

# Non-Normality Propagation among Latent Variables and Indicators in PLS-SEM Simulations

Ned Kock

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

---

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized administrator of DigitalCommons@WayneState.

# Non-Normality Propagation among Latent Variables and Indicators in PLS-SEM Simulations

**Ned Kock**

Texas A & M International University  
Laredo, Texas

---

Structural equation modeling employing the partial least squares method (PLS-SEM) has been extensively used in business research. Often the use of this method is justified based on claims about its unique performance with small samples and non-normal data, which call for performance analyses. How normal and non-normal data are created for the performance analyses are examined. A method is proposed for the generation of data for exogenous latent variables and errors directly, from which data for endogenous latent variables and indicators are subsequently obtained based on model parameters. The emphasis is on the issue of non-normality propagation among latent variables and indicators, showing that this propagation can be severely impaired if certain steps are not taken. A key step is inducing non-normality in structural and indicator errors, in addition to exogenous latent variables. Illustrations of the method and its steps are provided through simulations based on a simple model of the effect of e-collaboration technology use on job performance.

*Keywords:* E-Collaboration; Partial Least Squares; Latent Variable; Indicator; Non-Normal Data; Monte Carlo Simulation

---

## Introduction

Structural equation modeling (SEM) employing the partial least squares (PLS) method, or PLS-SEM for short, has been extensively used in business research (Hair, Ringle, & Sarstedt, 2011; Kock, 2010; 2014). It has also been increasingly used in a wide variety of fields; some closely related to business, including subfields, and others less so. Examples are information systems (Guo, Yuan, Archer, & Connelly, 2011; Kock & Lynn, 2012), marketing (Biong & Ulvnes, 2011), international business (Ketkar, Kock, Parente, & Vervielle, 2012), nursing

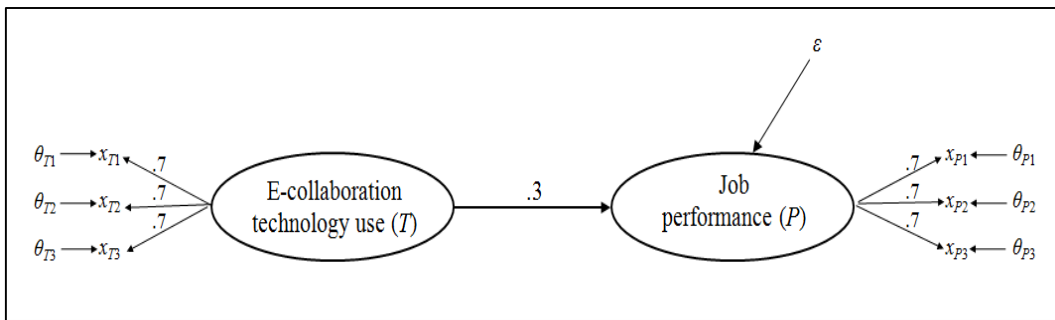
---

*Ned Kock is a Distinguished Professor and Chair of the Division of International Business and Technology Studies. Email him at: nedkock@gmail.com.*

## NON-NORMALITY PROPAGATION IN PLS-SEM

(Kim et al., 2012), medicine (Berglund, Lytsy & Westerling, 2012), and global environmental change (Brewer, Cinner, Fisher, Green, & Wilson, 2012).

One of the elements that characterize the PLS-SEM method is that it creates latent variables (sometimes referred to as latent “composites”) by means of weighted aggregations of their respective indicators, where the weights are obtained through iterative algorithms (Cirillo & Barroso, 2012; Lohmöller, 1989). The simple model shown in Figure 1 illustrates the main elements of any model used in PLS-SEM.



**Figure 1.** Structural model with two latent variables

\*Notes: latent variables within ovals; loadings next to indicator arrows.

Our simple model follows from past empirical research (Cassivi, Lefebvre, Lefebvre, & Léger, 2004; Chen, Chen, & Capistrano, 2013). It contains two latent variables, e-collaboration technology use (*T*) and job performance (*P*), which are measured indirectly through three indicators each. The unit of analysis is assumed to be a team of individuals who collaborate to accomplish work-related tasks in their respective organizations. E-collaboration technology use (*T*) measures the extent to which a team uses an integrated technology including e-mail and voice conferencing to facilitate the collaborative work of its members. Job performance (*P*) measures the performance of each team, as perceived by individuals who receive the outputs of the team to perform downstream work-related tasks.

The structural error  $\varepsilon$ , when properly weighted, accounts for the variance in the latent variable job performance (*P*) that is not explained by e-collaboration technology use (*T*). For e-collaboration technology use (*T*) the indicators are  $x_{T1}$ ,  $x_{T2}$  .and  $x_{T3}$  . For job performance (*P*) the indicators are  $x_{P1}$ ,  $x_{P2}$  .and  $x_{P3}$  . When properly weighted, the uncorrelated indicator errors  $\theta_{T1}$ ,  $\theta_{T2}$ ,  $\theta_{T3}$ ,  $\theta_{P1}$ ,  $\theta_{P2}$

and  $\theta_{p_3}$  account for the variances in the indicators that are not explained by their corresponding latent variables.

