Pak J. Statist.
2003 Vol. 19(2) pp 213 – 226

CUSTOMER SATISFACTION MEASUREMENT MODELS: GENERALISED MAXIMUM ENTROPY APPROACH

AMJAD. D. AL-NASSER
Department of Statistics, Faculty of Science
Yarmouk University, Irbid
Jordan
amjadn@yu.edu.jo

ABSTRACT

This paper presents the methodology of the Generalised Maximum Entropy (GME) approach for estimating linear models that contain latent variables such as customer satisfaction measurement models. The GME approach is a distribution free method and it provides better alternatives to the conventional method; Namely, Partial Least Squares (PLS), which used in the context of costumer satisfaction measurement. A simplified model that is used for the Swedish customer satis faction index (CSI) have been used to generate simulated data in order to study the performance of the GME and PLS. The results showed that the GME outperforms PLS in terms of mean square errors (MSE). A simulated data also used to compute the CSI using the GME approach.

KEYWORDS

Generalised Maximum Entropy, Partial Least Squares, Costumer Satisfaction Models.

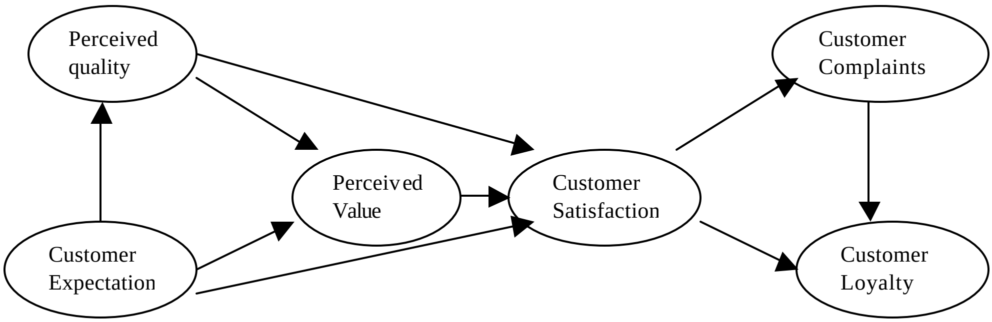
1. INTRODUCTION

Much has been written in the past few years on Customer Satisfaction measurement models in order to study the relationship between satisfaction and market share, and the impact of customer switching barriers (Fornell 1992) in terms of customer satisfaction Index (CSI). A Customer Satisfaction Index quantifies the level of profitable satisfaction of a particular customer base and specifies the impact of that satisfaction on the chosen measure(s) of economic performance. Index can be generated for specific businesses or market segments or "rolled-up" into corporate or divisional measures of performance. The index is used to monitor performance improvement and to identify differences between markets or businesses. The CSI score provides a baseline for determining whether the marketplace is becoming more or less satisfied with the quality of products or services provided by individual industry or company. Traditional approaches in estimating CSI from especial linear structural relationship models have raised two important issues; the first concerns with the Maximum Likelihood (ML) approach developed by Jöreskog (1973), which estimates the parameters of the model by the maximum likelihood method using Davidon-Fletcher-Powell algorithm. The other research issue concerns with the distribution free approach, namely, Partial Least Square (PLS). The PLS method was developed by Wold (1973, 1975) which he calls "soft modelling", or "Nonlinear Iterative Partial Least Square" (NIPLAS). After several versions in its development, Wold (1980) presented the basic design for the implementation of PLS algorithm. In the literature, the PLS method is usually presented by two equivalent algorithms. There are many authors who described PLS algorithms in their articles (Geladi and Kowalski (1986), Helland (1988), Helland(1990), Lohmoller (!989), Bremeton(1990) and Garthwaite(1994)). Appendix A is describe the PLS algorithm.

