is not present as this assumption is probably unrealistic in many situations. Approximate linearity is a much better assumption than assuming perfect linearity and is exactly what the skew-t modeling distribution uses. In this model, however, both the skew-t distribution model and the normal distribution model assume that $E(Y|X)$ is linear in terms of X. This again is because we do not model the distribution of X, we only model the distribution of

ε as a skew-t distribution and therefore linearity is assumed.

We generate two data sets. Both are of size 10000. The first data set is generated from a bivariate skew-t distribution where the residual variances are 1, the covariance is 0.5, the means are 0, the skew parameters are 3 and the df parameter is 3 as well. The second data set is generated also from a bivariate skew-t distribution but now the skew parameter is set to 0 for the X variable and the DF parameter is set to a large value, thus yielding a normal distribution for X and a skew-normal distribution for Y. The linearity of $E(Y|X)$ holds in the second data set but it does not hold in the first. It only holds approximately. Note again that neither of the above models, the skew-t or the normal regression model accommodate non-linearity in $E(Y|X)$, i.e., they are both wrong for the first data set but are correct for the second data set.

The results in the second data set are as follows. The log-likelihood for the skew-t model is -20609.632, the log-likelihood for the normal model -21104.013. Clearly the LRT test here would reject the normal model in favor of the skew-t model, however this is due only to modeling of the distribution of the residual. The coefficient β is estimated to 0.506 for the skew-t model and to 0.507 for the normal model. A formal LRT test to see if the skew-t model coefficient is different from the estimated value from the normal model yields a p-value of 0.92, i.e., there is no statistical evidence that the normal based model has a biased regression coefficient.

The results in the first data set are quite different. The log-likelihood for the skew-t model is -17557.873, the log-likelihood for the normal model -20061.821. Clearly again the LRT test would reject the normal model in favor of the skew-t model. However this is no longer due only to modeling of the distribution of

the residual. The coefficient β is estimated as 0.893 for the skew-t model and as 0.877 for the normal model. A formal LRT test to see if the skew-t model coefficient is different from the estimated value from the normal model yields a p-value of 0.0005, i.e., there is statistical evidence that the normal based model has a biased regression coefficient. The bias is due to the violation of the perfect linearity assumption. Neither of the two models is correct here. The skew-t model, however, is able to extract more information from the data and obtain more accurate estimates.

It is important to understand the limitations of the robustness of the ML estimation. It is also important to understand why the skew-t distribution yields different structural estimates from the normal distribution estimates. This example illuminates both points. It is not unusual to see similar differences also in real data sets. In the next section we illustrate this with one real data example.

4.4 The BMI mediation example

In the example described in Section 4.1 when the BMI17 variable is regressed on the mother's education predictor the normal model yields an estimate of $-0.388(0.063)$. When we analyze the effect of mother's education on BMI17, using mother's education as a multiple group variable, with the skew-t distribution the results show that there is no effect on the intercept and the degrees of freedom parameter while the skew and the variance parameter have a significant negative effect. Using formula (18) the overall effect on the mean is thus estimated to be $-0.444(0.066)$. In this example a linear effect based on the normal regression model is unreasonable because it would imply that higher mother's education not

only leads to reduced obesity problems but also that higher mother's education leads to an abnormally low range BMI associated with eating disorders. Instead, the model based on the skew-t distribution implies that higher mother's education leads to normal BMI range and reduced BMI variability.