The indicators store answers to question-statements in a questionnaire. The question-statements are redundant with one another, with respect to each latent variable, and are assumed to “reflect” the latent variable. That is, the indicators are assumed to measure only the latent variable to which they refer. This measurement carries a certain amount of imprecision, which is indicated by the loadings being lower than 1. This implies the existence of measurement error, which would be absent if at least one loading were to be equal to 1.

Because PLS-SEM algorithms are generally claimed to perform particularly well with small samples and non-normal data (Hair et al., 2011), it is necessary to test that claim by comparing the performance of a PLS-SEM algorithm, such as PLS regression (Kock, 2010), in terms of statistical power, against the performance of a non-PLS algorithm. A common choice of “control” non-PLS algorithm is one where indicators are aggregated to generate latent variable scores in a non-weighted fashion; i.e., indicators are aggregated using the same weight.

Performance analyses usually build on Monte Carlo simulations (Robert & Casella, 2005) whereby multiple samples are created and analyzed using the algorithms that are being compared. The samples are created based on true population coefficients. In this case, these are the standardized regression coefficient ( $\beta = .3$ ) and the loadings ( $\lambda_{Ti} = \lambda_{Pi} = .7, i = 1 \dots 3$ ), which are assumed to exist in the population from which the samples are taken. Both the standardized regression coefficient and the loadings are set by the researcher conducting the Monte Carlo simulations.

We address the issue of how one creates normal and non-normal data for such performance analyses. A simple and effective method is proposed for creating data for exogenous latent variables and errors directly, from which data for endogenous latent variables and indicators is subsequently derived. This method is similar to that proposed by Mattson (1997), incorporating elements that arguably make it simpler.

The discussion of the method places emphasis on the issue of non-normality propagation among latent variables and indicators in PLS-SEM simulations, showing that this propagation can be severely impaired if certain steps are not taken. A key step is to induce non-normality in structural and indicator errors, in addition to exogenous latent variables. This is illustrated through Monte Carlo simulations.

### A Method for Creating Normal and Non-Normal Data

Several methods exist to create normal and non-normal data for simulations (Headrick, 2010). Power methods relying on polynomial transformations are perhaps the most widely used (Fleishman, 1978; Headrick, 2002). A special case relies on squaring a standardized normal variable  $X$  to obtain a non-normal variable  $X_n$  as shown in (1) and (2). In these equations  $Rndn(N)$  is a function that returns a different normal random variable each time it is invoked, in the form of a vector with  $N$  elements, and  $Stdz(\cdot)$  is a function that returns a standardized variable.

$$X \leftarrow Stdz(Rndn(N)) \quad (1)$$

$$X_n \leftarrow Stdz(X^2) \quad (2)$$

This method of creating non-normal data has the advantages of introducing enough non-normality to be useful in robustness tests, and at the same time yielding data that follows a  $\chi^2$  distribution with 1 degree of freedom. A number of properties are known for this distribution, including probability limit skewness and kurtosis (a.k.a. excess kurtosis) values. These are  $\sqrt{8} \cong 2.828$  and 12, respectively, which combined can be seen as indications of severe non-normality.

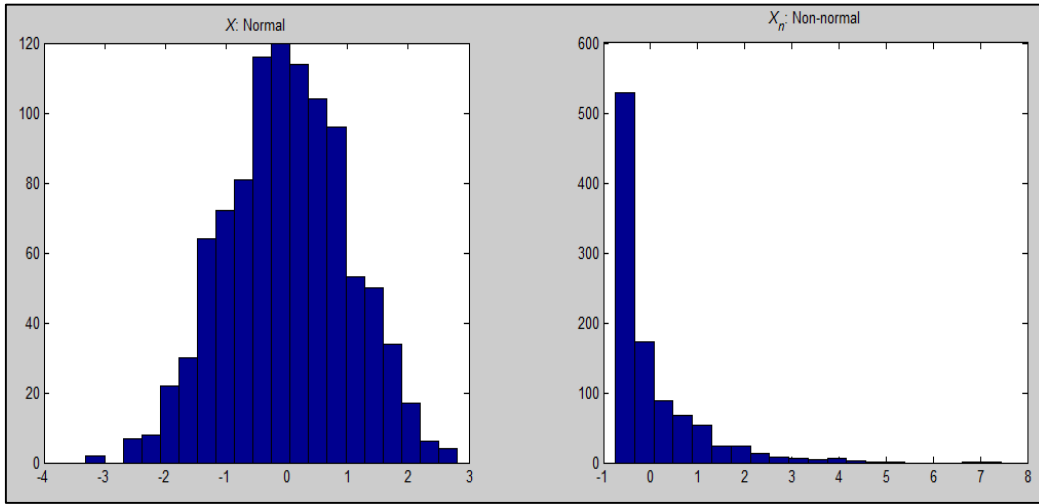
Figure 2 shows two histograms. The one on the left is for a normally distributed variable  $X$  created based on (1) with  $N = 1,000$ . The one on the right shows a variable  $X_n$  that follows a non-normal distribution created based on (2), applied to the normally distributed variable  $X$ . Both variables  $X$  and  $X_n$  are standardized.