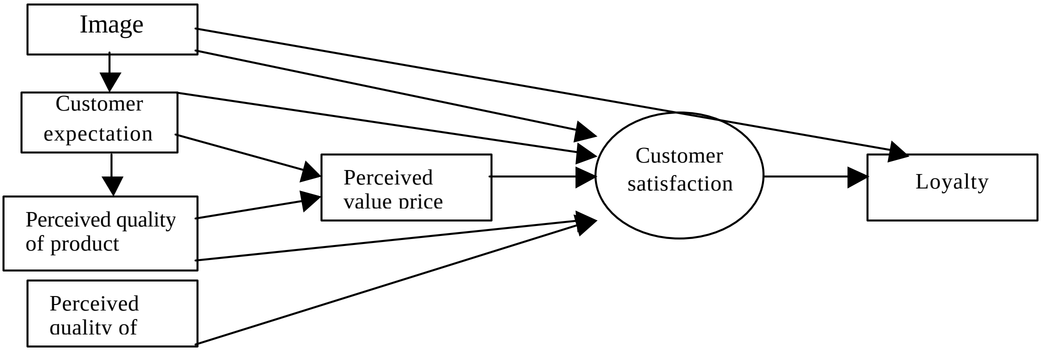
However, The Swedish CSI (Fornell 1992) and European's CSI (Gronhlodt et al 2000) models are used PLS method. This paper will discuss the GME estimation approach in solving the customer satisfaction models. A proposed method can be used to compute CSI based on statistical information about customer satisfaction measurements model.

2. COSTUMER SATISFACTION MEASUREMENT MODELS

Customer satisfaction model is a complete path model, which can be depicted in a path diagram to analyse a set of relationships between variables. It differs from simple path analysis in that all variables are latent variables measured by multiple indicators, which have associated error terms in addition to the residual error factor associated with the latent variable, a good examples on these models are the American customer satisfaction index (see Figure.1) which is a cross-industry measure of the satisfaction of customers in United States households with the quality of goods and services they purchase and use (Bryant 1995), and the European customer satisfaction index model, which is an economic indicator, represents in Figuer.2.



Figuer.1: The American Customer Satisfaction Framework



Figuer.2 The European Customer Satisfaction Framework

Many researchers from various disciplines have used Linear Structural Relationship (LISREL) as a tool for analysing customer satisfaction models. The general and formal model of customer satisfaction can be written as a series of equations represented by three matrix equations Jöreskog (1973):

$$\eta_{(m \times 1)} = B_{(m \times m)} * \eta_{(m \times 1)} + \Gamma_{(m \times n)} * \xi_{(n \times 1)} + \zeta_{(m \times 1)} \quad (1)$$

$$y_{(p \times 1)} = \Lambda_{y\ (p \times m)} * \eta_{(m \times 1)} + \varepsilon_{(p \times 1)} \quad (2)$$

$$x_{(q \times 1)} = \Lambda_{x\ (q \times n)} * \xi_{(n \times 1)} + \delta_{(q \times 1)} \quad (3)$$

The structural equation models given in (1-3) have two parts; the first part is structural model (1) that represents a linear system for the inner relations between the unobserved inner variables. The second part is the measurement model (2) and (3) that represents the outer relation between observed and unobserved or latent and manifest variables.

The structural equation model (1) refers to relations among exogenous variables (i.e; a variables that is not caused by another variable in the model), and endogenous variables (i.e; a variables that is caused by one or more variable in the model). The inner variables in this equation, **h** which is a vector of latent endogenous variables, and **x** which is a vector of latent exogenous variables are related by a structural relation. The parameters, **B** is a matrix of coefficients of the effects of endogenous on endogenous variables, and **G** is a matrix of coefficients of the effects of exogenous variables (**x**'s) on equations. However, **z** is a vector of residuals or errors in equations.

The inner variables are unobserved. Instead, we observe a number of indicators called outer variables and described by two equations to represent the measurement model (2) *and (3) which specify the relation between unobserved and observed, or latent and manifest variables. The measures in these two equations,* ***y*** *is a* ***p x 1*** *vector of measures of dependent variables, and* ***x*** *is a* ***q*** *x 1 vector of measures of independent variables. The parameters,* $\boldsymbol{L_y}$ *is a matrix of coefficients, or loadings, of* ***y*** *on unobserved dependent variables (****h****), and* $\boldsymbol{L_x}$ *is a* ***q*** *x n matrix of coefficients, or loadings, of* ***x*** *on the unobserved independent variables (****x****). Moreover,* ***e*** *is a vector of errors of measurement of* ***y****, and* ***d*** *is a vector of errors of measurement of* ***x****.*

The model given in (1-3) has many assumptions that may be perceived as restrictions, and these may be treated as hypotheses to be confirmed or disconfirmed and the rational of their specification in the model depend on methodological, theoretical, logical or empirical considerations, these assumptions:

(i) The elements of **h** and **x**, and consequently those of **z** also, are uncorrelated with the components of **e** and **d**. The later are uncorrelated as well, but the covariance matrices of **e** and **d** need to be diagonal. The means of all variables are assumed to be zero, which mean that the variables are expressed in the deviation scores. That is,

$$E(\eta) = E(\xi) = E(\zeta) = E(\varepsilon) = E(\delta) = 0$$

$$E(\varepsilon\varepsilon`) = \theta^2{}_\varepsilon \text{, and } E(\delta\delta`) = \theta^2{}_\delta$$

where $\theta^2{}_\varepsilon$ and $\theta^2{}_\delta$ are diagonal matrices.

