Information Management and Business Review
Vol. 7, No. 2, pp. 113-130, April 2015 (ISSN 2220-3796)

PLS-based SEM Algorithms: The Good Neighbor Assumption, Collinearity, and Nonlinearity

Ned Kock, Milton Mayfield
Texas A&M International University, USA
nedkock@gmail.com

Abstract: The partial least squares (PLS) method has been extensively used in information systems research, particularly in the context of PLS-based structural equation modeling (SEM). Nevertheless, our understanding of PLS algorithms and their properties is still progressing. With the goal of improving that understanding, we provide a discussion on the treatment of reflective and formative latent variables in the context of three main algorithms used in PLS-based SEM analyses –PLS regression, PLS Mode A, and PLS Mode B. Two illustrative examples based on actual data are presented. It is shown that the "good neighbor" assumption underlying modes A and B has several consequences, including the following: the inner model influences the outer model in a way that increases inner model coefficients of association and collinearity levels in tandem, and makes measurement model analysis tests dependent on structural model links; instances of Simpson's paradox tend to occur with Mode B at the latent variable level; and nonlinearity is improperly captured. In spite of these mostly detrimental outcomes, it is argued that modes A and B may have important and yet unexplored roles to play in PLS-based structural equation modeling analyses.

Keywords: *Partial Least Squares, Latent Variable, Formative Measurement, Reflective Measurement, PLS Regression, PLS Mode A, PLS Mode B, Collinearity*

1. Introduction

The partial least squares (PLS) method has been steadily used in information systems research, particularly in the context of PLS-based structural equation modeling (SEM). For example, in the period from 2002 to 2014, the journal *Information Systems Research* published 43 articles, out of a total of 548, employing PLS. In the same period, the journal *MIS Quarterly* published 34 articles, out of a total of 482, employing PLS. The *European Journal of Information Systems* published 46 articles, out of 633. While at first glance these numbers may appear to be small, at between 7 and 8 percent of all articles published, they become more impressive when we consider the multitude of empirical methods available to information systems researchers, and the fact that the totals include a variety of publications, including non-empirical pieces – e.g., conceptual, theory, opinion, and editorial articles. As an important tool in the information systems researcher's quantitative analysis arsenal, much has been said about the possible advantages and disadvantages of PLS-based SEM, especially when it is compared with the more traditional covariance-based SEM (Chin, 1998; see, also, Haenlein & Kaplan, 2004; Recker & La Rosa, 2012). These discussions, which usually revolve around advantages of PLS-based SEM, have been enhanced by cogent arguments suggesting that some advantages have been overstated (Goodhue et al., 2012).

Amid this debate, however, much less has been said in the past about how different algorithms routinely used in PLS-based SEM can lead to different requirements and interpretations of results in the context of information systems research. Even outside the field of information systems, comparative discussions about different PLS algorithms have been limited. Temme et al. (2006) widely cited report discusses different algorithms in the context of various software implementations, but only with passing reference to important aspects of the underlying mathematics that characterize different algorithms. Tenenhaus et al. (2005) mathematical discussion of PLS-based SEM, published in a statistics journal and building on Lohmöller's (1989) seminal book on PLS, is detailed and extensive, but nevertheless leaves out several aspects that may lead to somewhat surprising conclusions about the possible impacts of using different PLS algorithms on PLS-based SEM analyses – in the information systems field, and also in other applied fields. While PLS-based SEM analyses focus on latent variables, the most basic "material" of these analyses is the raw data; or the values of the manifest variables across the various rows that make up the raw data table. Manifest variables are

aggregated in the calculation of latent variable scores in PLS-based SEM, also known as PLS path modeling (Tenenhaus et al., 2005). Manifest variables are measured directly, often using Likert-type scales (e.g., "1 – Strongly disagree" to "5 – Strongly agree"), based on which individuals provide answers to question-statements in questionnaires (Chin, 1998; Lohmöller, 1989; Noonan & Wold, 1983). Latent variables are measured indirectly through the manifest variables associated with them; latent variable scores in fact do not actually exist until a PLS-based SEM analysis is conducted (Gefen et al., 2000; Fornell & Larcker, 1981; Wold, 1974).

Latent variables employed in PLS-based SEM are usually defined as belonging to one of two main types – reflective and formative (Diamantopoulos, 1999; Diamantopoulos & Siguaw, 2002; Petter et al., 2007). In a reflective latent variable, the corresponding manifest variables are designed to redundantly reflect the latent variable. For example, a latent variable *job satisfaction*, at the individual level of analysis (i.e., reflecting the satisfaction of an individual with his or her job), could be reflectively measured through the following question-statements: "I am very satisfied with my current job", "My present job gives me internal satisfaction", and "My job gives me a sense of fulfillment". The use of question-statements that are redundant has two main advantages. The first is that it allows for a confirmatory factor analysis to be conducted (Ehremberg & Goodhart, 1976; Schumacker & Lomax, 1996; Thompson, 2004). Through this type of analysis a researcher can assess the extent to which the respondents understood the question-statements in the same way that the researcher did, and the extent to which the respondents understood the question-statements in the same way that other respondents did. These and other related assessments refer to the validity and reliability of the question-statements used for latent variables in a SEM analysis (Nunnaly, 1978; Nunnally & Bernstein, 1994). The second advantage of the use of question-statements that are redundant is that, if they pass validity and reliability criteria, they are likely to reduce measurement error (Rencher, 1998; Rosenthal & Rosnow, 1991; Schumacker & Lomax, 1996).

In a formative latent variable, the corresponding manifest variables are designed to measure different facets, or dimensions, of the latent variable (Diamantopoulos & Siguaw, 2006). Also, formative manifest variables are designed specifically to avoid redundancy. Still, formative manifest variables are designed to be significantly associated with their latent variables (Petter et al., 2007). For example, a latent variable *electronic communication use*, at the team level of analysis and used to evaluate the degree to which a team used electronic communication tools in a project, could be formatively measured through the following question-statements: "The team used e-mail to fellow team members", "The team used a Web page dedicated to this project", and "The team used teleconferencing". The corresponding manifest variables measure different facets of electronic communication use, and are not redundant because the tools referred to in the question-statements are not expected to necessarily be used together. The use of e-mail, dedicated Web page, and teleconferencing can be reasonably expected to be used independently from one other (Kock, 2005; 2008; Roy, 2012). Like reflective latent variables, formative latent variables are expected to reduce measurement error when properly designed (Lohmöller, 1989). However, given their different nature, validity and reliability assessments of formative latent variables rely on different criteria (Diamantopoulos, 1999; Petter et al., 2007). More specifically, the weights linking formative manifest and latent variables must be statistically significant, and collinearity levels among manifest variables must be low. The former can be tested through resampling in PLS-based SEM (Chin, 1998; Martin, 2007), while the latter can be tested via the calculation of variance inflation factors and their comparison against a threshold (Diamantopoulos & Siguaw, 2006; Petter et al., 2007).

This paper provides a discussion on the treatment of reflective and formative latent variables in the context of three main algorithms; namely the PLS regression algorithm, and two main variations of that algorithm, known as PLS modes A and B. Two illustrative examples based on actual data are presented, the first focusing on a model with two linked reflective latent variables, and the second on a model with one reflective latent variable pointing at one formative latent variable. The models are purposely simple, to illustrate in a generic way important aspects of the algorithms, and are not aimed at making any empirical contribution. The paper ends with a discussion and recommendations on the use of the algorithms with reflective and formative latent variables. There are two main reasons why this paper is relevant for information systems researchers. The first is that it clarifies a number of issues that appear to be largely unknown to many researchers in the

field of information systems, who tend to use PLS-based SEM extensively in their investigations. For example, many of those researchers might be surprised to know that, by using the PLS modes A and B, their hypotheses directly influence the results of their analysis in such a way as to increase the level of collinearity among variables that are linked. The second reason why this paper is relevant for information systems researchers is that some of the recommendations that follow from the discussion of the three algorithms have direct practical consequences in the context of investment in and use of information technologies, as will be seen later.

2. PLS Regression and its use in SEM

This section discusses a fundamental form of PLS-based SEM analysis, where latent variable scores are generated entirely based on manifest variable scores. Following Temme et al.'s (2006) terminology, we refer to the underlying algorithm used to conduct this fundamental form of PLS-based SEM analysis simply as the "PLS regression" algorithm. The PLS regression algorithm appears to be most often used in chemometrics applications (Wold et al., 2001), in nonlinear SEM analyses (Guo et al., 2011), and in SEM analyses where collinearity must be minimized (Temme et al., 2006). We provide a discussion where a full set of PLS-related algorithms are expressed in modern matrix algebra notation (Harville, 1997; Penney, 2004). This provides a complement to previous discussions and, given the modern matrix algebra notation used (including left and right division, as well as pseudo inverses), facilitates the implementation of the algorithms through various publicly available mathematical programming languages and packages such as MATLAB, GNU Octave, FreeMat and R. The modern matrix algebra notation used is very similar to the general notation used in these mathematical programming languages and packages.