Data generated through this method, as well as variations discussed here, is initially standardized. Unstandardization can be easily accomplished by multiplying by  $\sigma$  and adding  $\mu$ , where  $\sigma$  and  $\mu$  are the standard deviation and mean of the desired unstandardized distribution, respectively. Rounding to the closest integer within an ordinal scale (e.g., 1...7) yields unstandardized data on a Likert-type scale.

Not only does the non-normal variable  $X_n$  present significant positive skewness (i.e., longer tail on the right) and positive kurtosis (i.e., leptokurtosis, or “peakedness”), but it also contains more extreme outliers than  $X$ . As noted in other graphs, this is a common feature of non-normal data created through this method. This makes it useful in robustness stress tests; where claimed robustness

in the presence of non-normality is tested under non-normality conditions that are more extreme than usually found in empirical data.

The univariate method described above can be easily extended to the multivariate case. Multiple exogenous latent variables and errors (i.e., error variables) can be created in the same general way, and non-normality can be propagated from exogenous latent variables and structural errors to endogenous latent variables and indicators. This is discussed in the following sections.



**Figure 2.** Transforming normal into non-normal data

\* Notes: both variables  $X$  and  $X_n$  are standardized;  $X$  follows a normal distribution;  $X_n$  follows a  $\chi^2$  distribution with 1 degree of freedom;  $X_n$  was created based on  $X$ .

Data with less severe non-normality can be created using the same general method, by increasing the number of degrees of freedom of the  $\chi^2$  distribution used. This can be carried out by adding more than one squared standardized normal variable to generate the non-normal variable, as indicated in (3) and (4).

$$X_i \leftarrow Stdz(Rndn(N)) \tag{3}$$

$$X_n \leftarrow Stdz\left(\sum_{i=1}^k X_i^2\right) \tag{4}$$

## NON-NORMALITY PROPAGATION IN PLS-SEM

The number  $k$  of standardized normal variables  $X_i (i = 1 \dots k)$  used to generate the non-normal variable  $X_n$  equals the number of degrees of freedom of the resulting  $\chi^2$  distribution. The probability limit skewness and kurtosis of such a distribution are given by  $\sqrt{8/k}$  and  $12/k$ , respectively. Therefore, we can create data with varying degrees of skewness and kurtosis using various values of  $k$  through this generalized version of the method.

For example, if  $k = 3$  the non-normal variable  $X_n$  will have the following probability limit values for skewness and kurtosis:  $\sqrt{8/3} \cong 1.633$  and  $12/3 = 4$ , respectively. Data created with these distributional properties could be used in a robustness test for an intermediated condition that could be referred to as one with “moderate” non-normality, and whose results might be contrasted with those for two other conditions: normal, where  $Rndn(N)$  would be used with no transformation; and severely non-normal, where a transformation with  $k = 1$  would be used.

### Creating Normal and Non-Normal Data for Latent Variables

The method is illustrated based on the simple model presented earlier, which contains only two latent variables, and applies to more complex models, with any number of latent variables. In all cases, latent variables and structural errors are created first, and indicators and corresponding errors are created afterwards.

In this model, the normal data for the exogenous latent variable e-collaboration technology use ( $T$ ) is created according to (5). This is the predictor latent variable in the model. The non-normal data for this same latent variable ( $T_n$ ) is created according to (6). Analogously, the normal data for the structural error  $\varepsilon$  is created according to (7). The corresponding non-normal data for the structural error ( $\varepsilon_n$ ) is created according to (8).

$$T \leftarrow Stdz(Rndn(N)) \quad (5)$$

$$T_n \leftarrow Stdz(T^2) \quad (6)$$

$$\varepsilon \leftarrow Stdz(Rndn(N)) \quad (7)$$

$$\varepsilon_n \leftarrow Stdz(\varepsilon^2) \quad (8)$$

Both  $T$  and  $\varepsilon$  have probability limit values of 0 and 0 for skewness and kurtosis, respectively. Conversely, the non-normal variables  $T_n$  and  $\varepsilon_n$  have both probability limit values of  $\sqrt{8} \cong 2.828$  and 12 for skewness and kurtosis, respectively. As discussed earlier, these values refer to a  $\chi^2$  distribution with 1 degree of freedom.