If we incorporate the mother's education predictor in model (76) and consider the total, the direct, and the indirect effect from the mother's education predictor to BMI17 we reach substantively different conclusions using normal versus skew t-distributions. With the standard normality-based model the results indicate that all of the effect is an indirect effect. The direct effect is insignificant and virtually zero. With the skew-t based model which allows the predictor to affect the skew and the variance of the variables we obtain completely different results. Following the discussion in Section 3.7 we estimate the following model and the implied effects

$$BMI12 = \alpha_1 + \varepsilon_1, \tag{81}$$

$$BMI17 = \alpha_2 + \beta_1 BMI12 + \varepsilon_2, \tag{82}$$

where

$$\varepsilon_1 \sim rMST(0, a_1 + b_1 X, a_2 + b_2 X, \nu), \tag{83}$$

$$\varepsilon_2 \sim rMST(0, a_3 + b_3 X, a_4 + b_4 X, \nu), \tag{84}$$

where X represents the mother's education predictor. To compute $E[Y(x, M(x^*))]$ from formula (68) for BMI17 we note that this is the marginal mean of $rMST(\mu, \sigma, \delta, \nu)$ where

$$\mu = \alpha_2 + \beta_1 \alpha_1, \tag{85}$$

$$\sigma = \beta_1^2(a_1 + b_1 x^*) + a_3 + b_3 x, \tag{86}$$

$$\delta = \beta_1(a_2 + b_2 x^*) + a_4 + b_4 x. \tag{87}$$

Using formula (18) we obtain

$$E[Y(x, M(x^*))] = \alpha_2 + \beta_1\alpha_1 + \left(\beta_1(a_2 + b_2 x^*) + a_4 + b_4 x\right)\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}\sqrt{\frac{\nu}{\pi}}. \tag{88}$$

We can now compute the direct effect as

$$E[Y(x, M(x^*))] - E[Y(x^*, M(x^*))] = b_4(x - x^*)\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}\sqrt{\frac{\nu}{\pi}} \tag{89}$$

and the indirect effect as

$$E[Y(x, M(x))] - E[Y(x, M(x^*))] = \beta_1 b_2(x - x^*)\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}\sqrt{\frac{\nu}{\pi}}. \tag{90}$$

In the BMI example the above skew-t model results indicate that the direct effect is 85% of the total effect and the indirect effect is only 15% of total effect. This drastically different structural result illustrates the modeling opportunities when we look beyond the mean and variance modeling used with standard SEM.

4.5 Simulation study of a path analysis model with covariates

In this section we describe a simulation study using the following path analysis model

$$Y1 = \alpha_1 + \beta_1 Y2 + \beta_2 X + \varepsilon_1, \tag{91}$$

$$Y2 = \alpha_2 + \beta_3 X + \varepsilon_2. \tag{92}$$

We generate 100 samples of size 5000 and analyze the above model using the skew-t distribution and the normal distribution. We do not model the distribution of X, i.e., it is treated as a true covariate while the distribution of ε_1 and ε_2 are modeled as uncorrelated residuals. The covariate X is generated as a normally distributed variable with mean 0 and variance 1. The variables ε_1 and ε_2 are generated from a skew-t distribution. In this simulation study the conditional expectation $E(Y2|X)$ is linear in terms of X and $E(Y1|Y2, X)$ is linear in terms of X, but not linear in terms of Y_2. This last non-linearity violates the assumption of the standard regression model and thus we may expect to see biased estimates. On the other hand the skew-t distribution model is the same as the generating model and thus we should see unbiased estimates.

The true parameters of the skew t-distribution used for the generation are given in Table 2. Table 2 also contains the bias and the coverage for the parameter estimates when we analyze the model assuming the skew-t distribution. Clearly the estimates are unbiased and the coverage is near the nominal level of 95%. Because $E(Y2|X)$ and $E(Y1|Y2, X)$ are linear in terms of X, it is relevant to compare the skew-t and normal estimates for the β_2 parameter and for the β_3 parameter. The β_1 parameter estimates, however, are not comparable due to the non-linearity. When we analyze the data assuming normality we find that the parameter β_2 is severely biased. The ML estimate has a bias of -0.44. The estimate of β_3 is unbiased, the bias is 0, however the MSE of the ML estimate is almost twice as large as the estimate under the skew-t model. The MSE for β_3 is 0.060 using the ML normality assumption while it is only 0.032 under the

Table 2: Absolute bias and coverage for the skew T model

Parameter	True value	Bias(Coverage)
α_1	5	.01(.96)
α_2	0	.00(.95)
β_1	1	.01(.96)
β_2	0	.00(.97)
β_3	1	.00(.96)
δ_1	2	.03(.96)
δ_2	4	.00(.99)
DF	4	.01(.94)

skew-t assumption, i.e., even though the β_3 estimate is unbiased it is still much less accurate that under the more flexible skew-t model.