The PLS regression algorithm (Gerlach et al., 1979; Lohmöller, 1989; Noonan & Wold, 1983; Wold, 1974) does not require the existence of an inner model, which is one of its key defining characteristics in the context of PLS-based SEM. It starts with arbitrary values being assigned to W and L, and the initialization of F. The column vectors (i.e., one-column matrices) W and L store the standardized weights and loadings, respectively, linking the manifest variable matrix X and the column vector F containing the scores for a latent variable. The manifest variable matrix X stores the original data, and is made up of several column vectors X_1, X_2 … that have a length that is equal to the size of the sample used in the analysis (as does F). The algorithm then proceeds with the iterative solutions of the equations in the steps shown below, where $Stdz(\quad)$ denotes the standardization function. The algorithm stops when the change in W in two consecutive iterations falls below a small threshold fraction, such as .001. After the algorithm stops, latent variable scores and loadings must be re-calculated based on the final values assumed by the weights. Inner model coefficients of association are then calculated based on the re-calculated values of the latent variable scores.

$$\mathrm{L} = \hat{\mathrm{X}}^{\mathrm{T}} \,/\, \hat{\mathrm{F}}^{\mathrm{T}} \qquad \text{[Step 1]}$$

$$\hat{\mathrm{F}} = Stdz\left(\hat{\mathrm{X}}\,\mathrm{L}^{\mathrm{T}}\right) \qquad \text{[Step 2]}$$

$$\mathrm{W} = \hat{\mathrm{X}} \setminus \hat{\mathrm{F}} \qquad \text{[Step 3]}$$

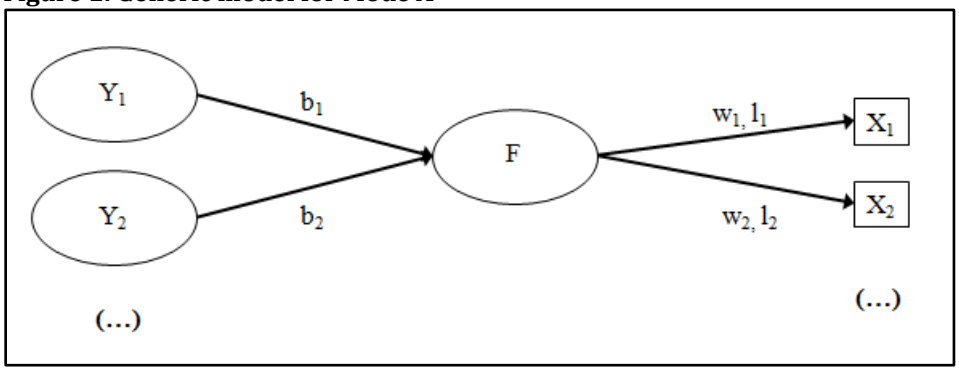
The hat symbol refers to the standardized version of the corresponding variable (e.g., $\hat{\mathrm{F}}$ is the standardized version of F). In the specific case of $\hat{\mathrm{F}}$, it can also be seen as the iteratively estimated, or expected, value of F, given X. Latent variable scores are always standardized. Corresponding un-standardized values can be obtained by multiplying standardized scores by standard deviation multiplication and mean addition, where these values (i.e., standard deviation and mean) are obtained based on the latent variable indicators (e.g., using a weighted average). The superscript ^{T} refers to the transpose matrix operation (e.g., X^{T} is the transpose of X). The symbols \ and / refer to the left and right matrix division operations. For example, considering two matrices A and B, the left division is defined as: $\mathrm{A} \setminus \mathrm{B} = \mathrm{A}^{+}\,\mathrm{B}$. The right division is defined as: $\mathrm{A} \,/\, \mathrm{B} = \mathrm{A}\,\mathrm{B}^{+}$. The matrices A^{+} and B^{+} are the pseudoinverses, or generalized inverses (Rao & Mitra, 1971), of the matrices A and B. In this fundamental algorithm an endogenous latent variable is completely determined by the manifest variables with which it is associated; that is, the estimation of its score is not influenced by any predictor latent variable pointing at it in the inner model. The algorithm stops when two conditions are met: (a) the latent variable is an exact linear combination of the manifest variables with which it is associated; and (b) the error terms of the equations where the latent variable is the predictor of the indicators are minimized. Since these criteria are met while a latent variable is completely determined by the manifest

variables with which it is associated, this algorithm is arguably one of the most effective at minimizing collinearity among latent variables in a model with multiple latent variables (Hair et al., 2009; Kock & Lynn, 2012; Lohmöller, 1989). This applies to both reflective and formative latent variables, whose nature does not depend on the algorithm used to implement them, as will be discussed in more detail later. In a PLS-based SEM analysis, this algorithm would be used to calculate the scores for all the latent variables included in the model, after which inner model coefficients would be estimated through path analysis (Duncan, 1966; Mueller, 1996; Wright, 1934; 1960).

The Good Neighbor Assumption (GNA): The good neighbor assumption (GNA) is defined here as the fundamental assumption, made in the PLS-based SEM algorithms known as PLS Mode A and PLS Mode B (Lohmöller, 1989; Wold, 1974), that the weights and loadings linking latent variable scores and their indicators should be estimated in such a way as to maximize the strength of the associations among latent variables that are causally linked in a structural model. The term "good neighbor", as used in this paper, is derived from Adelman & Lohmoller's (1994, p. 357) description of the GNA, which is that a latent variable should be estimated "… so that it is a good neighbor in its neighborhood. That is, estimate the [latent variable] so that it is well predicted by its predecessors in the path diagram and is a good predictor for its followers in the diagram." In essence, the GNA implies that the inner model, where hypothesized causal relationships among latent variables are defined, should be taken into consideration during the calculation of latent variable scores (Lohmöller, 1989). The latter, the latent variable scores, are also defined in part by the outer model, where associations among manifest and latent variables are specified (Adelman & Lohmoller, 1994; Tenenhaus et al., 2005).

The PLS modes A and B: There are two main "modes" used in PLS-based SEM, which could be seen as PLS regression variations, and which are usually referred to as modes A and B (Vittadini et al., 2007). A third mode, often referred to as MIMIC (Lohmöller, 1989), can be seen as a combination of modes A and B (Tenenhaus et al., 2005). In both modes, A and B, the GNA is employed; this is a key difference between these modes and PLS regression. Each mode can be fully expressed based on a latent variable block, which is essentially a model containing several latent variables, whose scores are stored in a matrix Y, pointing at a latent variable whose scores are stored in a column vector F. The PLS Mode A algorithm starts with arbitrary values being assigned to W and L, and the initialization of F. The terms used here have generally the same meanings as in PLS regression. The vector W stores standardized weights, and the vector L stores loadings. These link the manifest variable matrix X and the column vector F containing the scores for a latent variable. The matrix Y contains the scores of the latent variables that point at the latent variable F. The column vector B stores the standardized regression coefficients b_1, b_2 … linking Y and F. In Mode A the matrix X "reflects" F, which is illustrated in Figure 1 through the arrows pointing from F out to X_1, X_2 …, which are the elements of the matrix X. PLS Mode A also depends on the coefficients associated with links going from the latent variable to other latent variables in the model, which are omitted here for simplicity.

Figure 1: Generic model for Mode A



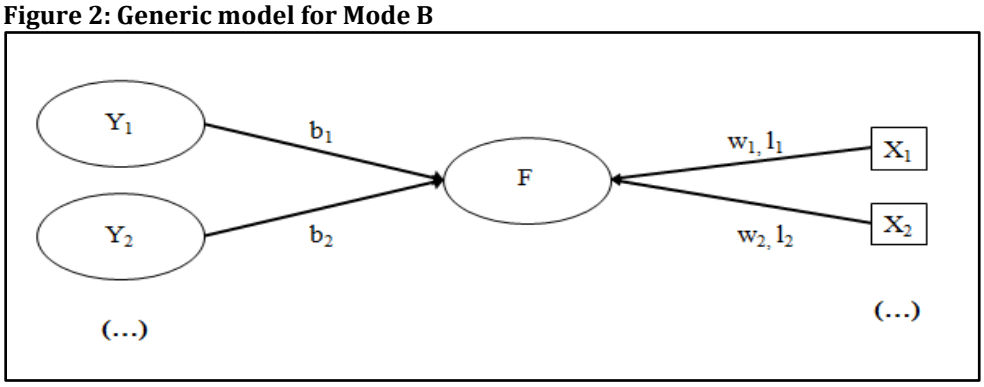
The matrix of latent variable scores Y itself is made up of several column vectors Y_1, Y_2 … that have a length that is equal to the size of the sample used in the analysis (as does F). The relevant column vectors Y_1, Y_2 … are specified through the inner model. They represent latent variables that in turn are associated with their

own manifest variables. The algorithm then proceeds with the iterative solutions of the equations in the steps outlined below. The algorithm stops when the change in W in two consecutive iterations falls below a small threshold fraction. After the algorithm stops, latent variable scores and coefficients of association must be re-calculated, and loadings calculated, based on the final values assumed by the weights.

