DOI: 10.20472/IAC.2017.030.015

GLORIA GHENO

Free University of Bozen-Bolzano, Italy

A NEW SEMIPARAMETRIC APPROACH FOR MEDIATION ANALYSES

Abstract:

In economics it is very important to understand the mechanisms which determine the phenomena of interest. For example in marketing it is very important to analyze the factors which determine the future choices of the customer such that it is possible to influence the sales of a product. In most of the economic applications, to make this the researcher supposes the knowledge of the mathematical relations among the variables of interest and in particular these relations are supposed linear. In this work I propose a new estimation method for mediation models without supposing the form of the mathematical relations but aggregating many different equations. My new estimation method derives by Bauer's work (2005) about the semiparametric approach for SEM (structural equation models). To apply my estimation method, I propose new formulas to calculate the direct, indirect and total effects. I apply my method both to simulated data and to marketing data.

Keywords:

direct effect, marketing, mediation, indirect effect, SEM, semiparametric approach, total effect

JEL Classification: C40, C14, C15

Introduction

The analysis of the relationships among variables is very useful in business and in many other sectors to understand the mechanism which regulates and causes the phenomena which are object of study. Among many statistical methods proposed to investigate these relationships I mention the SEM (structural equation model) method, which has the advantage of being able to study the relationships both among directly observed variables and among unobserved variables, but derivable from other variables. To this end, the SEM incorporates various statistical concepts such as confirmatory factor analysis, path analysis, multiple regression, ANOVA and simultaneous equation models. The SEM shows a structural part, which defines the direct relationships among unobserved variables, and a measurement part, in which the unobserved variables, or latent, are derived from observed variables, called indicators (Bollen, 1989). The measurement part originates from the explanatory factor analysis (Sperman, 1904), while the structural part is linked to the causal diagrams (Wright, 1921), in which a diagram is proposed to explain the causal relationships among variables. The two parts were unified in the seventies by Jöreskog, Keesling and Wiley.

Until the middle ‘80s, the structural part of the SEM considered only the linear-in-parameters and linear-in-variables functional form. Subsequently Kenny and Judd’s paper (1984) started to analyze nonlinear relationships among variables. Since then, many researchers have studied this problem proposing new estimation method (as Bollen and Paxton, 1998; Hensler and Chin, 2010) and different types of indicators for the non-linear part (such as Ping, 1995; Jöreskog and Yang, 1996; Marsh et al., 2004). Moosbrugger and Klein (2000) proposed a new method which does not require the specification of indicators for the nonlinear part. All these works, however, require the knowledge, or the supposed knowledge of the functional form. Bauer (2005) suggested, however, the use of the finite mixture structural equation model (SEMM) to examine a nonlinear SEM, but without defining the functional form. His method is defined semi-parametric approach because it uses a particular combination of parameters to find the functional forms which are not specified a priori. The SEMM (Jedidi, Jagpal and DeSarbo, 1997), unlike the SEM, assumes that the data come from different classes and estimates, for each of these classes, a different model so as to can find the unobserved heterogeneity which could lead to biased estimates if it is not considered. The following simplified example clarifies the difference between these two methods (SEM, SEMM). Analyzing how education affects income in different states, the SEM considers all together, while the SEMM aggregates all the more resemble states in different classes and then estimates a different model for each class. The method proposed by Bauer has the advantage of not specifying the functional form, as for example that the variable age influences parabolically the variable income.

The SEM, analyzing multiple relationships simultaneously, can specify simple or complex mediation models. A mediation model considers that the effect of one variable on another variable is due to other variables, defined mediators (Hayes and Preacher, 2010; Hayes, 2013; Hayes and Preacher, 2014). The nonlinearity in the SEM is mainly evaluated in models without mediation. Only Coenders et al. (2008), Chen and Cheng (2014) and Gheno (2015, 2016) have considered models in which the nonlinearity occurs in the mediation, specifically examining the interaction between two mediators. The interaction, mathematically represented by the product of two variables, is a nonlinear effect implying that the influence of a variable depends on the value of another variable, and vice versa. In the mediation models, however, the estimation method so far used for the nonlinearity has always been parametric.