(ii) It is assumed that the inner variables (η, ξ) are not correlated with the error terms (ε, δ), but they may be correlated with each other. Moreover, ξ and ζ are uncorrelated. That is,

$$E(\eta\varepsilon`) = E(\xi\delta`) = E(\xi\zeta`) = 0$$

(iii) B is nonsingular with zeros in its diagonal elements.

Given information about the variables $x_{(q\ x\ 1)}$ and $y_{(p\ x\ 1)}$, the objective in this article is to recover the unknown parameters $B_{(m\ x\ m)}$, $\Gamma_{(m\ x\ n)}$, $\Lambda_{y\ (p\ x\ m)}$, $\Lambda_{x\ (q\ x\ n)}$ and the disturbances $\zeta_{(m\ x\ 1)}$, $\varepsilon_{(p\ x\ 1)}$, $\delta_{(q\ x\ 1)}$ by using the GME principle.

3. GENERALIZED MAXIMUM ENTROPY (GME) ESTIMATION APPROACH

Conventional work in information theory concerns with developing a consistent measure of the amount of uncertainty. Suppose we have a set of events $\{x_1, x_2, \ldots, x_k\}$ whose probabilities of occurrence are $p_1, p_2, \ldots, p_k$ such that $\sum_{i=1}^{k} p_i = 1$. These probabilities are unknown but that is all we know concerning which event will occur. Using an axiomatic method to define a unique function to measure the uncertainty of a collection of events, Shannon (1948) defines the entropy or the information of entropy of the distribution (discrete events) with the corresponding probabilities P = $\{p_1,p_2,\ldots,p_k\}$, as

$$H(P) = -\sum_{i=1}^{k} p_i \ln(p_i) \qquad (4)$$

where 0ln(0) = 0.

The amount $(-\ln(p_i))$ is called the amount of self information of the event x_i. The average of self-information is defined as the entropy. The best approximation for the distribution is to choose p_i that maximizes (4) with respect to the data Consistency constraints and the Normalization-additivity requirements. Golan et al (1996) developed GME procedure for solving the problem of recovering information when the underling model is incompletely known and the data are limited, partial or incomplete. Al-Nasser et al (2000) developed the GME method for estimating Errors-In-Variables models and Abdullah et al (2000) used the same approach to study the functional relationship Between Image, customer satisfaction and loyalty.

3.1. RE-PARAMETERISATION

In order to illustrate the use of GME in estimating the model given in (1-3) we rewrite this model as:

$$y = \Lambda_y \Lambda_x^{-1} \Gamma (I - B)^{-1} (x - \delta) + \Lambda_y (I - B)^{-1} \zeta + \varepsilon \qquad (5)$$

where I is the identity matrix, and Λ_x^{-1} is the generalised inverse of Λ_x.

The GME principle stated that one chooses the distribution for which the information (the data) is just sufficient to determine the probability assignment. Hence the GME is to recover the unknown probabilities, which represents the distribution function of the random variable. However, the unknown parameters in customer satisfaction model are not in the form of probabilities and their sum does not represent the unity, which is the main characteristic of the probability density function. Therefore, in order to recover the unknowns in the model we need to rewrite the unknowns in terms of probabilities values. In this context we need to reparametrized the unknowns as expected values of discrete random variable with two or more sets of points, that is to say;

$$\boldsymbol{b}_{jk} = \sum_{s=1}^{S} z_{jks} b_{jks} \quad , \sum_{s=1}^{S} b_{jks} = 1, \text{ j = 1,2,…,m , k = 1,2,…,m}$$

$$\boldsymbol{g}_{ij} = \sum_{l=1}^{L} g_{ijl} f_{ijl} \quad , \sum_{l=1}^{L} f_{ijl} = 1, \text{ j = 1,2,…,m , i = 1,2,…, n}$$