$$B = \hat{Y} \setminus \hat{F} \qquad \text{[Step 1]}$$
$$\hat{F} = Stdz(B\,\hat{Y}) \qquad \text{[Step 2]}$$
$$L = \hat{X}^T / \hat{F}^T \qquad \text{[Step 3]}$$
$$\hat{F} = Stdz(\hat{X}\,L^T) \qquad \text{[Step 4]}$$
$$W = \hat{X} \setminus \hat{F} \qquad \text{[Step 5]}$$

There is one main difference between the PLS Mode A and the PLS regression algorithms. The difference is that in Mode A the weights (W) and loadings (L) linking the manifest variable matrix X and the latent variable vector F are influenced by B, which also changes in successive iterations of the algorithm. This is an algorithmic peculiarity that follows directly from the GNA. The step through which this happens, indicated as Step 1 above, is known as the "inside approximation" (Lohmöller, 1989; Tenenhaus et al., 2005). The Mode B variation of the PLS regression algorithm starts with arbitrary values being assigned to W and L, and the initialization of F. The vectors W, L, F have the same meaning as in Mode A. Also analogously to Mode A, the column vector B in Mode B stores the standardized regression coefficients b_1, b_2 … linking Y and F. However, in Mode B, the matrix X "forms" F, which is illustrated in Figure 2 through the arrows pointing in to F from X_1, X_2 …, which are the elements of the matrix X. Like PLS Mode A, PLS Mode B also depends on the coefficients associated with links going from the latent variable to other latent variables in the model, which are omitted here for simplicity.

Figure 2: Generic model for Mode B



After arbitrary values are assigned to W and L, and F is initialized, the algorithm for Mode B proceeds with the iterative solutions of the equations in the steps outlined below. The algorithm stops when the change in W in two consecutive iterations falls below a small threshold fraction. After the algorithm stops, latent variable scores and inner model coefficients of association must be re-calculated, and loadings calculated, based on the final values assumed by the weights.

$$B = \hat{Y} \setminus \hat{F} \qquad \text{[Step 1]}$$
$$\hat{F} = Stdz(B\,\hat{Y}) \qquad \text{[Step 2]}$$
$$W = \hat{X} \setminus \hat{F} \qquad \text{[Step 3]}$$
$$\hat{F} = Stdz(\hat{X}\,W) \qquad \text{[Step 4]}$$
$$L = \hat{X}^T / \hat{F}^T \qquad \text{[Step 5]}$$

There are two main differences between the PLS regression and PLS Mode B algorithms. The first difference is that in Mode B the weights (W), and not the loadings (L), are used in the calculation of F. Before that, the weights (W) are calculated through the column vector F being regressed on X; with X being fixed (as it is the original data) and F changing in successive iterations. The second difference is that in Mode B, as with Mode A, the weights (W) linking the manifest variable matrix X and the latent variable vector F are influenced by B, with the latter also varying across iterations until convergence. This follows directly from the GNA. The main

difference between modes A and B is that in Mode A the loadings (L) are used in the estimation of the latent variable vector F across successive iterations, whereas in Mode B the weights (W) are used in the estimation of F. In other words, in Mode A the vector F predicts its manifest variables (X), whereas in Mode B the manifest variables (X) predict F. This can be somewhat confusing because, regardless of the algorithm used, the vector F is always calculated as an exact linear combination of its manifest variables (X). A common element of modes A and B is that they both implement the GNA, which can be problematic due to the resulting influence that the inner model places on the outer model. On one hand, researchers specify the inner model based on hypotheses that they intend to test. On the other hand, researchers define outer model elements by associating manifest variables with latent variables. Researchers carry out the latter task based on data collection instrument (e.g., questionnaire) design assumptions about latent-manifest variable relationships.

The interdependence of outer and inner model elements resulting from the GNA leads to a "stealth" violation of data collection instrument (e.g., questionnaire) design assumptions about latent-manifest variable relationships. It increases collinearity levels and may lead to unexpected changes in signs in the weights (Tenenhaus et al., 2005), in turn leading to the phenomenon known as Simpson's paradox, with respect to the outer model (Pearl, 2009; Wagner, 1982). The occurrence of Simpson's paradox may be associated with the existence of improbable causal relationships (Pearl, 2009; Wagner, 1982), or with instability caused by collinearity (Kock & Lynn, 2012), among other problems. With respect to the outer model, or at the latent variable level, the Simpson's paradox phenomenon is characterized by a negative contribution of one or more manifest variables to the variance explained in the latent variable to which they are assigned. That is, one or more manifest variables ends up detracting from the variance explained in the latent variable.

3. Modes and Latent Variable Types

Are the modes A and B necessary for handling reflective and formative latent variables, respectively? Even though it is widely assumed among PLS software users that the correct answer to this question is "yes" (Hardin et al., 2008; Subhadip, 2008), it could be argued that the correct answer is in fact "no" (Chatelin et al., 2002; Lohmöller, 1989; Temme et al., 2006). We present evidence here that the nature of a latent variable is independent of the algorithm used. In other words, a properly designed latent variable should either fundamentally reflective or formative, and the latent variable will "behave" as such, regardless of whether a researcher uses the PLS regression, PLS Mode A, or PLS Mode B algorithms in a PLS-based SEM analysis. The main reason for this is that latent variables are normally "designed" by researchers, typically through the researchers' design of question-statements to be answered by questionnaire respondents, to be reflective or formative. However, the theoretical scenario above is complicated by a practical reality, which is that some latent variables behave as though they were not of the type assumed by design, and some even behave as formative-reflective hybrids. These are often due to conflicting views about question-statements by questionnaire designers and respondents. A latent variable may be designed as formative, but pass only reflective measurement assessment criteria. Conversely, it may be designed as reflective, but pass only formative measurement assessment criteria. Some latent variables pass both types of criteria, thus apparently "behaving" as formative and reflective at the same time; these are the ones that are referred to here as hybrids.

In reflective latent variables, the manifest variables are designed to be redundant with one another and with their latent variable, but not with the manifest variables associated with other latent variables (which indicates good discriminative power). In formative latent variables, the manifest variables are designed so that they are *not* redundant with one another or their latent variable, and yet are significantly associated with their latent variable on a multivariate basis (i.e., when controlling for one another, which indicates good formative power). Such design may well be done subconsciously, without the researcher even knowing the difference between reflective and formative measurement. For example, let us assume that a researcher wants to measure the respondents' satisfaction with a meal that includes appetizer, main course, and dessert. The measurement will be reflective if the researcher asks redundant questions of the type: "I am satisfied with this meal", "This meal was great", and "I really enjoyed this meal". The measurement will be formative if the researcher asks non-redundant questions of the type: "I enjoyed the appetizer", "I enjoyed the main course", and "I enjoyed the dessert". Latent variable measurement assessment criteria, used for reliability and

validity testing, vary depending on whether a latent variable is reflective or formative. The criteria for reflective latent variables are aimed at ensuring both redundancy and discriminative power; for formative latent variables they are aimed at avoiding redundancy and ensuring formative power (Diamantopoulos & Siguaw, 2006; Lohmöller, 1989; Petter et al., 2007). If a latent variable does not satisfy either set of criteria, then it should not be used in PLS-based SEM analyses.

Two Illustrative Analyses: In this section data from two actual studies, with only partial results presented, are used to illustrate the main points made in this paper. MATLAB scripts implementing the algorithms discussed here were used in the analyses, and the results were double-checked with the following publicly available multivariate statistical analysis software tools: PLS-Graph, Smart PLS, SPSS, and WarpPLS. The focus of this section is not on the presentation of domain-specific empirical results, but rather on a comparison across algorithms (PLS regression and PLS modes and A and B) of inner model coefficients of association, weights, loadings, contributions by individual manifest variables to the variance explained in the latent variables to which they are assigned, and collinearity estimates between latent variables shown as linked in the inner model. Hence only partial results are presented. The first study features two linked reflective latent variables; the second, one reflective latent variable pointing at one formative latent variable.

The First Illustrative Study: The data for the first study was from a survey-based investigation of 193 professionals, from across the United States. The professionals surveyed used social networking sites, such as Facebook and LinkedIn, to different degrees during the time the data was collected. Females contributed 53.4 percent of the data. The average age of the professionals was 27.2 years. In terms of their education level, 14.5 percent of the respondents had only completed high school, 13.5 percent had a 2-year college degree, 43 percent had a 4-year college degree, 23.8 percent had a master's degree, and 4.7 percent had a doctoral degree. The two latent variables in the first study are *job satisfaction* and *job performance*. *Job satisfaction*, a reflective latent variable, is the degree to which a professional is satisfied with his or her job. *Job performance*, also a reflective latent variable, is the self-assessed performance level of a professional at his or her job. In the model investigated, *job satisfaction* points at *job performance*; that is, it is hypothesized that professionals that are highly satisfied with their jobs would also tend to have high performance levels.