The analysis of causality, and therefore of the effects of one variable on another variable, is more complicated in nonlinear mediation models, requiring the use of special methods to calculate the effects. Using Pearl's definitions of the effects (2001, 2009, 2012, 2014), Gheno (2015, 2016) proposed new formulas to calculate them in models with two mediators and interaction. Unlike the theory proposed by Pearl, where it is possible to measure always the effects only for uncorrelated mediators, these new formulas allow to calculate them in many models even if the mediators are correlated.

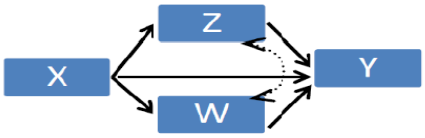
In this work I propose a new semi-parametric estimation method for nonlinear mediation models. I modified this method, deriving by approach of Bauer (2005), so that it is applicable to mediation models. The causal analysis plays a fundamental role in the analysis of mediation models, whatever type of estimate is used and thus I propose to use a modification of the formulas proposed by Gheno (2015, 2016) to calculate the effects using the parameters obtained by this new method estimation. This paper, therefore, before explains Bauer's method (2005) with my changes, then shows the application in simulated data in order to verify the goodness of the method being possible to compare the real values with the estimated ones, and in conclusion it presents a practical application in the field of marketing.

Finite mixture structural equation model as semi-parametric approach

The parametric methods require a functional form, specified a priori, both of linear-in-variables relationships and of the nonlinear-in-variables ones. Bauer (2005) proposes a semi-parametric approach which does not require the specification of the functional for a priori and then he uses the SEMM method (mixture structural equation model). In the traditional parametric SEM, the researcher has to decide at the time of estimation the influence of the variable X on the variable Y, for example he can put $Y = \beta X^2$. In the method proposed by Bauer this choice a priori is not required and therefore the same estimation process defines the functional relationship between Y and X, i.e. $Y = f(X)$. In the traditional parametric SEMM, the parameters of the measurement part and

those of the structural part are different in each class in order to catch the heterogeneity, but the functional form has to be given a priori yet. Bauer's method, however, considers different in each class only the parameters of the structural part not wanting to catching this heterogeneity, which must be expressed a priori, but only the functional form. In a simple model, in which only the unobserved

Figure 1: parallel multiple mediators



Source:Own path diagrams

variable X affects the unobserved variable Y, the relationship between the two variables is

$$Y = \alpha_{2k} + \beta_k X + e_k \quad (1)$$

where e_k represents the error term which is normally distributed with mean 0 and variance φ_{2k} and the subscript k identifies the relationship in the class k, consequently the relationship between the two variables is linear for each class k. The expected value of the total relationship between the two unobserved variables is obtained by aggregating the formula (1):

$$E(Y|X) = \sum_{k=1}^{K} P(k|X) E_k(Y|X) = \sum_{k=1}^{K} P(k|X)\,(\alpha_{2k} + \beta_k X) \quad (2)$$

where P (k | X) is the probability of being in class k given the observed value of the variable X. This probability is calculated as follows

$$P(k|X) = \frac{P(k)\phi_k(X, \alpha_{1k}, \varphi_{1k})}{\sum_{k=1}^{K} P(k)\phi_k(X, \alpha_{1k}, \varphi_{1k})} \quad (3)$$

where $\phi_k(X, \alpha_{1k}, \varphi_{1k})$ represents the distribution of the normal variable X with mean α_{1k} and variance φ_{1k}. I note that the mean and the variance of variable X are different

in each class, as well as the variance of the error term and of the variable Y. Through the aggregation (2), Bauer's method is able to insert the nonlinearity in the relationship between the two variables obtaining the true functional form. Bauer (2005) recommended to fix $\varphi_{1k} = \varphi_1$.