The normal data for the endogenous latent variable job performance ( $P$ ) is created according to (9). This is the criterion latent variable in the model. The non-normal data associated with this latent variable can either propagate exclusively from  $T_n$  according to (10), or from both  $T_n$  and  $\varepsilon_n$  according to (11). As will become clear, the latter approach, using (11), is the most advisable of the two. In these equations the structural errors are properly weighted (i.e., given the weight  $\sqrt{1-\beta^2}$ ) to account for the variance in  $P$  that is not explained by  $T$ .

$$P = \beta T + \sqrt{1-\beta^2} \varepsilon \quad (9)$$

$$\dot{P}_n = \beta T_n + \sqrt{1-\beta^2} \varepsilon \quad (10)$$

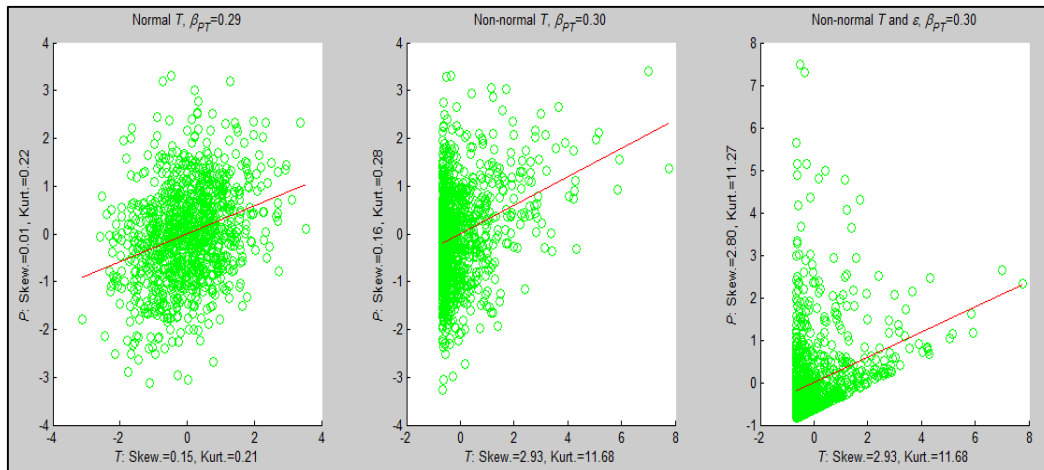
$$P_n = \beta T_n + \sqrt{1-\beta^2} \varepsilon_n \quad (11)$$

Figure 3 shows data points and regression lines for three samples, where the predictor latent variable is plotted on the horizontal axis and the criterion latent variable on the vertical axis, and in which: (left) both the predictor latent variable, e-collaboration technology use ( $T$ ), and structural error are normal ( $\varepsilon$ ); (middle) the predictor is non-normal ( $T_n$ ) but the error is normal ( $\varepsilon$ ); and (right) both the predictor and error are non-normal ( $T_n$  and  $\varepsilon_n$ , respectively). The sample sizes are 1,000 for the three samples. The data was created based on the foregoing equations, with  $\beta = .3$  as specified in our model.

At the top of the graphs are the true sample values of the standardized regression coefficients for each case. Their values are relatively stable across graphs, and close or identical to the true population value ( $\beta = .3$ ) implying robustness in the presence of severe non-normality and outliers. The robustness observed is a characteristic of regression methods in general (Haas & Scheff, 1990; Knez & Ready, 1997), and is one of the reasons why PLS-SEM is also a robust method. PLS-SEM builds heavily on regression methods.



## NON-NORMALITY PROPAGATION IN PLS-SEM



**Figure 3.** Normal and non-normal data for latent variables

\* Notes: scales are standardized; left - predictor latent variable and error are normal; middle - predictor is non-normal but error is normal; right - predictor and error are non-normal.

It should be emphasized that these standardized regression coefficients are not calculated based on the indicators. They are calculated directly based on the latent variable scores, which are available in the simulation method we describe here. Therefore, these true sample standardized regression coefficients are not distorted by measurement error. This is a type of error discussed earlier, whose existence is implied by the loadings being lower than 1.

As can be inferred from the graphs, when the predictor latent variable is non-normal but the error is normal (middle), the propagation of non-normality from the predictor latent variable  $T_n$  to the criterion latent variable job performance  $\hat{P}_n$  is severely impaired. In this case, while skewness and kurtosis for  $T_n$  are 2.93 and 11.68 respectively, the criterion latent variable  $\hat{P}_n$  is essentially normal (skewness = .16, kurtosis = .28).

Using this approach to create non-normal data in Monte Carlo simulations to test a PLS-SEM algorithm would lead to results supporting the conclusion that the algorithm is robust to non-normality when that may not be the case. In other words, if non-normality propagation is severely impaired, robustness tests would be largely meaningless, and may lead to incorrect conclusions.

However, if the approach associated with the graph at the far right is used, where both the predictor and error are non-normal (right), the propagations of non-normality from the predictor latent variable  $T_n$  and error  $\varepsilon_n$  to the criterion latent variable  $P_n$  is largely unimpaired. Here the same values of skewness and

kurtosis for  $T_n$  lead to 2.80 and 11.27 for  $P_n$ , because a large amount of the non-normality comes from the non-normal error  $\varepsilon_n$ .