Figure 3 shows the model that refers to the first study. The only inner model coefficient shown is b, the standardized partial regression coefficient for the link between *job satisfaction* and *job performance*. The weights and loadings linking the manifest variables Prf1, Prf2 and Prf3 and the latent variable *job performance* are indicated, respectively, as w_1, w_2 and w_3 and l_1, l_2 and l_3. The coefficients indicated as CVE_1, CVE_2 and CVE_3 refer to the contributions of each of the manifest variables to the variance explained in the latent variable *job performance*. The VIF coefficient refers to the variance inflation factor (Hair et al., 2009; Kline, 1998) between *job satisfaction* and *job performance*, a measure of *lateral* collinearity between these two latent variables (Kock & Lynn, 2012). Table 1 summarizes the results obtained from the analysis of the first study's two-latent variable model. The three columns refer to the algorithms used: PLS regression (indicated as PLSR), PLS Mode A (PLSR-A), and PLS Mode B (PLSR-B). Shown in a shaded cell is an instance of Simpson's paradox (Pearl, 2009; Wagner, 1982), whereby the manifest variable Prf1 makes a negative contribution to the explained variance in the latent variable to which it was assigned, namely *job performance*.

Figure 3: First Study Model

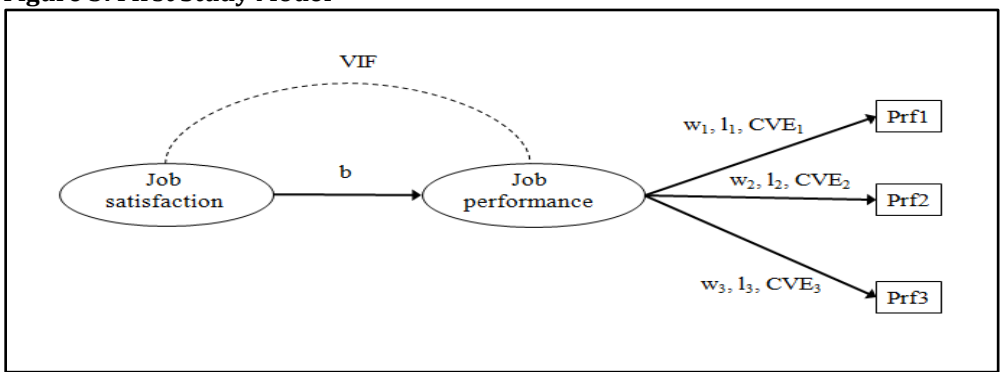


Table 1: First study's algorithms and related coefficients

	PLSR	PLSR-A	PLSR-B
B	.433	.438	.455
w_1	.339	.281	-.142
l_1	.885	.867	.698
CVE_1	.300	.244	-.099
w_2	.367	.394	.626
l_2	.958	.965	.979
CVE_2	.352	.380	.613
w_3	.365	.391	.501
l3	.953	.961	.971
CVE_3	.348	.376	.486
VIF	1.231	1.237	1.261

Simpson's paradox occurs only with PLSR-B. The standardized partial regression coefficient for the link between *job satisfaction* and *job performance* (b) goes up as we move from PLSR to PLSR-A and then to PLSR-B. So does the level of collinearity between *job satisfactions* pointing at *job performance*, measured through the VIF coefficient. Again, the type of collinearity we refer to here is *lateral* collinearity, or predictor-criterion collinearity (Kock & Lynn, 2012).

The Second Illustrative Study: The data for the second study was from a survey-based investigation of 290 new product development teams from 66 organizations located in Northeastern United States. These were teams that developed new products, such as new toothbrushes and car parts. The teams were moderately to highly geographically distributed; enough to make face-to-face communication difficult, and thus strongly stimulate the use of electronic communication tools. Team size ranged from 3 to 300 individuals, with a mean of approximately 24 individuals, and a median of 10 individuals. The two latent variables in the second study are *degree of coordination* and *electronic communication use*. *Degree of coordination*, a reflective latent variable, is the degree to which a team employed progress-tracking techniques to coordinate its work. *Electronic communication use*, a formative latent variable, is the degree to which a team used e-mail, a web page dedicated to the project, and teleconferencing to communicate. In the model investigated, *degree of coordination* points at *electronic communication use*; that is, it is hypothesized that teams that coordinate their activities to a high degree tend to make more general use of the electronic communication tools that make up the latent variable *electronic communication use*: e-mail, a web page dedicated to the project, and teleconferencing.

The model that refers to the second study is depicted in Figure 4. As with the first study, the only inner model coefficient shown is b, the standardized partial regression coefficient associated with the link between *degree of coordination* and *electronic communication use*. The weights and loadings linking the manifest variables Ecu1, Ecu2 and Ecu3 and the latent variable *electronic communication use* are indicated, respectively, as w_1, w_2 and w_3 and l_1, l_2 and l_3. The coefficients indicated as CVE_1, CVE_2 and CVE_3 refer to the contributions of each of the manifest variables to the variance explained in the latent variable *electronic communication use*. The VIF coefficient refers to the variance inflation factor (Hair et al., 2009; Kline, 1998) between *degree of coordination* and *electronic communication use*, a measure of collinearity between these two latent variables. The results obtained from the analysis of the second study's two-latent variable model are summarized in Table 2. As with the first study, the three columns refer to the algorithms used: PLS regression (PLSR), PLS Mode A (PLSR-A), and PLS Mode B (PLSR-B). Shown in a shaded cell is an instance of Simpson's paradox (Pearl, 2009; Wagner, 1982), whereby the manifest variable Ecu3 makes a negative contribution to the explained variance in the latent variable to which it was assigned, namely *electronic communication use*.

Figure 4: Second study model

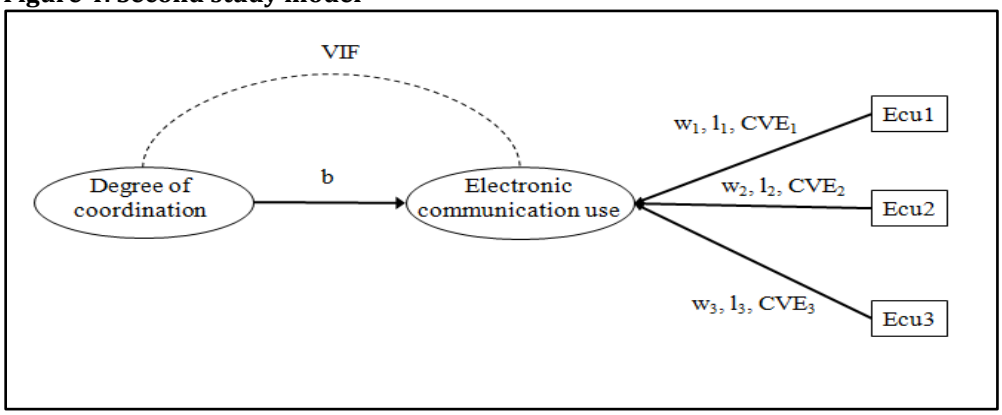


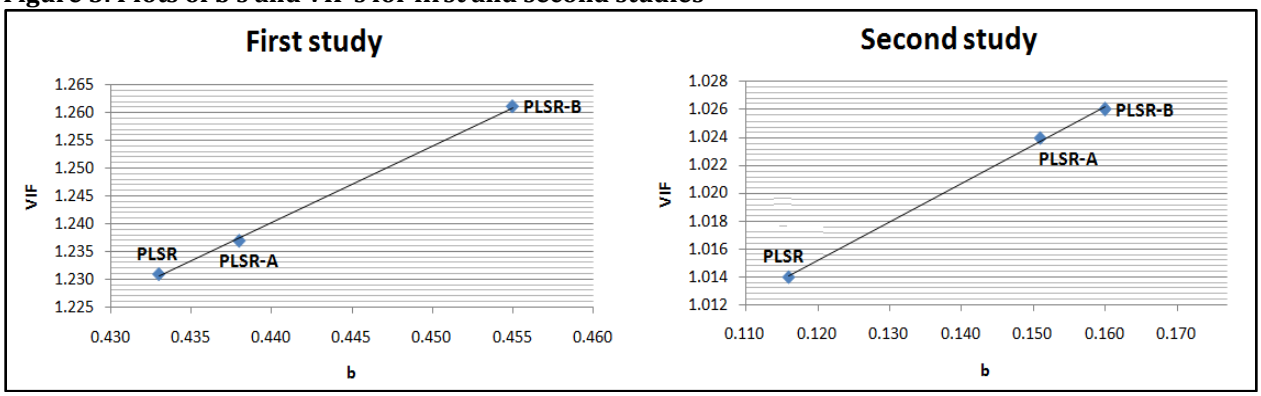
Table 2: Second study's algorithms and related coefficients

	PLSR	PLSR-A	PLSR-B
b	.116	.151	.160
w_1	.504	.265	.005
l_1	.750	.560	.316
CVE_1	.378	.148	.002
w_2	.427	.152	-.013
l_2	.627	.371	.181
CVE_2	.268	.056	-.002
w_3	.490	.838	1.001
l_3	.723	.950	1.000
CVE_3	.354	.796	1.001
VIF	1.014	1.024	1.026

Again, as with the first study, Simpson's paradox occurs only in Mode B. Also similarly to the first study, the standardized partial regression coefficient for the link between *degree of coordination* and *electronic communication use* (b) goes up as we move from PLSR to PLSR-A and then to PLSR-B; and so does the level of collinearity between these two latent variables, measured through the VIF coefficient.