Semi-parametric approach for mediation model

In this paper I propose to adapt the method proposed by Bauer (2005) to a mediation model as that represented in Fig.1, because so far a nonparametric method has never been applied to complex models. The variable X affects the variable Y both directly and indirectly through the mediators Z and W, which are correlated. In the graph (Fig.1), the unidirectional arrows represent the direct effects, while the bidirectional arrow the covariances. The measurement part, where the unobserved variables are obtained by the observed ones, remains constant for K classes:

$$\begin{aligned} X_i &= \lambda_{ix} X + \varepsilon_{ix} \quad i = 1,2,3 \\ Z_i &= \lambda_{iz} Z + \varepsilon_{iz} \quad i = 1,2,3 \\ W_i &= \lambda_{iw} W + \varepsilon_{iw} \quad i = 1,2,3 \\ Y_i &= \lambda_{iy} Y + \varepsilon_{iy} \quad i = 1,2,3 \end{aligned} \tag{4}$$

The unobserved variables X, Z, W and Y, defined latent, are obtained by the observed variables X_i, Z_i, W_i, Y_i, defined indicators. The errors ε are uncorrelated with each other and with the latent variables. The structural part, however, is different in each class k

$$\begin{aligned} Y_k &= \alpha_{ky} + \beta_{kx} X + \beta_{kw} W + \beta_{kz} Z + e_{yk} \\ Z_k &= \alpha_{kz} + \gamma_{kz} X + e_{zk} \\ W_k &= \alpha_{kw} + \gamma_{kw} X + e_{wk} \end{aligned} \tag{5}$$

where the errors e_{lk} with l = y, w, z are normally distributed with mean 0 and variance φ_{lk}. The error e_{wk} is correlated with the error e_{zk} and both are uncorrelated with e_{yk}. To get a better estimate I advice keeping fixed the variances of the errors of the mediators Z and W, i.e. $\varphi_{wk} = \varphi_w$ and $\varphi_{zk} = \varphi_z$, and allowing the covariance to vary.

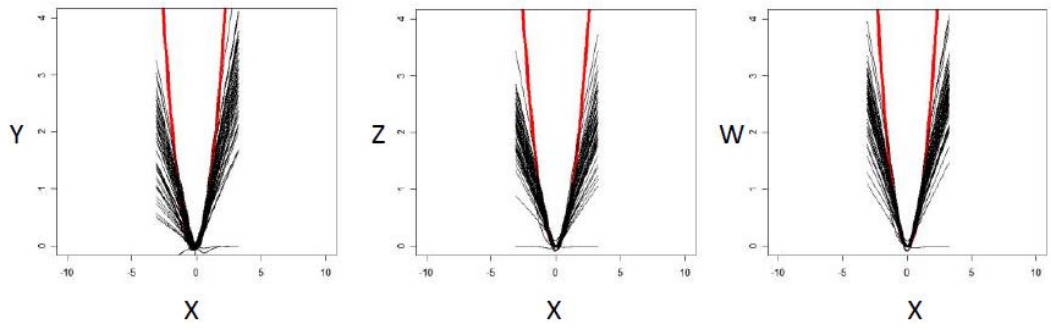
To calculate how the variable X influences the variable Y both directly and indirectly, i.e. through other variables called mediators, I use a modification of the formulas proposed by Gheno (2015, 2016) who, using Pearl's causal theory (2001,2009,2012, 2014), provided formulas to measure the effects, in the mediation models, when the errors are correlated and K = 1. The direct effect (DE), indirect (IE) and total (TE) for the change of the variable X from x to x ' are

$$DE_{x,x'} = \sum_{e_z,e_w} \sum_{Z,W} [E(Y|x',z,w) - E(Y|x,z,w)]P(Z,W|x,e_z,e_w)P(e_z,e_w)$$
$$IE_{x,x'} = \sum_{e_z,e_w} \sum_{Z,W} E(Y|x,z,w)[P(Z,W|x',e_z,e_w) - P(Z,W|x,e_z,e_w)]\,P(e_z,e_w) \quad (6)$$
$$TE_{x,x'} = DE_{x,x'} - IE_{x'x}$$