Why is the propagation so severely impaired when the predictor latent variable is non-normal ( $T_n$ ) but the error is normal ( $\varepsilon$ )? As it will be clear from our discussion of non-normality propagation from latent variables to indicators, the reason is the magnitude of the propagation coefficient that links the latent variables.

In this case, this propagation coefficient is the standardized regression coefficient  $\beta$ , whose value is .3 in the model. This value is small compared with the propagation coefficient for the error ( $\sqrt{1-\beta^2} = \sqrt{1-.3^2} \cong .954$ ). Small propagation coefficients tend to impair non-normality propagation.

Small propagation coefficients are likely to be commonly found in PLS-SEM models, because standardized partial and full regression coefficients tend to be relatively small (or small enough to impair propagation) in models that are free from vertical and lateral collinearity (Kock & Lynn, 2012). The same applies to path models in general, with or without latent variables, and multiple regression models.

### Creating Normal and Non-Normal Data for Indicators

Consider the creation of normal and non-normal data for indicators by creating normal and non-normal data for each of the six indicator errors, expressed generally as  $\theta_{T_i}, \theta_{P_i}, \theta_{T_n}$ , and  $\theta_{P_n}$  ( $i = 1 \dots 3$ ).

The normal data for the indicators associated with the exogenous latent variable e-collaboration technology use ( $T$ ) and the endogenous latent variable job performance ( $P$ ) are created according to (12) and (13), respectively.

$$x_{T_i} = \lambda_{T_i} T + \sqrt{1 - \lambda_{T_i}^2} \theta_{T_i} \tag{12}$$

$$x_{P_i} = \lambda_{P_i} P + \sqrt{1 - \lambda_{P_i}^2} \theta_{P_i} \tag{13}$$

Analogously, the non-normal data for the indicators associated with the non-normal versions of the same latent variables, the exogenous latent variable e-collaboration technology use ( $T_n$ ) and the endogenous latent variable job performance ( $P_n$ ), are created according to (14) and (15), respectively.

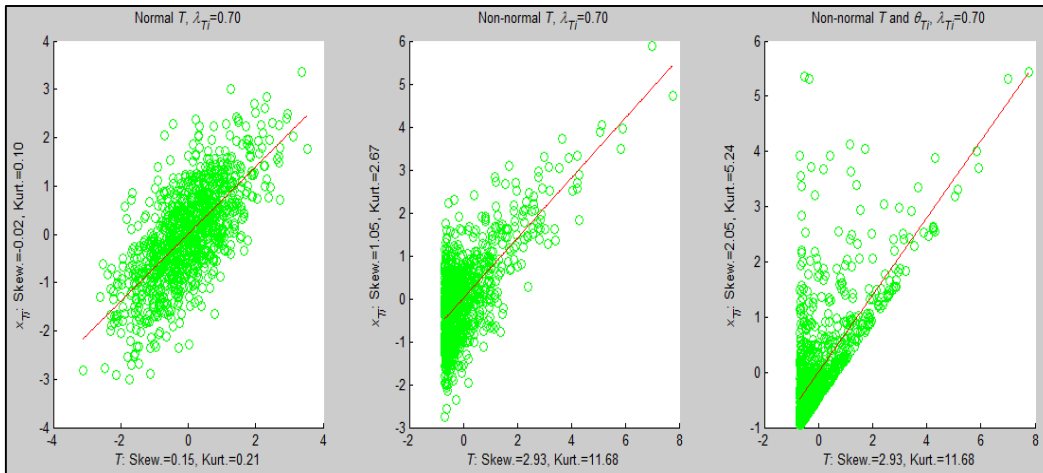
## NON-NORMALITY PROPAGATION IN PLS-SEM

$$x_{T_i} = \lambda_{T_i} T_n + \sqrt{1 - \lambda_{T_i}^2} \theta_{T_i} \quad (14)$$

$$x_{P_i} = \lambda_{P_i} P_n + \sqrt{1 - \lambda_{P_i}^2} \theta_{P_i} \quad (15)$$

Unlike the structural error weight, used in the creation of the endogenous latent variable, the weights of the indicator errors will tend to have magnitudes that are similar to the magnitudes of the loadings. In some cases, where measurement precision is high (i.e., high loadings), the weights of the indicator errors will be significantly lower than those of the indicator errors.

For example, a loading of .7 will lead to an indicator error weight of  $\sqrt{1 - .7^2} \cong .714$ , whereas a loading of .9 will lead to an indicator error weight of  $\sqrt{1 - .9^2} \cong .436$ . In the former case, the degree of non-normality propagation, measured through the corresponding coefficients of propagation (loading of .7 and weight of .714), will be about the same from the latent variable and the indicator error. In the latter case, the degree of non-normality propagation from the latent variable (loading of .9) will be much greater than from the indicator error (weight of .436).