Coefficients of Association and Collinearity Levels: The measures of collinearity between predictor and criterion latent variable (the VIF coefficients) and the measures of the strength of the association between these latent variables (the b coefficients) are plotted in Figure 5, for the first and second studies. The line shown is the best fitting regression line passing through the points. As can be seen, the relationship between VIF and b appears to be very close to a linear relationship, with VIF increasing proportionally to increases in b.

Figure 5: Plots of b's and VIF's for first and second studies



In summary, in both the first and the second studies modes A and B led to increases in the levels of collinearity between predictor and criterion latent variables. The observed increases in collinearity levels were directly proportional to increases in inner model coefficients of association between predictor and criterion latent variables. Also, in both the first and the second studies Mode B has led to instances of Simpson's paradox (Pearl, 2009; Wagner, 1982) at the latent variable level, in which one manifest variable ended up detracting from the variance explained in the latent variable to which it was assigned. It should be noted that the increases in collinearity levels occurred, in the illustrated studies presented here, without actually making the latent variables in question actually collinear. In PLS-based SEM two or more latent variables would be defined as actually collinear if their VIFs were greater than 3.3 (Kock & Lynn, 2012; see also: Petter et al., 2007), a threshold far higher than the highest VIF obtained (in the first study). Nevertheless, the increase in the level of collinearity causes an increase in the coefficients of association, thus increasing the likelihood that type I errors (false positives) will be committed.

4. Discussion

This section provides a discussion organized in various sub-sections, each addressing a key topic based on the preceding arguments and illustrative examples. The key topics addressed here are the GNA and the exploratory nature of analyses with modes A and B, the confirmatory nature of the PLS regression algorithm, the GNA and the model specificity of confirmatory factor analyses in modes A and B, Simpsons' paradox and measurement model assessment with Mode B, and nonlinear PLS regression as an alternative to the GNA in terms of capturing nonlinearity. These topics are discussed in light of the practical tradeoffs that clearly exist between committing types I and II errors, particularly when technology cost and competitive advantage effects are considered.

Exploratory SEM Analyses and PLS Modes A and B: Under the GNA, the inner model influences the outer model in a way that increases inner model coefficients of association, through a small increase in the levels of collinearity, thus making type I errors more likely (Hair et al., 2009; Petter et al., 2007). This applies to both modes A and B; particularly to the latter (Mode B). Moreover, under the GNA, the researcher's hypotheses influence the data analysis results, specifically by influencing the latent variable scores used to calculate inner model coefficients of association. The interdependence between inner and outer model, caused by the GNA, is analogous to the effects that sharing manifest variables or proxies would have on the coefficient of association in a predictor-criterion latent variable pair. For instance, let us say that in a predictor-criterion latent variable pair one manifest variable from the predictor is taken, a small amount of error is added to it (to avoid problems stemming from full linear dependence), and then the manifest variable is also used in the calculation of the criterion latent variable. This would artificially increase the level of collinearity between the two latent variables, and thus artificially inflate the coefficient of association between the latent variables.

Whether the increase in the level of collinearity in this scenario is high enough for one to consider the latent variables to be actually collinear is irrelevant to the point being made here. A measure of the level of collinearity among latent variables, such as a variance inflation factor, would have to be above a certain threshold (e.g., greater than 3.3) to be suggestive of actual "pathological" collinearity (Kline, 1998; Kock & Lynn, 2012). The point that is being made here, however, is that even a small increase in the level of collinearity, leading to a variance inflation factor going up from 1.231 to 1.237, would also increase the corresponding inner model coefficient of association. This appears to be one of the consequences of the GNA, although more so in Mode B than Mode A. Another way of looking at the situation created by the GNA in modes A and B though, is that the GNA is an attempt at capturing nonlinearity between predictor-criterion latent variables. This may well be the reason why modes A and B have frequently been referred to as variations of a *nonlinear* iterative partial least squares algorithm, whose acronym is NIPALS (Wold, 1974). In fact, capturing nonlinearity is probably the most promising underlying reason for the GNA. While promising, there appear to be problems with it, which are discussed in more detail later in this section.

Nevertheless, the incremental increase in collinearity levels caused by the GNA may not always be a bad thing, since there are often practical tradeoffs between making type I and II errors. One could argue that, under the GNA, PLS-based SEM algorithms (i.e., modes A and B) become more sensitive to small effects. This

may be seen as an advantage when there is not as much a concern about making type I errors, as long as type II errors (false negatives) are avoided. In technology-mediated effects investigations, this would refer to situations in which missing a small effect that could become large (e.g., a small short-term effect, which could become large in the long term) is a much bigger problem than finding an effect that does not actually exist. Relatively low cost technologies that can greatly increase market share, such as customer-oriented web service-based features on an existing company website, would be examples. Based on the above discussion, it seems reasonable to recommend that researchers should clearly acknowledge the exploratory nature of their PLS-based SEM analysis when using PLS modes A and B. Moreover, arguably they should consider the use of modes A and B over PLS regression, particularly when "sensitive" methods for standard error estimation are used (e.g., the "stable" P value calculation method discussed by Kock, 2013), in situations where type I errors are not as much a source of concern as type II errors. This is likely to be the case when technology-mediated effects appear to be small and the range of variation in the criterion variable is large. Examples would be scenarios involving technologies whose cost is relatively low and whose potential impact over time is significant. In these cases, missing a technology investment opportunity may be significantly more costly than making it and not obtaining the expected return. For example, the investment in not very costly disruptive technology innovations may lead to small gains in the short term, but significant gains over many years (Dos Santos & Peffers, 1995).

Confirmatory SEM Analyses and PLS Regression: While some would be tempted to call the GNA a form of "cheating", others may see it as an assumption that renders a SEM analysis "exploratory", as opposed to "confirmatory", as argued above. That is, a PLS-based SEM analysis employing the GNA, using Mode A or Mode B or a combination of them, could be seen as an exploration of the data at hand under specific hypothetical assumptions. A problem that must be recognized by researchers is that the hypothesis-testing results under the GNA would be influenced by those specific hypothetical assumptions. And that influence would be in favor of the hypotheses; that is, when links are hypothesized, the corresponding coefficients of association would tend to increase under the GNA. As noted before, this type of influence may lead to type I errors (Hair et al., 2009; Rosenthal & Rosnow, 1991). In an equivalent analysis using the PLS regression algorithm, a hypothesized inner model link between two latent variables would have no influence on the test of that hypothesis. This would possibly render an analysis using the PLS regression algorithm to have confirmatory elements, in contrast to the more exploratory nature of analyses employing modes A or B.

One could argue that hypothesized relationships among latent variables, specified through the inner model, always influence coefficients of association in latent variable blocks with more than one predictor latent variable, regardless of the algorithm used. That is indeed correct, because in latent variable blocks with more than one predictor latent variable each of the predictors "competes" for the explained variance in the block's criterion latent variable (Hair et al., 2009; Rencher, 1998). In this scenario, however, in the absence of the GNA the inner model influences the inner model (not the outer model), and in a way that reduces the likelihood of type I errors. The reason is that the coefficients of association linking each of the predictors with the criterion latent variation in a block will tend to *decrease* under this competition, as correlations among the predictors are controlled for (Hair et al., 2009; Lohmöller, 1989; Rosenthal & Rosnow, 1991). Based on the above discussion, one could contend that information systems researchers should favor the PLS regression algorithm over PLS modes A and B whenever the purpose of their investigations is clearly academic and/or confirmatory in nature. One example would be an investigation whose main goal is to falsify a theoretical model, as opposed to finding a small technology-mediated effect that would be costly to miss. Moreover, researchers should favor the use of PLS regression over modes A and B, particularly when "sensitive" methods for standard error estimation are used, in situations where type II errors are not as much a source of concern as type I errors. This is likely to be the case when the technologies involved are very costly – e.g., enterprise systems (Davenport, 2000).