From (6) I can see that the total effect $TE_{x,x'}$ is equal to the direct effect of the change from x to x' minus the indirect effect of the change from x' to x. With a reworking of these formulas, I get the effects for a model with two mediators and with K classes:

$$DE_{x,x'} = \sum_k \beta_{kx} x' P(k|x') - \beta_{kx} x P(k|x)$$
$$IE_{x,x'} = \sum_k [\beta_{kz}(\alpha_{kz} + \gamma_{kz}x') + \beta_{kw}(\alpha_{kw} + \gamma_{kw}x')]P(k|x') \quad (7)$$
$$-[\beta_{kz}(\alpha_{kz} + \gamma_{kz}x) + \beta_{kw}(\alpha_{kw} + \gamma_{kw}x)]P(k|x)$$

Figure 2: comparison between the true relations and the estimated relations



Source: Simulated data

The sum of the direct and indirect effects, both of the change from x to x', gives the total effect of the same variation. It is not possible to decompose the indirect effect depending on the mediators, because with this method it is possible also to find the interaction effect between the mediators without having to specify it. In these formulas the intercepts α_{kz} and α_{kw} are also included, although they should not create a causal effect. The introduction of the intercepts can be justified remembering that the estimated model does not represent the true model and each parameter can be important to represent the nonlinearity through the weighted sum of the linear models. The intercept of the Y, α_{ky} , is not considered because, being estimated different from 0 especially in models with interaction and correlation between the errors of the two mediators, is a measure of the covariance between the errors of mediators, which is

not causal. Recalling that in a model with K = 1 and $\alpha_{kz} = \alpha_{kw} = 0$ the expected value of the interaction ZW, i.e. the product of the two mediators, depends on the covariance between e_z and e_w, the relation between the intercept and the covariance is clarified .

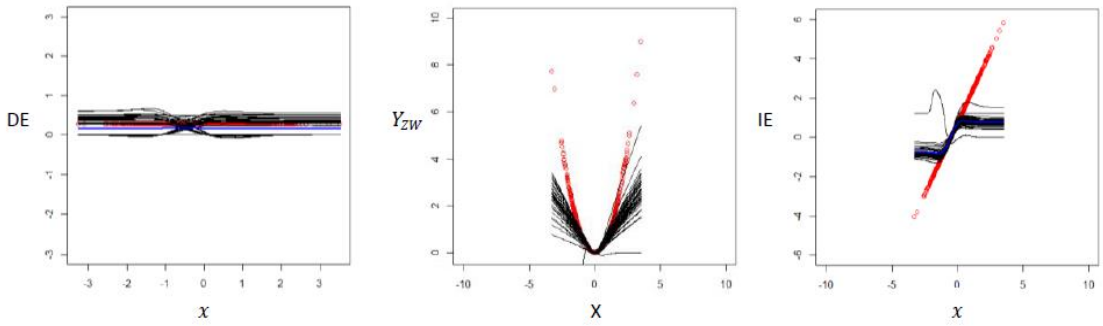
Numerical studies: simulations and real dataset

To test the goodness of my semi-parametric estimation method, I use several simulated datasets so as to can compare the estimated model with the real one. I submit to the analysis 50 datasets of sample size equal to 1000 generated by a same mediation model, such as that shown in Fig. 1, in which the direct effect of the variable X on the variable Y influences relatively little the total effect and the variables X affects parabolically the variables Z and W, i.e.

$$\begin{aligned} Y &= 0.27\,X + 0.57\,W + 0.45\,Z + e_y \\ Z &= 0.63\,X^2 + e_z \\ W &= 0.77\,X^2 + e_w \end{aligned} \tag{8}$$