**Figure 4.** Normal and non-normal data for indicators

\* Notes: scales are standardized; latent variable -  $T$ ; indicator -  $x_{T_i}$ ; left - latent variable and indicator error are normal; middle - latent variable is non-normal but error is normal; right - latent variable and error are non-normal.

Figure 4 shows data points and regression lines for three samples, where the latent variable is plotted on the horizontal axis and the indicator on the vertical axis, and in which: (left) both the latent variable and the indicator error are normal; (middle) the latent variable is non-normal but the indicator error is normal; and (right) both the latent variable and the indicator error are non-normal. As with the graphs for latent variables, the sample sizes here are 1,000 for the three samples. The data were created based on the foregoing equations with the loadings as specified in our model.

Data for only one latent variable and one indicator are used in these graphs. These variables serve as an illustration of what would happen with any pair of latent variable and corresponding indicator in our model. At the top of the graphs are the true sample values of the loadings for each case.

Non-normality propagation is different for the cases in which the latent variable is non-normal but the indicator error is normal (middle) and both the latent variable and the indicator error are non-normal (right). In the former case, skewness and kurtosis for the latent variable are 2.93 and 11.68 respectively, and 1.05 and 2.67 for the indicator. In the latter case, the same values of skewness and kurtosis for the latent variable lead to 2.05 and 5.24 for the indicator. In neither case non-normality propagates fully; both are examples of partial propagation.

These results bring to the fore two interesting characteristics of non-normality propagation. One is that there is always some loss in the propagation among linked variables; be the propagation among latent variables, or among latent variables and indicators. The other interesting characteristic of non-normality propagation is that the magnitude of the loss is strongly dependent on the propagation coefficients (path coefficients, loadings, and error weights), with the loss increasing steeply in response to decreases in those coefficients.

From these results it seems that this problem is more pronounced in the non-normality propagation from latent variables to indicators, as long as non-normal errors are used – otherwise propagation losses are greater among linked latent variables, because path coefficients tend to be generally lower in magnitude than loadings.

It could be argued that this loss in propagation is not a characteristic of the non-normal data creation method used, but stems from assumptions underlying the common factor model (MacCallum & Tucker, 1991). In it, the propagation of variance (and thus non-normality) happens only from latent variables to indicators, via loadings, and not the other way around.

Skewness and kurtosis values are not usually found in empirical data as extreme as those created. In empirical data, non-normality is often found, but of a

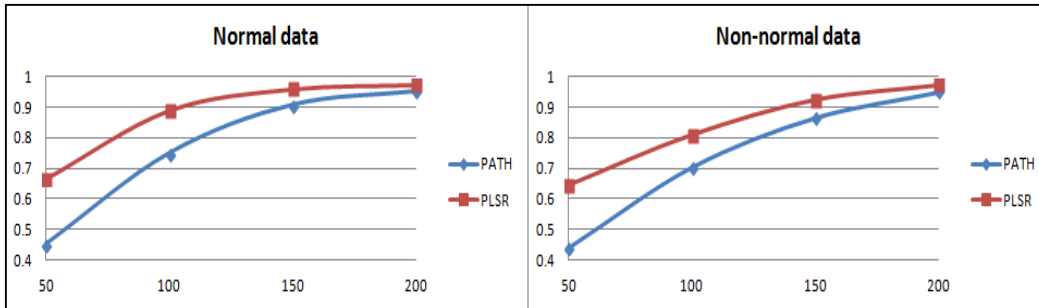
less extreme nature. Therefore, it is possible that the loss in non-normality propagation that we see in our analyses reflects a corresponding phenomenon in actual populations.

### Monte Carlo Simulation Results

The results of a set of Monte Carlo simulations are shown in Figure 5 where the performance of a relatively new and increasingly popular PLS-SEM algorithm, namely PLS regression (Kock, 2010), is shown against a control non-PLS algorithm in the context of our simple model. We used parametric path analysis as the control non-PLS algorithm. WarpPLS version 4.0, was used to analyze the data in our Monte Carlo simulations. The focus of our performance analysis is on statistical power, which is the probability of avoiding false negatives. We created and analyzed 500 samples (or replications) with normal and severely non-normal data. The data were created using the method described in the preceding sections, for each of the following sample sizes: 50, 100, 150, and 200.

The  $p$ -value calculation method used for PLS regression is the stable method (Kock, 2013). This heuristic method employs a built-in table of standard errors generated through bootstrapping and jackknifing (Chiquoine & Hjalmarsson, 2008; Diaconis & Efron, 1983; Efron et al., 2001), but instead of generating resamples it obtains standard errors based on nonlinear fitting using the built-in table. This significantly increases computational efficiency, particularly when large samples are used. In the parametric path analysis algorithm, which is our “control” non-PLS algorithm, indicators are aggregated to generate latent variable scores using the same weight of 1 for all indicators. The  $p$ -value calculation method used for parametric path analysis is the “parametric” method (Kock, 2013). This method calculates standard errors based on a Student’s  $t$ -distribution.