Model Specificity of Confirmatory Factor Analyses with PLS Modes A and B: Because of the GNA, weights and loadings become dependent on the inner model structure when modes A and B are employed. Confirmatory factor analyses generate loadings and cross-loadings, which are used, together with other coefficients, in measurement instrument validity assessments (Fornell & Larcker, 1981; Hair et al., 2009; Kline, 1998). Therefore, confirmatory factor analyses must be re-done each time an inner model structure

changes when modes A and B are employed. This is not the case with the PLS regression algorithm, since in it the inner model does not influence the outer model coefficients in any way. For example, let us assume that modes A or B are used in a study that involves the comparison of multiple configurations with the same latent variables but different inner model structures. In this example, a separate confirmatory factor analysis must be conducted for each of the configurations. With the PLS regression algorithm, a confirmatory factor analysis may be conducted without an inner model; i.e., with a "model" without links among latent variables. This is not the case with PLS modes A and B. If PLS modes A or B are used, a model without links will effectively have no solution. As seen earlier, in these modes the iterative algorithms contain steps in which path coefficients are estimated, and then subsequently weights or loadings are estimated based on them. Without the ability to estimate path coefficients, the algorithms in PLS modes A and B cannot yield solutions without significant adaptations – e.g., defaulting to the PLS regression algorithm in the absence of an inner model.

Given the above, one could argue that investigators should be mindful of, if not explicitly acknowledging in writing, the fact that confirmatory factor analysis results are model-specific when PLS modes A or B are used. In addition, if multiple variations of an SEM model are investigated, even if the latent variables in the models are the same, separate confirmatory factor analyses should be conducted and their results reported for each model. Appendices could be used when reporting these, to avoid "crowding" reports with numeric tables. With modes A or B (particularly the latter), certain inner model structures may yield loadings, weights and other coefficients (e.g., composite reliability coefficients) that do not pass measurement model assessment criteria. As those inner model structures change, criteria-acceptable coefficients may be obtained.

Simpsons' Paradox and Measurement Model Assessment with PLS Mode B: Formative latent variables do not change in nature if mode B is not used. That is, if either the PLS regression algorithm or its Mode A variation are used, formative latent variables do not "become reflective" latent variables. Manifest variables associated with properly designed formative latent variables are expected to have statistically significant weights and present a low level of redundancy among them (i.e., the manifest variables). Typically formative latent variables will also have lower loadings than reflective latent variables associated with their manifest variables. These properties should generally hold regardless of the algorithm used (Cenfetelli & Bassellier, 2009; Petter et al., 2007), be it PLS regression, PLS Mode A, or PLS Mode B. These differences are illustrated in Table 3 where the following manifest variable coefficients, calculated through the PLS regression algorithm, are shown side-by-side for a reflective (*job performance*) and a formative (*electronic communication use*) latent variable: loadings (L), P values for loadings {P(L)}, weights (W), P values for weights {P(W)}, and variance inflation factors (VIF).

Table 3: Reflective and Formative Latent Variable Coefficients (PLSR)

Job performance (reflective)						**Electronic communication use (formative)**					
	L	**P(L)**	**W**	**P(W)**	**VIF**		**L**	**P(L)**	**W**	**P(W)**	**VIF**
Prf1	.885	<.001	.339	<.001	2.378	Ecu1	.750	<.001	.504	<.001	1.142
Prf2	.958	<.001	.367	<.001	6.921	Ecu2	.723	<.001	.490	<.001	1.129
Prf3	.953	<.001	.365	<.001	6.645	Ecu3	.627	<.001	.427	<.001	1.070
Mean	*.932*		*.357*		*5.315*	*Mean*	*.700*		*.474*		*1.114*

As can be seen, even though mode B was not used in calculations related to the formative latent variable *electronic communication use*, its manifest variables had statistically significant weights, indicated by P values lower than .001; and presented a low level of redundancy among them, indicated by fairly low individual variance inflation factors (approaching the minimum possible value of 1) and also a low average variance inflation factor of 1.114. The average variance inflation factor for the reflective latent variable *job performance* was a much higher 5.315, suggesting redundancy (Hair et al., 2009; Kline, 1998), as expected. This was a consequence of the individual variance inflation factors being much higher than those for the formative latent variable, contributing to that higher average variance inflation factor. Also as expected, the loadings for the reflective latent variable were generally higher (.932 on average), than the loadings for the formative latent variable (.700 on average). Mode B however, seems to have led to instability in the weights, causing instances of Simpson's paradox (Pearl, 2009; Wagner, 1982) at the latent variable level. This was characterized by one manifest variable detracting from the variance explained in the latent variable to which

it was assigned, as demonstrated in the previous analyses under the section covering the two illustrative studies. Interestingly, this occurred for both reflective (*job performance*) and formative (*electronic communication use*) latent variables.

The above findings related to PLS Mode B, however, have a possible silver lining. One can reasonably argue that Mode B is more sensitive to Simpson's paradox. The most widely held interpretation of Simpson's paradox is that it indicates that a link is either implausible or reversed (Pearl, 2009; Wagner, 1982). Given this, Mode B could be useful in the identification of indicators that should be removed from a formative latent variable; or that are predicted by a latent variable, as opposed to predicting it. That is, Mode B can be useful in more strict assessments of formative latent variable measurement. These would have to be conducted in combination with other tests, such as indicator collinearity tests, because Simpson's paradox can also be caused by collinearity (Kock & Lynn, 2012). These tests would have to also be conducted in the context of specific models, because, as noted by Tenenhaus et al. (2005) and is clear from the algorithms discussed earlier, without the inner model one would not be able to estimate weights in PLS Mode B. Even if one were to modify Mode B's algorithm, so as to skip the algorithmic step where inner model coefficients are estimated, all weights would be identical in this modified version of Mode B. This is why PLS regression does not have a "formative mode" per se. The above discussion lead us to recommend that researchers should consider employing PLS Mode B in strict assessments of formative latent variable measurement, in the context of specific models, and in combination with other tests. This would arguably be particularly important when latent variables are designed as formative but behave as reflective under PLS regression or PLS Mode A, or are designed as reflective and behave as formative under those algorithms.

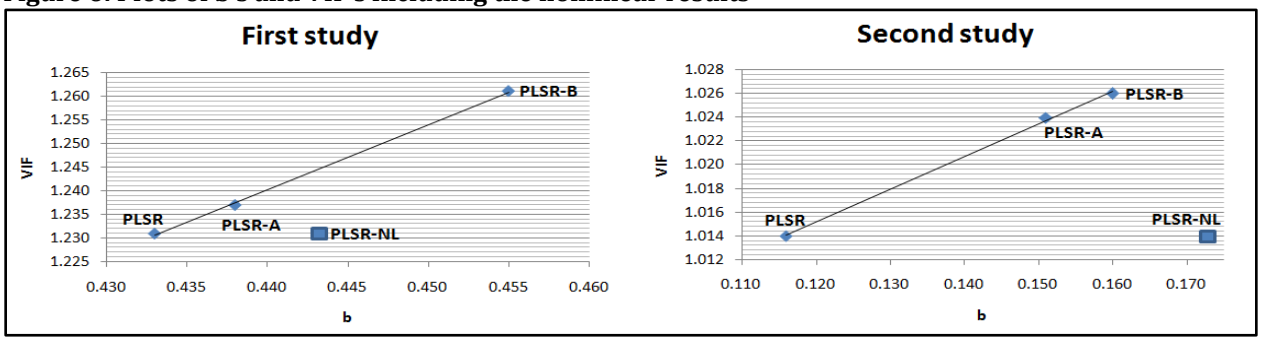
Nonlinear PLS Regression versus PLS Modes A and B: PLS-based SEM tends to yield inner model coefficients of association that are generally lower than those generated by the classic covariance-based approaches to SEM (Goodhue et al., 2012; Haenlein & Kaplan, 2004). Could the GNA have been an attempt to "correct" this situation? In other words, could the lower coefficients justify the use of an algorithmic assumption, such as the GNA, that would slightly inflate inner model coefficients of association in PLS-based SEM, and thus "level the playing field" somewhat? It is difficult to argue in favor a "yes" answer to the question above, unless the algorithmic assumption captured something that would increase inner model coefficients of association based on justifiable reasons. This "something" could be nonlinearity among latent variables that are linked. The vast majority of multivariate statistical analysis methods estimate coefficients of association between variables, including latent variables, assuming relationships to be linear (Bauer et al., 2012; Hair et al., 2009). Even in moderating effects analyses, where moderator variables may be seen as sources of nonlinearity, the influence of the moderating variables is normally modeled as linear (Arnold, 1982; Carte & Russel, 2003; Chin et al., 2003). Yet, most relationships in natural and behavioral phenomena are nonlinear (Giaquinta, 2009; Kaiser, 2010; Masujima, 2009). As mentioned before in this section, the GNA might have been originally an attempt at capturing any existing nonlinearity between predictor-criterion latent variables. However, the discussion below suggests that the ideas that Mode A and/or Mode B meaningfully contribute to capturing nonlinearity are questionable.