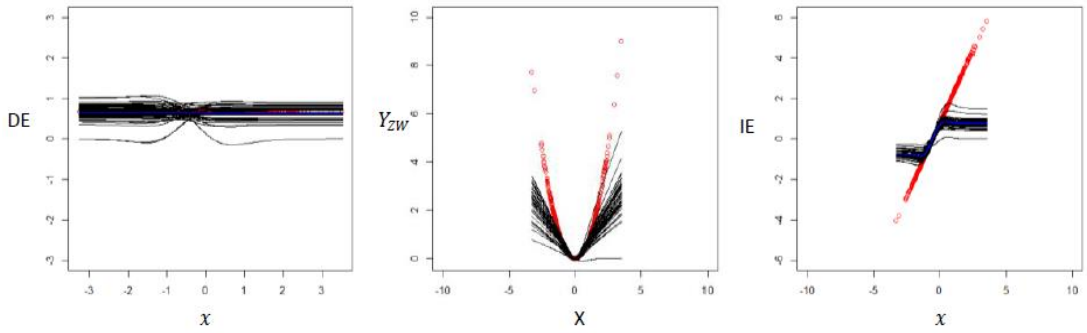
where the errors e_y, e_z and e_w are marginally distributed as a normal with mean zero and variances equal to 0.52, 0.48 and 0.58. The error e_y is uncorrelated with e_z and e_w, which in turn are correlated with covariance equal to 0.4. The comparison between

Figure 3: comparison between the true and the estimated effects, $\boldsymbol{\beta_x = 0.27}$



Source: Simulated data

Figure 4: comparison between the true and the estimated effects, $\beta_x = 0.67$



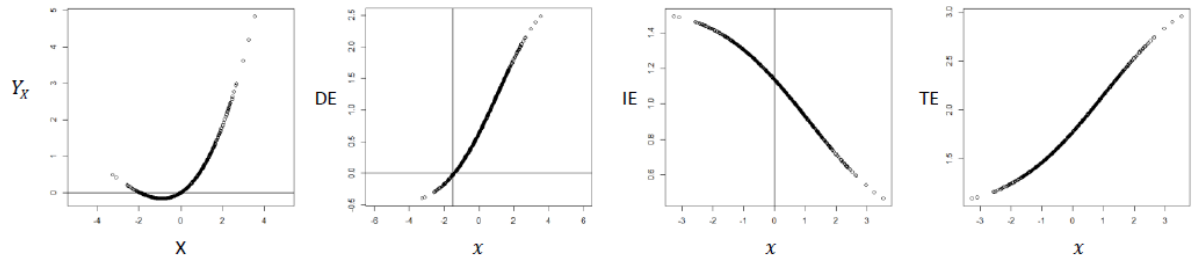
Source: Simulated data

the true relationships of X with Z, W and Y (red line) and the estimated ones with K = 2 in the 50 datasets (one black line for each dataset), is represented in Fig. 2. In most of these simulated datasets the estimated model finds the true parabolic trend and only in very few datasets "mistakenly" one branch of parabola. In this model the effect of X on Y takes place directly or through its effect on the mediators Z and W, then the system of equations (8) can be rewritten as

$$Y = \underbrace{0.27\,X}_{\substack{part\ due\ to\\ direct\ effect\\ Y_X}} + \underbrace{0.7224\,X^2}_{\substack{part\ due\ to\\ indirect\ effect\\ Y_{ZW}}} + \underbrace{0.57e_w + 0.45\,e_z + e_y}_{error\ part} \qquad (9)$$

The intercept α_{ky} is estimated different from 0 only in one dataset confirming the previous theory, which asserts that it is different from 0 when the true model has an interaction between the mediators and their errors are correlated. In the true model the direct effect of X on Y is equal to 0.27ΔX and the indirect effect is equal to 0.7224 ΔX^2 using the formulas of the direct and indirect effects proposed from Gheno (2015, 2016). Considering a change of X equal to 1, the comparison between the true direct effect (red line) and the estimated one in the 50 datasets (black lines) is represented in the first graph of Fig 3, where the direct effect is a function of the initial value x (i.e.

Figure 5: direct, indirect and total effects in real dataset



Source: Own data

$DE_{x,x+1}$). In some of these datasets the direct effect is 0, however, the average value in the 50 datasets (blue line) is very close to the real value. The indirect effect analysis is represented in the second and third graph of Fig. 3. The first graph shows the influence of X on Y only due to the mediation (Y_{ZW} = 0.7224 X²) and I note that in almost all the datasets the estimated model finds a parabolic curve; in the second graph the true indirect effect is compared with the estimated one, considering a unitary change of X (i.e. $IE_{x,x+1}$). I note that in the central part, most of the lines are superimposed demonstrating that the estimated model is able to find the true indirect effect. The average indirect effect (blue line) is superimposed on the true effect in the middle part.