Note that in this example there were no violations of linearity for X. Both $E(Y2|X)$ and $E(Y1|Y2, X)$ are linear in terms of X. Despite that the normality based ML estimator, which is assumed to be robust for β_2 and β_3, did not perform well. The reason this happened is because of the non purely linear relationship between $Y1$ and $Y2$. This impurity creeps in to the effects of the X variable. Thus we conclude here that misspecification in one part of the model may affect seemingly unrelated parameters and variables. Again, the perfect linearity assumptions of the ML based normality estimator resulted in this poor performance.

In a model such as the one given by equations (91-92) to test for a significant effect we can still use the standard T-test for coefficients β_2 and β_3 since X is a covariate. For coefficient β_1 the T-test will work also as long as δ_1 or δ_2 is 0. If both skew parameters are non-zero then we have to use either equation (19), (22), or equation (26) depending on which distribution has been used to test that

the covariance between the two dependent variables is 0. However, even if the covariance is 0, $E(Y_1|Y_2)$ and $Var(Y_1|Y_2)$ may still depend on Y_2.

Another perspective on the non-linearity issue is as follows. We generate a bivariate sample of size 5000 using a multivariate skew t-distribution with the following parameters $\mu_i = 0$, $\sigma_{ii} = 5$, $\sigma_{12} = 2$, $\delta_i = 4$, $DF = 4$. We split the sample in 5 groups by the order of Y_1, i.e., we order the observations by the value of Y_1 and use the first 1000 observations with the lowest Y_1 to form group 1. The second group is formed by the next 1000 observations etc. We estimate a linear regression model of Y_2 on Y_1 in each of the 5 groups and in all 5 groups together. The results for the regression coefficient are given in Table 3. Clearly some of the differences in the 5 groups are not significant but some of them are. The relationship between Y_1 and Y_2 for the lowest values of Y_1 is not as strong as for the other groups and if we estimate the groups together not only will we miss this fact but for the rest of the observations the β coefficient will be underestimated because that single coefficient is averaging the relationship over all the observations and essentially will need to compensate the strength of the relationship from one group to another. If one indeed has a sample size of 5000 and the level of variation shown in Table 3 it is seems quite insufficient to attempt to describe the relationship between Y_1 and Y_2 with one coefficient.

4.6 Simulation study of a factor analysis model

In this section we present a factor analysis model simulation. The model has 1 factor η, 5 indicator variables Y_i, i=1,...,5; and one covariance. In this simulation we generate data using the skew-normal distribution. The model is given by these

Table 3: Nonlinearity in linear regression

Group	β
1	0.35(.05)
2	0.82(.15)
3	.93(.15)
4	.71(.12)
5	.89(.02)
all	.84(.02)

two equations

$$Y_i = \alpha_i + \lambda_i \eta + \varepsilon_i \tag{93}$$

$$\eta = \beta X + \xi. \tag{94}$$

To generate the data we use the following values for the parameters, $\alpha_i = 5$, $\lambda_i = 1$, $\beta = 1$, $\theta_i = Var(\varepsilon_i) = 5$, $\psi = Var(\xi) = 5$, the skew parameter for $\delta_\eta = 4$. We also generate the data with a skew parameter for Y_1, i.e., in this structural equation model the skewness of the data is not completely explained by the skewness of the factor η. For indicator variables $Y_2 - Y_5$ the skewness is explained by the skewness of the factor η but not for Y_1. The skewness parameter for Y_1 is $\delta_{Y_1} = 2$. We generate 100 data sets of size 5000. The data are analyzed with the same model that generates the data, i.e., with the factor analysis model and estimating the skew parameters for η and Y_1. The results for a subset of the parameter estimates are presented in Table 4. The bias in the parameter estimates is almost non-existent and the coverage is near the nominal 95% level. In addition when the model is tested against the saturated skew normal model the average chi-square statistic has an average value of 12.389 which matches the 12 degrees of