Skewness and excess kurtosis were calculated, and normality tested, for all indicators in each of the generated samples. This was done with the goal of ensuring that, with non-normal data, sample non-normality propagation to indicators occurred to the extent that all indicators followed truly non-normal distributions. Two tests of normality were used, each taking as inputs skewness and excess kurtosis values: the classic Jarque-Bera test (Jarque & Bera, 1980; Bera & Jarque, 1981) and Gel and Gastwirth’s (2008) robust modification of this test. Both tests, when applied to non-normal data, indicated statistically significantly non-normality in all indicators.



**Figure 5.** Monte Carlo simulation results

\* Notes: vertical axis - statistical power values (probabilities of avoiding false negatives); horizontal axis - sample sizes; PLSR = PLS regression; PATH = parametric path analysis.

As we can see from the results, PLS regression performed better in terms of statistical power than parametric path analysis with both normal and non-normal data, particularly so with small sample sizes. For example, PLS regression reached the widely accepted power threshold of .8 (yielding false negatives 20 percent of the time) with a sample size of approximately 75 with normal data, and with a slightly greater sample size with non-normal data.

Overall both algorithms suffered small performance losses with non-normal data, compared with their performance with normal data. The fact that those losses were small suggests that both algorithms are fairly robust to deviations from normality. This is not surprising because regression techniques in general and related *p*-value calculation methods are generally believed to be remarkably robust to deviations from normality (Haas & Scheff, 1990; Knez & Ready, 1997). PLS-SEM builds heavily on those techniques and methods.

As a side note, we should clarify that the PLS regression algorithm is referred to as “new” in the context of PLS-SEM because it has been more commonly used in the past in chemometrics applications (Wold et al., 2001) not involving PLS-SEM per se. The use of this algorithm in PLS-SEM is growing. It appears to offer some advantages over other PLS algorithms. One of the advantages is the de-coupling of the estimation of coefficients for the structural and measurement models (Kock, 2010), reducing the likelihood of capitalization on error. The advantages tend to become particularly clear when PLS regression is compared with the more widely used PLS mode A (Lohmöller, 1989) in PLS-SEM applications.

## Conclusion

A simple and effective method was proposed for the creation of non-normal data that follows a  $\chi^2$  distribution with 1 degree of freedom. This gives access to a number of properties, as this is a well known distribution, including probability limit skewness and kurtosis (a.k.a. excess kurtosis) values. These are  $\sqrt{8} \cong 2.828$  and 12, respectively, which reflect severe non-normality and are thus useful in robustness tests. It was shown how less severely non-normal data can be generated using the same general approach, by increasing the degrees of freedom of the  $\chi^2$  distribution used.

It was shown that proper propagation of non-normality requires the use of non-normal latent variables and errors, which can be created through the same  $\chi^2$  distribution approach. It was demonstrated that propagation of non-normality is severely impaired when propagating non-normal latent variables are used in combination with normal errors, and thus that it is important to use errors that are also non-normal. This applies to both structural errors and indicator errors.

Simulation researchers may be tempted to rescale the indicators directly to obtain non-normal data for use in PLS-SEM and other SEM simulations, since the indicators form the “raw material” that is used to compare different SEM techniques. The problem with this approach is that it removes the interdependence between latent variables and indicators, which in turn prevents true sample analyses and comparisons.

The method discussed here generates data for latent variables and errors directly, and then for indicators, preserving that interdependence. It gives full control of the samples, and the ability to calculate a variety of true sample coefficients that are not available from the specified true population model. In fact, this method permits creation of very large samples (e.g., with  $N = 10^6$ ), from which various traits of the population can be ascertained. In samples this large sampling error is very small, and thus coefficients tend to very similar to those found in the population from which samples are taken. Although the parameterized population model used to create data in simulations allows the true population path coefficients and loadings to be known, it does not inform the shape of the relationship between loadings and weights or the degree of collinearity among latent variables.

The former, the shape of the relationship between loadings and weights, could help us develop better PLS algorithms (Kock, 2010), with unbiased loadings and weights (Cassel et al., 1999). The latter, the degree of collinearity

among latent variables, could help understand the impact that PLS algorithms have on full collinearity variance inflation factors (Kock & Lynn, 2012).