To check how my approach estimates the direct effect, I simulate 50 datasets of sample size equal to 1000 from the model of the formula (8), replacing only the value 0.67 to the value 0.27 of the parameter of the variable X. The relation (9) then becomes

$$Y = \underbrace{0.67\,X}_{\substack{part\ due\ to \\ direct\ effect}} + \underbrace{0.7224\,X^2}_{\substack{part\ due\ to \\ indirect\ effect}} + \underbrace{0.57e_w + 0.45\,e_z + e_y}_{error\ part} \qquad (10)$$

Applying the formulas of the direct and indirect effects proposed by Gheno (2015, 2016) in the true model, the direct effect of X on Y is equal to 0.67ΔX and the indirect effect is equal to 0.7224 ΔX², resulting unchanged from the previous example. The comparison between the estimated effects in the 50 datasets and the estimated ones is shown in the graphs of Fig. 4. The first shows the comparison between the estimated direct effect and the true one considering ΔX = 1 (i.e. $DE_{x,x+1}$) and only three of the 50 datasets estimate the direct effect in a way not good. The average direct effect calculated in the 50 datasets (blue line) and the true value overlap almost exactly showing that when the direct effect weights mostly on the total effect, my estimation method is exact. The second graph of Fig. 4 compares the part of Y explained indirectly by X (0.7224X²) with the estimated one. The parabolic part of Y

due to the two mediators is properly found by the estimated model. The third graph of Fig. 4 represents the comparison between the estimated indirect effect and the true one considering a unitary change of X (i.e. $IE_{x,x+1}$) and shows that the estimated indirect effect in the 50 datasets (black lines) and the average (blue line) overlap almost exactly the real effect (red line) in the middle part. This my estimation method, therefore, is able to obtain the true trend even without knowing it a priori.

I apply my estimation method to analyze marketing data having shown that it can find well the functional form of the true model. The analyzed datasets are collected by interviewing 395 customers in many stores of a known chain of jewelers. I consider how the atmosphere affects the customer loyalty both directly and through the positive and negative emotions in the jewelers situated in town centres. The analyzed model is shown in Fig. 1, where the variable X represents the atmosphere, the variable Z the positive emotions, the variable W the absence of negative emotions and Y the loyalty. The four variables are not observed directly but each through three indicators which are centralized, i.e. the average value is set equal to 0. The direct and indirect effects are depicted in Fig. 5, which shows that the atmosphere in the store does not affect linearly loyalty both directly and indirectly through the emotions. The nonlinearity is demonstrated by the effects, both direct and indirect, which become functions of the atmosphere. The first graph underlines that the direct influence of the atmosphere on the loyalty has a parabolic trend. If a customer gives a value to the atmosphere above the average, an increase of the consideration of the atmosphere leads to a more than proportional increase of the loyalty. The same trend is observed if the customer assesses the atmosphere from -1 to 0. For values smaller than about -1, an increase of the goodness of the atmosphere leads to the lowering of the loyalty. This analysis is also shown in the second graph of Fig. 5, which represents the direct effect of the change of one unit for each initial value x. With initial values x less than -1.5, an increase of one unit of the value of the atmosphere leads to a lowering of the loyalty; for values more than -1.5, an increase of one unit of the value of the atmosphere leads to an increase of the loyalty. In the first graph the change of the sign occurs at the point -1, while in the second at the point -1.5, because in the second graph I analyze a unitary change. If I consider an infinitely small variation of X, the point at which there is the change of sign is -1 in both graphs. The third graph shows that for every initial value x, an increase of one unit leads to an increase of the loyalty, thus indirectly the atmosphere affects always positively. The total effect of the atmosphere on the loyalty, however, is positive and then, if a manager improves the atmosphere, he will increase the customers loyalty.