freedom for this test of fit. The chi-square rejected the factor analysis model 7% of the time which is sufficiently near the nominal level of 5% rejections. Thus the chi-square test concludes that this factor analysis model fits the data well. On the other hand when we analyze the data assuming a normal distribution using the MLR estimator and the robust chi-square test of fit we obtain an average statistic of 54.291 and 100% rejection rate. In addition the β coefficient estimate is biased. The average estimate across the 100 replications is 1.17, where the true value is 1. Both of these problems are entirely due to the additional/residual skewness of Y_1. If we generate the data without that residual skewness the MLR estimate for the β coefficient is unbiased, bias is zero and the coverage is 96%. In that case also the MLR chi-square test of fit rejects the factor analysis model only 4% of the time.

In the above skew-SEM model the latent variable η does not have a zero mean as in standard SEM, even if we eliminate the covariate from the model and the residual skewness. The mean of η has to be computed using formula (18). In the factor analysis model this fact is not very important. The shift in the mean is absorbed by the the intercepts of the observed indicator variables and if you compare that model to a standard SEM model you will find that the intercept parameters are different. This is just a constant shift and does not have any significance. In some SEM models where the means of the indicators are also structured via a longitudinal feature or a multiple group feature the implications of the non-zero factor mean should be considered carefully.

This again confirms our findings that the robust ML estimation can deal well with non-normality of an individual factor or a residual but it will not work well when more complicated relationships are found in the data. The concept

of residual skewness will be found in real data examples as long as the skewness parameters for the indicator variables in the saturated model are not proportional to the factor loadings of the factor model. Alternatively to check if residual skewness exist for a particular factor model we can estimate the model where all indicator variables have estimated skew parameters while the factor skew parameter is fixed to 0. If in that model the skewness parameters are proportional to the loading parameters we can safely assume that the factor model can explain all of the skewness in the data. If the skewness parameters are significantly not proportional to the loading parameters then residual skewness will exist and the factor alone will not be able to explain all the skewness in the data. Alternatively we can use the LRT test to compare the these two models. Let's call the factor analysis model with all δ_{Y_i} free the H1 model and the let's call the factor analysis model with all δ_η free the H0 model. Testing for residual skewness is equivalent to testing the H1 model against the H0 model. This can be done using the LRT test statistic

$$T = 2(LL_{H1} - LL_{H0}). \tag{95}$$

We illustrate this with a real data example using the Australian Institute of Sports Data described in Lin et al. (2013). We estimate a one-factor model on the subset of 102 males in the sample. The factor model has 11 indicator variables. The log-likelihood for the H1 model is -973.986. The log-likelihood for the H0 model is -1304.343. The chi-square test has 10 degrees of freedom and the test statistic value is 660.714. The test clearly rejects the hypothesis that the skewness in the observed data is due entirely to the skewness of the factor. A more detailed analysis should follow at this point to determine which indicator

Table 4: Absolute bias and coverage for the skew factor analysis model

Parameter	Bias(Coverage)
α_1	.03(.97)
λ_2	.00(.92)
θ_1	.03(.95)
β	.01(.94)
ψ	.08(.96)
δ_η	.01(.96)
δ_{Y_1}	.03(.93)

variables need to have residual skewness estimated. This ad hoc evaluation should be done to achieve these three goals: minimize the number of residual skewness needed, maximize the skewness explained by the factor, and still get a model that fits the data as well as the H1 model. This detailed analysis goes beyond the scope of this paper.