We conducted a nonlinear analysis with the multivariate statistical analysis software tool WarpPLS, where the algorithms "PLS regression" and "Warp3" were chosen (Kock, 2013). These algorithms conduct the analysis through two main sub-algorithmic phases. The first phase is an implementation of the PLS regression algorithm, as describe here, through which outer model coefficients are calculated without letting the inner model influence them. The second phase consists in the calculation of standardized inner model coefficients of association (b_1, b_2 … b_k). This is conducted by identifying functions f_1, f_2 … f_k, from a finite pool of functions, that best approximate the distributions of points among linked predictor ($\hat{Y}_1$, $\hat{Y}_2$ … $\hat{Y}_k$) and criterion ($\hat{F}$) variables in each latent variable block, and then solving the matrix equation below. The finite pool of functions include common functions such as logarithmic, hyperbolic decay, exponential decay, exponential, quadratic, as well as more complex functions that are derived from these common functions through single indefinite integration (Kock, 2013).

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} f_1(\hat{Y}_1) \\ f_2(\hat{Y}_2) \\ \vdots \\ f_k(\hat{Y}_k) \end{bmatrix} \backslash \hat{F}$$

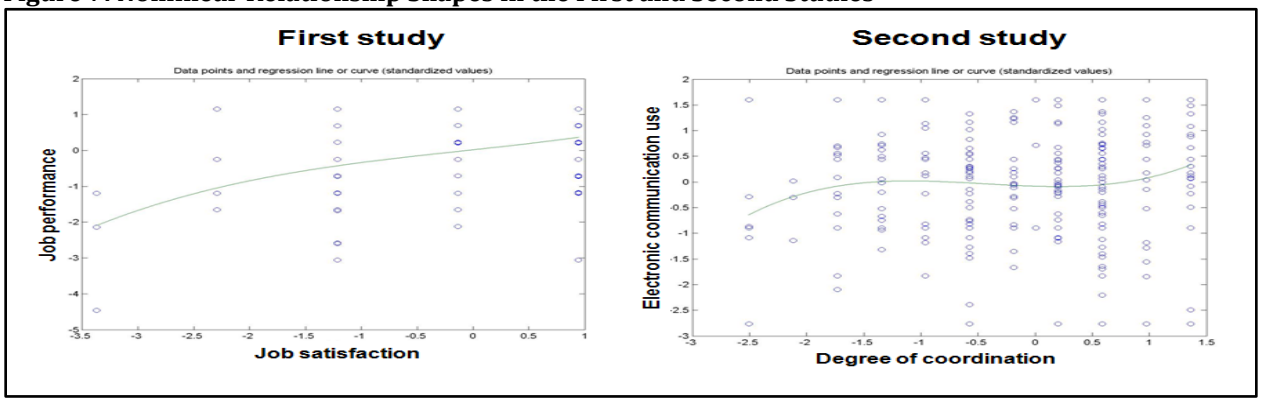
When the nonlinear algorithm (i.e., Warp3) was employed in combination with the PLS regression algorithm, the resulting collinearity and association coefficients in the first study were 1.231 and .443, respectively. In the second study these coefficients were 1.014 and .172. The coefficients of association between latent variables are standardized partial regression coefficients, as in the previous illustrative analyses. Figure 6 show these results (indicated as PLSR-NL) vis-à-vis the previous results obtained with strictly linear analyses employing the PLS regression algorithm (indicated as PLSR) and PLS modes A and B (indicated as PLSR-A and PLSR-B, respectively).

Figure 6: Plots of b's and VIF's including the nonlinear results



It is clear from these results that the coefficients of association increased significantly when nonlinearity was taken into consideration in their calculation, in conjunction with the use of the PLS regression algorithm, and this occurred without any increase in collinearity levels. In the second study, the nonlinear PLS regression algorithm coefficient of association even exceeded that resulting from the use of linear Mode B, still without any increase in collinearity levels compared with the linear PLS regression algorithm. As it can be seen from the graphs, the relative increase in the coefficient of association between the latent variables was significantly higher in the second study, pushing the coefficient beyond those obtained via modes A and B. In the first study, the increase was not as marked, with the nonlinear coefficient falling in between those from modes A and B. A possible underlying reason for this is illustrated in Figure 7, where the degrees of nonlinearity in the first and second studies appear to be markedly different. The degrees of nonlinearity can be seen as the degrees to which the shapes of the curves on the graphs deviate from that of a straight line. In the second study, the shape of the curve describing the relationship between the latent variables appears to deviate more from that of a straight line than in the first study.

Figure 7: Nonlinear Relationship Shapes in the First and Second Studies



It is conceivable that the GNA is the reason why modes A and B are said to be based on a *nonlinear* iterative partial least squares algorithm (Wold, 1974). That is, what has been described by seminal authors such as

Wold (1974) and Lohmöller (1989) as an attempt to ensure that linked latent variables were "good neighbors" might have well been an attempt at capturing nonlinearity among linked latent variables. While the nonlinear algorithm employed above is only one of many possible such algorithms (see, e.g., Bauer et al., 2012; Wold et al., 2001), the results suggest that the GNA may not per se be the best approach to capture nonlinearity. Based on the above, we contend that researchers interested in capturing nonlinearity among linked latent variables in the calculation of coefficients of association should favor the use of PLS regression in combination with algorithms that identify common nonlinear functions, instead of using algorithms that incorporate the GNA. However, further methodological research (and software implementations) should be conducted to explore the possibility that GNA-based algorithms such as PLS Mode A and Mode B could help in the identification and modeling of nonlinearity when used in conjunction with algorithms that discover best-fitting nonlinear functions associated with predictor-criterion latent variable pairs.

Limitations and Future Research: The formative-reflective measurement dichotomy is intimately related to a characteristic shared by the PLS-based SEM algorithms discussed here. These algorithms generate approximations of factors via exact linear combinations of indicators, without explicitly modeling measurement error. Recently new PLS-based SEM algorithms have been proposed that explicitly model measurement error (Kock, 2014). These new algorithms suggest that formative and reflective latent variables may be conceptually the same, but at the ends of a reliability scale, where reliability can be measured through the Cronbach's alpha coefficient (Cronbach, 1951; Kline, 2010). That is, a properly designed formative latent variable would typically have a lower reliability than a properly designed reflective latent variable. Nevertheless, both reliabilities would have to satisfy the same criterion – be above a certain threshold (e.g., .7). While reflective latent variables can achieve high reliabilities with few indicators (e.g., 3), formative latent variables require more indicators (e.g., 10). This mathematical property is in fact consistent with formative measurement theory, where many different facets of the same construct should be measured so that the corresponding formative latent variable can be seen as a complete depiction of the underlying formative construct. Future research opportunities stem from the above discussion, leading to important methodological questions. What is the best measure of reliability to be used? It is possible that the composite reliability coefficient (Kock, 2013) is a better choice than the Cronbach's alpha coefficient. Will the new PLS-based SEM algorithms that explicitly model measurement error (Kock, 2014) obviate the need for the classic algorithms discussed here, or will the new algorithms have a more limited scope of applicability? Will formative measurement be re-conceptualized as being at the low end of a reliability scale that also includes reflective measurement, providing a unified view of what could be seen as an artificial dichotomy? These and other related methodological questions give a glimpse of the exciting future of PLS-based SEM.

5. Conclusion

The preceding sections provided a thorough discussion of the treatment of reflective and formative latent variables in the context of three widely used PLS-based SEM algorithms; namely PLS regression, PLS Mode A, and PLS Mode B. Two illustrative examples, based on actual data, highlight some of the key outcomes, including pros and cons, which the use of modes A and B may have on the results of PLS-based SEM analyses, especially when compared with PLS regression. It is shown that the under the GNA (the "good neighbor" assumption underlying PLS modes A and B), the inner model influences the outer model in such a way as to increase inner model coefficients of association; this is done through a small increase in collinearity levels, thus making type I errors more likely. Viewed from a benign perspective, this would suggest that modes A and B should be seen as exploratory, while the PLS regression algorithm should be seen as confirmatory. The GNA also leads to the need for separate confirmatory factor analyses to be conducted for each configuration of a model with the same latent variables but different inner model structures. In addition, evidence is presented suggesting that Mode B has the potential to cause instability in the calculation of weights in PLS-based SEM analyses, leading to instances of Simpson's paradox at the outer model (a.k.a. measurement model). This paradox is characterized by one manifest variable detracting from the variance explained in the latent variable to which it is assigned.