Conclusions

In this paper I propose a new semi-parametric method to estimate mediation models. Its originality lies in not having to specify in advance the functional form with which the variables affect other variables, because it is found from the estimation process, with the aggregation of the various linear functions. A practical problem for the use of my

method in a mediation model can be born from the analysis of the causal effects being unspecified the functional form, and then I propose a method to calculate them. To demonstrate the goodness of my method of estimation and calculation of the causal effects I use simulated datasets in order to compare the estimated values with the real ones with which I simulated. The result is very good, indeed the method recognizes the true functional form and the true causal effects. Some problems can arise from extreme values, having been aggregated only two classes to ensure that there are always global solutions, without additional constraints on the parameters. In conclusion I present an application of my method and the advantage of its use only in the field of marketing, but it can be used in all other economic fields.

References

BAUER, D. J. (2005). A semiparametric approach to modeling nonlinear relations among latent variables. *Structural Equation Modeling: A Multidisciplinary Journal* 12(4), pp 513-535

BOLLEN, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley and Sons

BOLLEN, K. A., & PAXTON, P. (1998). Interaction of latent variables in structural equation models. *Structural equation modeling: a multidisciplinary journal*, vol. 5(3), pp 267-293

COENDERS, G., BATISTA-FOGUET, J. M., SARIS, W. E. (2008). Simple, efficient and distribution-free approach to interaction effects in complex structural equation models, *Quality & Quantity*, vol. 42, pp 369-396

CHEN, S-P, and CHENG, C-P. (2014). Model specification for latent interactive and quadratic effects in matrix form, *Structural equation modeling: A Multidisciplinary Journal*, vol. 21, pp 94-101

GHENO, G. (2015). S*tructural equation models with interacting mediators: theory and empirical results*. Phd Thesis, University of Padua.

GHENO, G. (2016): Nonlinear SEM: comparison between endogenous and exogenous interaction, *International Journal of Development Research*, vol. 06(10), pp. 9850-9857

HAYES, A. F. (2013). *Introduction to mediation, moderation and conditional process analysis: a regression-based approach*, New York: The Guilford press.

HAYES, A. F., & PREACHER, K. J. (2010). Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. *Multivariate Behavioral Research* 45, pp 627-660

HAYES, A. F., & PREACHER, K. J. (2014). Statistical mediation analysis with multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology* 67, pp 451-470

HENSELER, J., and CHIN, W.W., (2010). A comparison of approaches for the analysis of interaction effects between latent variables using partial least squares path modeling, *Structural equation modeling: a multidisciplinary journal,* vol. 17(1), pp 82-109

KENNY, D. A. & JUDD, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables, *Psychological Bulletin,* vol 96, pp 201-210

JEDIDI,K., JAGPAL, H., & DESARBO, W. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, vol. 16(1), pp 39-59

JÖRESKONG, K. G., & YANG, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G. A. Marcoulides &R. E. Schumacker (Eds.) *Advanced structural equation modeling techniques* (pp. 57-88). Hillsdale, NJ: Lawrence Erlbaum.

MARSH, H. W.,WEN, Z. & HAU, K. T. (2004). Structural equation models of latent interact interactions: evaluation of alternative estimation strategies and indicator construction, *Psychological Methods*, 9, pp 275-300

MOOSBRUGGER, H., & KLEIN, A. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method, *Psychometrika*, vol. 65(4), pp 457-474

PEARL, J. (2001). Direct and indirect effects. In J. Breese and D. Koller ed. *Proceedings of seventeenth conference on uncertainty and artificial intelligence*, San Francisco: Morgan Kaufman, pp. 411-420

PEARL, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys* 3, pp. 96–146.

PEARL, J. (2012). The mediation formula: a guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, and L. Bernardinelli ed. *Causality: Statistical Perspectives and Applications*, Chirchester: John Wiley & Sons, pp. 151–179.

PEARL, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods* 19, pp. 459-481

PING, R. A. (1995). A parsimonious estimating technique for interaction and quadratic latent variables. *Journal of Marketing Research*, vol. 32, pp 336-347.

SPEARMAN, C. (1904). General intelligence, objectively determined and measured. *American journal of Psychology,* 15, pp 201-293

WRIGHT, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(7), pp. 557-585