4.7 Missing data

It is well known that modeling with missing data via the FIML estimator is not robust to the normality assumptions. In fact early interest in the t-distribution was based on generalizing the EM-algorithm used to estimate sample means and variance in the presence of missing and non-normal data, see Liu and Rubin (1995). In this section we illustrate the effect of normality assumption violation with a simple simulated example. We generate a sample using the skew normal distribution with 5 variables and the following parameters: $\mu_i = 0$, $\sigma_{ii} = 1$, $\sigma_{ij} = 0.4$ for $i \neq j$, $\delta_i = 3$. We induce missing data for the first variable using the

following MAR missing data mechanism

$$P(Y_1 \text{ is missing}) = \frac{1}{1 + Exp(-1 + Y2 + Y3 + Y4 + Y5)}. \tag{96}$$

This is a simple logistic regression for the missing data indicator for Y_1 on the rest of the observed variables. We use one sample of size 100000 so that the estimates have no or minimal variation across samples and we can easily see the bias in the estimates. With this data generation, the true mean for Y_1 is $3\sqrt{2/\pi} \approx 2.4$. The standard FIML estimator assuming normality estimates the mean of Y_1 as 2.1. Using the correct distributional assumption and estimating the saturated skew-normal model we get the sample mean estimate to be 2.4. This example illustrates the heavy dependence on the normality assumption of the FIML estimator when there is missing data. Even as simple a value as the average for a variable can be misestimated. More advanced parameters such as structural parameters may be even more vulnerable.

4.8 Mixture models

With mixture models the situation is somewhat different when it comes to modeling with skew-t distributions. First, we do not need to be concerned with linearity and non-linearity of the relationships in the variables. The relationships are already non-linear because it is a mixture of models. Second, we can get the main benefit from the skew-t distribution already by simply allowing latent variables to have a skew distribution, and avoid complications arising from residual skewness. The main benefit of the skew-t distribution in mixture models is the ability to relax the within-class normality assumption for the observed variables

and to be able to accommodate skewed or heavy tails in the distributions. This is important because if the normality assumption is used, then classes will have to be formed to thicken the tail of the distribution if such tails are indeed observed. We use mixture models to discover latent subpopulations that have structurally different relationships between the variables. We are generally not interested in discovering latent classes that are formed simply to match the observed distribution curvature.

To illustrate this concept we reanalyze the BMI quadratic latent growth mixture model described in Nonnemaker et al. (2009), using the sample of black females; see also Muthén and Asparouhov (2014b). The sample size is 1160. BMI is observed at 12 time points spanning ages 12 to 23. The three latent variables, random intercept, random slope, and a quadratic term, are modeled either as normal variables or as skew-t variables. The residuals of the observed variables are assumed normal. We use the BIC criterion for model selection and to determine the number of latent classes. The BIC values are presented in Table 5 for the skew-t distribution model and the normally distributed model. We choose the number of classes for which BIC attains its minimum value. Using the normal distribution leads to four classes. If we use the the skew-t distribution for the latent growth factors, however, we conclude that there are only two classes. Thus the skew t-distribution helps eliminate two of the classes that may be spurious and lacking proper interpretation. In addition to that the 2-class skew-t model yields a better BIC than the 4-class normal model, i.e., not only did we eliminate potentially spurious classes but we actually found a better fitting model.

Table 5: BIC for BMI quadratic latent growth model

Classes	Normal	Skew-T
1	34168	31411
2	31684	**31225**
3	31386	31270
4	**31314**	
5	31338	

5 Conclusion

Liu and Rubin (1995) state that "Current computational advances for the multivariate t-distribution will make it routinely available in practice in the near future". Twenty years later it appears they were correct. Implementing the skew-t distribution for general structural equation models in Mplus will hopefully make this a reality. Applications are possible where structural equation models are built on more than just sample means and sample covariances. It is perhaps time to break out of a factor analysis and path analysis modeling framework invented about 100 years ago (see, e.g., Spearman, 1904; Wright, 1918, 1928) before the advent of computers and even calculators. The skew-t distribution is not the final word by any means but it is definitely a good step in that direction.