Finally, a nonlinear analysis suggests that, while it is conceivable that the GNA is the reason why modes A and B are said to be *nonlinear* algorithms, based on the underlying assumption that the GNA helps capture

nonlinearity, the results suggest that the GNA may not be the best single approach to capture nonlinearity. As illustrated through a nonlinear analysis, a better alternative may well be to conduct PLS-based SEM analyses in two stages. The first stage would entail the use of the PLS regression algorithm for the calculation of latent variable scores. The second stage would entail the use of algorithms that discover best-fitting nonlinear functions associated with predictor-criterion latent variable pairs, and then calculate multivariate coefficients of association (path coefficients, a.k.a. beta coefficients in PLS-based SEM) based on the application of those functions to the predictor latent variables. The preceding discussion does not mean that modes A and B should be completely abandoned, as these modes may have interesting applications that are yet to be explored. For instance, a possibility that should be explored is that Mode A could help in the identification and modeling of nonlinearity when used in conjunction with algorithms that discover best-fitting nonlinear functions associated with predictor-criterion latent variable pairs. Even Mode B's inherent instability may have interesting applications. For example, the occurrence of Simpson's paradox may in some cases be associated with the existence of improbable causal relationships (Pearl, 2009; Wagner, 1982). In other words, if under Mode B a manifest variable is found to make a negative contribution to the explained variance in its latent variable, to which it points, then maybe either the causal link should be removed or its direction changed (Pearl, 2009). This opens the door for the use of Mode B in more rigorous validity assessments of formative latent variables. It may well be that unstable weights are a sign that formative latent variables are not properly designed, or that the inner model structure influencing them contains improbable causal relationships, or both.

References

Adelman, I. & Lohmoller, J. B. (1994). Institutions and development in the nineteenth century: A latent variable regression model. *Structural Change and Economic Dynamics*, 5(2), 329-359.

Arnold, H. J. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organization Behavior and Human Performance*, 29(4), 143-174.

Baskerville, R. L. & Myers, M. D. (2002). Information systems as a reference discipline. *MIS Quarterly*, 26(1), 1-14.

Bauer, D. J., Baldasaro, R. E. & Gottfredson, N. C. (2012). Diagnostic procedures for detecting nonlinear relationships between latent variables. *Structural Equation Modeling*, 19(2), 157-177.

Carte, T. A. & Russel, C. J. (2003). In pursuit of moderation: Nine common errors and their solutions. *MIS Quarterly*, 27(3), 479-501.

Cenfetelli, R. & Bassellier, G. (2009). Interpretation of formative measurement in information systems research. *MIS Quarterly*, 33(4), 689-708.

Chatelin, Y. M., Vinzi, V. E. & Tenenhaus, M. (2002). State-of-art on PLS path modeling through the available software. Storrs, CT: Department of Economics, University of Connecticut.

Chin, W. W. (1998). Issues and opinion on structural equation modeling. *MIS Quarterly*, 22(1), 7-16.

Chin, W. W., Marcolin, B. L. & Newsted, P. R. (2003). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study. *Information Systems Research*, 14(2), 189-218.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Davenport, T. H. (2000). Mission critical: Realizing the promise of enterprise systems. Boston, MA: Harvard Business School Press.

Diamantopoulos, A. (1999). Export performance measurement: Reflective versus formative indicators. *International Marketing Review*, 16(6), 444-457.

Diamantopoulos, A. & Siguaw, J. A. (2002). Formative vs. reflective indicators in measure development: Does the choice of indicators matter? Ithaca, NY: School of Hotel Administration, Cornell University.

Diamantopoulos, A. & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management*, 17(4), 263–282.

Dos Santos, B. L. & Peffers, K. (1995). Rewards to investors in innovative information technology applications: First movers and early followers in ATMs. *Organization Science*, 6(3), 241-259.

Duncan, O. D. (1966). Path analysis: Sociological examples. *The American Journal of Sociology*, 72(1), 1-16.

Ehremberg, A. S. C. & Goodhart, G. J. (1976). Factor analysis: Limitations and alternatives. Cambridge, MA: Marketing Science Institute.
Fornell, C. & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research,* 18(1), 39-50.
Gefen, D., Straub, D. W. & Boudreau, M. C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the AIS*, 4(7), 1-76.
Gerlach, R. W., Kowalski, B. R. & Wold, H. (1979). Partial least-squares path modeling with latent variables. *Analytica Chimica Acta*, 112(4), 417-421.
Giaquinta, M. (2009). Mathematical analysis: An introduction to functions of several variables. New York, NY: Springer.
Goodhue, D. L., Lewis, W. & Thompson, R. (2012). Does PLS have advantages for small sample size or non-normal data? *MIS Quarterly*, 36(3), 981-1001.
Grover, V., Gokhale, R., Lim, J., Coffey, J. & Ayyagari, R. (2006). A citation analysis of the evolution and state of information systems within a constellation of reference disciplines. *Journal of the Association for Information Systems*, 7(5), 270-325.
Guo, K. H., Yuan, Y., Archer, N. P. & Connelly, C. E. (2011). Understanding no malicious security violations in the workplace: A composite behavior model. *Journal of Management Information Systems*, 28(2), 203-236.
Haenlein, M. & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics*, 3(4), 283-297.
Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. (2009). Multivariate data analysis. Upper Saddle River, NJ: Prentice Hall.
Hardin, A. M., Chang, J. & Fuller, M. A. (2008). Formative vs. reflective measurement: Comment on Marakas, Johnson, and Clay (2007). *Journal of the Association for Information Systems*, 9(9), 519-534.
Harville, D. A. (1997). Matrix algebra from a statistician's perspective. Springer-Verlag: New York, NY.
Kaiser, H. M. (2010). Mathematical programming for agricultural, environmental, and resource economics. Hoboken, NJ: Wiley.
Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York, NY: The Guilford Press.
Kock, N. (2005). Business process improvement through e-collaboration: Knowledge sharing through the use of virtual groups. Hershey, PA: Idea Group Publishing.
Kock, N. (2008). Encyclopedia of e-collaboration. Hershey, PA: Information Science Reference.
Kock, N. (2013). *WarpPLS 4.0 User Manual*. Laredo, Texas: ScriptWarp Systems.
Kock, N. (2014). *A note on how to conduct a factor-based PLS-SEM analysis*. Laredo, TX: ScriptWarp Systems.
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.
Martin, M. A. (2007). Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties. *Computational Statistics & Data Analysis*, 51(12), 6321-6342.
Masujima, M. (2009). Applied mathematical methods in theoretical physics. Weinheim. Germany: Wiley-VCH.
Mueller, R. O. (1996). Basic principles of structural equation modeling. New York, NY: Springer.
Nevitt, J. & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, 8(3), 353-377.
Noonan, R. & Wold, H. (1983). Evaluating school systems using partial least squares. *Evaluation in Education*, 7(3), 219-364.
Nunnaly, J. (1978). Psychometric Theory. New York, NY: McGraw Hill.
Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric theory. New York, NY: McGraw-Hill.
Pearl, J. (2009). Causality: Models, reasoning, and inference. Cambridge, England: Cambridge University Press.
Penney, R. C. (2004). Linear algebra: Ideas and applications. John Wiley & Sons: Hooboken, NJ.
Petter, S., Straub, D. & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly*, 31(4), 623-656.

Rao, C. R. & Mitra, S. K. (1971). Generalized inverse of matrices and its applications. New York, NY: Wiley.
Recker, J. & La Rosa, M. (2012). Understanding user differences in open-source workflow management system usage intentions. *Information Systems*, 37(3), 200-212.
Rencher, A. C. (1998). Multivariate statistical inference and applications. New York, NY: John Wiley & Sons.
Rosenthal, R. & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis. Boston, MA: McGraw Hill.
Roy, S. R. (2012). Digital mastery: The skills needed for effective virtual leadership. *International Journal of e-Collaboration*, 8(3), 58-68.
Schumacker, R. E. & Lomax, R. G. (1996). A beginner's guide to structural equation modeling. Mahwah, NJ: Lawrence Erlbaum.
Subhadip, R. (2008). Measuring formative constructs in management research: Definitions, distinctions and measurement. *ICFAI Journal of Management Research*, 7(1), 18-34.
Temme, D., Kreis, H. & Hildebrandt, L. (2006). PLS path modeling – A software review. Berlin, Germany: Institute of Marketing, Humboldt University Berlin.
Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M. & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, 48(1), 159-205.
Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. Washington, DC: American Psychological Association.
Vittadini, G., Minotti, S. C., Fattore, M. & Lovaglio, P. G. (2007). On the relationships among latent variables and residuals in PLS path modeling: The formative-reflective scheme. *Computational Statistics & Data Analysis*, 51(12), 5828-5846.
Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician*, 36(1), 46–48.
Walker, H. M. (1940). Degrees of freedom. *Journal of Educational Psychology*, 31(4), 253–269.
Wold, H. (1974). Causal flows with latent variables: Partings of the ways in the light of NIPALS modeling. *European Economic Review*, 5(1), 67-86.
Wold, S., Trygg, J., Berglund, A. & Antti, H. (2001). Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 131-150.
Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3), 161-215.
Wright, S. (1960). Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics*, 16(2), 189-202.