## References

- Bera, A. K., & Jarque, C. M. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters*, 7(4), 313-318. doi:10.1016/0165-1765(81)90035-5
- Berglund, E., Lytsy, P., & Westerling, R. (2013). Adherence to and beliefs in lipid-lowering medical treatments: A structural equation modeling approach including the necessity-concern framework. *Patient Education and Counseling*, 91(1), 105-112. doi:10.1016/j.pec.2012.11.001
- Biong, H., & Ulvnes, A. M. (2011). If the supplier's human capital walks away, where would the customer go? *Journal of Business-to-Business Marketing*, 18(3), 223-252. doi:10.1080/1051712X.2011.541375
- Brewer, T. D., Cinner, J. E., Fisher, R., Green, A., & Wilson, S. K. (2012). Market access, population density, and socioeconomic development explain diversity and functional group biomass of coral reef fish assemblages. *Global Environmental Change*, 22(2), 399-406. doi:10.1016/j.gloenvcha.2012.01.006
- Cassel, C., Hackl, P., & Westlund, A. H. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics*, 26(4), 435-446. doi:10.1080/02664769922322
- Cassivi, L., Lefebvre, E., Lefebvre, L. A., & Léger, P. M. (2004). The impact of e-collaboration tools on firms' performance. *The International Journal of Logistics Management*, 15(1), 91-110. doi:10.1108/09574090410700257
- Chen, J. V., Chen, Y., & Capistrano, E. P. S. (2013). Process quality and collaboration quality on B2B e-commerce. *Industrial Management & Data Systems*, 113(6), 908-926. doi:10.1108/IMDS-10-2012-0368
- Chiquoine, B., & Hjalmarsson, E. (2008). *Jackknifing stock return predictions*. Washington, DC: Federal Reserve Board.
- Cirillo, M. A., & Barroso, L. P. (2012). Robust regression estimates in the prediction of latent variables in structural equation models. *Journal of Modern Applied Statistical Methods*, 11(1), 42-53. Available at: <http://digitalcommons.wayne.edu/jmasm/vol11/iss1/4>
- Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 249(1), 116-130.



## NON-NORMALITY PROPAGATION IN PLS-SEM

Efron, B., Rogosa, D., & Tibshirani, R. (2001). Resampling methods of estimation. In N. J. Smelser, & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 13216-13220). New York, NY: Elsevier.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4), 521-532. doi:10.1007/BF02293811

Gel, Y. R., & Gastwirth, J. L. (2008). A robust modification of the Jarque-Bera test of normality. *Economics Letters*, 99(1), 30-32. doi:10.1016/j.econlet.2007.05.022

Guo, K. H., Yuan, Y., Archer, N. P., & Connelly, C. E. (2011). Understanding nonmalicious security violations in the workplace: A composite behavior model. *Journal of Management Information Systems*, 28(2), 203-236. doi:10.2753/MIS0742-1222280208

Haas, C. N., & Scheff, P. A. (1990). Estimation of averages in truncated samples. *Environmental Science & Technology*, 24(6), 912-919. doi:10.1021/es00076a021

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *The Journal of Marketing Theory and Practice*, 19(2), 139-152. doi:10.2753/MTP1069-6679190202

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics and Data Analysis*, 40(4), 685-711. doi:10.1016/S0167-9473(02)00072-5

Headrick, T. C. (2010). *Statistical simulation: Power method polynomials and other transformations*. Boca Raton, FL: CRC Press.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3), 255-259. doi:10.1016/0165-1765(80)90024-5

Ketkar, S., Kock, N., Parente, R., & Verville, J. (2012). The impact of individualism on buyer-supplier relationship norms, trust and market performance: An analysis of data from Brazil and the U.S.A. *International Business Review*, 21(5), 782-793. doi:10.1016/j.ibusrev.2011.09.003

Kim, M. J., Park, C. G., Kim, M., Lee, H., Ahn, Y.-H., Kim, E., Yun, S.-N., & Lee, K.-J. (2012). Quality of nursing doctoral education in Korea: Towards policy development. *Journal of Advanced Nursing*, 68(7), 1494-1503. doi:10.1111/j.1365-2648.2011.05885.x

Knez, P. J., & Ready, M. J. (1997). On the robustness of size and book-to-market in cross-sectional regressions. *The Journal of Finance*, 52(4), 1355-1382. doi:10.2307/2329439

Kock, N. (2010). Using WarpPLS in e-collaboration studies: An overview of five main analysis steps. *International Journal of e-Collaboration*, 6(4), 1-11. doi:10.4018/jec.2010100101

Kock, N. (2013). *WarpPLS 4.0 User Manual*. Laredo, TX: ScriptWarp Systems.

Kock, N. (2014). Advanced mediating effects tests, multi-group analyses, and measurement model assessments in PLS-based SEM. *International Journal of e-Collaboration*, 10(1), 1-13. doi:10.4018/ijec.2014010101

Kock, N., & Lynn, G. S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, 13(7), 546-580.

Lohmöller, J.-B. (1989). *Latent variable path modeling with partial least squares*. Heidelberg, Germany: Physica-Verlag.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, 109(3), 502-511. doi:10.1037/0033-2909.109.3.502

Mattson, S. (1997). How to generate non-normal data for simulation of structural equation models. *Multivariate Behavioral Research*, 32(4), 355-373. doi:10.1207/s15327906mbr3204\_3

Robert, C. P., & Casella, G. (2005). *Monte Carlo statistical methods*. New York, NY: Springer.

Wold, S., Trygg, J., Berglund, A., & Antti, H. (2001). Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 131-150. doi:10.1016/S0169-7439(01)00156-3