Structural equation models should be built to illuminate processes, structural pathways and relationships in the data and not simply to fit the means and the covariances of the variables. Fitting means and covariances should be viewed as an auxiliary goal. The main goal is to find a structural model that represents and acknowledges the data. If the means and the covariances are our only interest then simply using the sample values should be enough and no structural model would

be needed. As the examples in this article illustrate standard SEM models based only on means and covariances are limited when used for prediction and inference on the individual level. A model based on aggregate characteristics can be used only for prediction and inference on the aggregate characteristics. Individual level inference and implications are out of reach for the standard SEM models. A standard SEM model can fit the means and covariances well and still be completely detached from the data. The concept of robust estimation for standard SEM models gives a false comfort. It eliminates the normality assumption by replacing it by equally unrealistic assumptions of pure linearity, pure independence between predictors and residuals, homoscedasticity of residual variables, and homogeneity in the relationships of the variables in the entire scope of the distributions. The skew-SEM framework can be used to challenge the conclusions obtained with standard SEM models and to enhance and illuminate our understanding of the data.

While univariate skewness and kurtosis are easy to visualize, comprehend and include in a model, the multivariate deviations from normality are more intricate to test and model. The modeling framework presented here based on the skew-t distribution is one possible option which may or may not be appropriate for a particular application. We can use BIC as a guide to the best fitting model, however, BIC will lead to correct results only if the true model is in consideration. If all the models we consider, skew-SEM and standard SEM, are inadequate then BIC can be misleading. More real data applications are needed to truly evaluate the implications of the skew-SEM framework and to guide in future development and extensions. The examples described in this article indicate that extending the framework to allow covariates to have an effect not just on the mean parameters

but also on the skew parameters would be very useful. Another useful extension would be to incorporate the unrestricted skew-t distributions into a general SEM model.

References

Arellano-Valle, R. B., and Genton, M. G. (2010). Multivariate extended skew-t distributions and related families. Metron-International Journal of Statistics, 68, 201-234.

Bauer, D.J., & Curran, P.J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. Psychological Methods, 8, 338-363.

Boik R. J. and Robison-Cox, J. F. (1998). Derivatives of the incomplete Beta function. Journal of Statistical Software, 3, 1-19.

Lee S., McLachlan G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. Journal of Statistics and Computing, 24, 181-202.

Lin T., Wu P. H., McLachlan G. J., Lee S. X. (2013). The skew-t factor analysis model. Submitted. http://arxiv.org/abs/1310.5336

Liu, C., Rubin, D.B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statistica Sinica, 5, 19-39.

MacKinnon, D.P. (2008). Introduction to statistical mediation analysis. New York, NY: Lawrence Erlbaum Associates.

Muthén, B. & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics, 55, 463-469.

Muthén, B. & Asparouhov, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (eds.), Longitudinal Data Analysis, pp. 143-165. Boca Raton: Chapman & Hall/CRC Press.

Muthén, B. & Asparouhov, T. (2014a). Causal effects in mediation modeling: An introduction with applications to latent variables. Forthcoming in Structural Equation Modeling.

Muthén, B. & Asparouhov, T. (2014b). Growth mixture modeling with non-normal distributions. Submitted for publication.

Nonnemaker JM, Morgan-Lopez AA, Pais JM, Finkelstein EA. (2009). Youth BMI trajectories: evidence from the NLSY97. Obesity, 17, 1274-1280.

Satorra, A. (2002). Asymptotic robustness in multiple group linear-latent variable models. Econometric Theory, 18, 297-312.

Schork, N.J. & Schork, M.A. (1988). Skewness and mixtures of normal distributions. Communications Statistics Theory Methods 17, 3951-3969.

Spearman, C. (1904). General intelligence, objectively determined and measured. American Journal of Psychology, 15, 201-293.

VanderWeele T.J. & Vansteelandt S. (2009). Conceptual issues concerning mediation, interventions and composition. Statistics and Its Interface, 2, 457-468.

Wright, S. (1918). On the nature of size factors. Genetics, 3, 367-374.
Wright, S. (1934). The method of path coefficients. Annals of Mathematical Statistics, 5, 161-215.