$$\boldsymbol{l}^{x}_{qi} = \sum_{a=1}^{A} L^{x}_{qia}\, d^{x}_{qia} \,, \sum_{a=1}^{A} d^{x}_{qia} = 1,\ q = 1,2,\ldots,\ \mathbf{q}\,,\ i = 1,2,\ \ldots,\ n$$

$$\boldsymbol{l}^{y}_{pj} = \sum_{c=1}^{C} L^{y}_{pjc}\, d^{y}_{pjc} \,, \sum_{c=1}^{C} d^{y}_{pjc} = 1,\ p = 1,2,\ldots,\mathbf{p}\,,\ j = 1,2,\ldots,m$$

$$\boldsymbol{z}_{j} = \sum_{t=1}^{T} v_{jt}\, w_{jt} \quad, \sum_{t=1}^{T} w_{jt} = 1,\ j = 1,2,\ldots,m$$

$$\boldsymbol{d}_{q} = \sum_{r=1}^{R} v^{x}_{qr}\, w^{x}_{qr} \,, \sum_{r=1}^{R} w^{x}_{qr} = 1,\ q = 1,2,\ldots,\ \mathbf{q}$$

$$\boldsymbol{e}_{p} = \sum_{e=1}^{E} v^{y}_{pe}\, w^{y}_{pe} \,, \sum_{e=1}^{E} w^{y}_{pe} = 1,\ p = 1,2,\ldots,\mathbf{p}$$

Using these re-parameterisation expressions the model (5) can be rewritten as

$$y_p = \psi(b,f,d^x,d^y,w^x,w^y,w)$$

where

$$\psi(b,f,d^x,d^y,w^x,w^y,w) =$$

$$\left\{\left(\sum_j\sum_c L^{y}_{pjc}\, d^{y}_{pjc}\right)\left(1-\sum_i\sum_k\sum_s z_{iks}\, b_{iks}\right)^{-1}\right\} *$$

$$\left\{\left(\sum_q\sum_i\sum_a L^{x}_{qia}\, d^{x}_{qia}\right)\left(\sum_i\sum_j\sum_l g_{ijl}\ f_{ijl}\right)\left(\sum_q\left[x_q-\sum_r v^{x}_{qr} w^{x}_{qr}\right]\right)+\left(\sum_j\sum_t v_{jt}\, w_{jt}\right)\right\} \quad (6)$$

$$+\sum_e v^{y}_{pe}\, w^{y}_{pe}$$

The weight support of the disturbance parts (v^x,v^y,v) will be chosen such that they are symmetric around zero for all j, q and p. However, the choice of the support of the other parameters are chosen to span the possible parameter space for each parameter (Golan et al (1997) Golan et al(1996) and Al-Nasser and Abdullah (2000)).

3.2 REFORMULATION AND SOLUTION

Given the re-parameterisation, the GME system can be expressed as a non-linear programming problem subject to linear constraints. Its objective function can be stated in scalar summation notations, maximising this function subject to the consistency and the add-up normalisation constraints can solve the problem. The model reformulation using the GME is given by:

Maximize $\quad H(b,f,d^x,d^y,w^x,w^y,w) =$

$$-\sum_j\sum_k\sum_s b_{jks}\ln(b_{jks})-\sum_i\sum_j\sum_l f_{ijl}\ln(f_{ijl})-\sum_q\sum_i\sum_a d^x_{qia}\ln(d^x_{qia})$$

$$-\sum_p\sum_j\sum_c d^y_{pjc}\ln(d^y_{pjc})-\sum_j\sum_t w_{jt}\ln(w_{jt})-\sum_q\sum_r w^x_{qr}\ln(w^x_{qr})$$

$$-\sum_p\sum_e w^y_{pe}\ln(w^y_{pe})$$

Subject to

(i) $y_p = \psi(b,f,d^x,d^y,w^x,w^y,w)$

(ii) $\sum_{s=1}^{S} b_{jks} = 1$, j = 1,2,…,m , k = 1,2,…,m

(iii) $\sum_{l=1}^{L} f_{ijl} = 1$, j = 1,2,…,m , i = 1,2,…, n

(iv) $\sum_{a=1}^{A} d^x_{qia} = 1$, q = 1,2,…, **q** , i = 1,2, …, n

(v) $\sum_{c=1}^{C} d^y_{pjc} = 1$, p = 1,2,…,**p** , j = 1,2,…,m

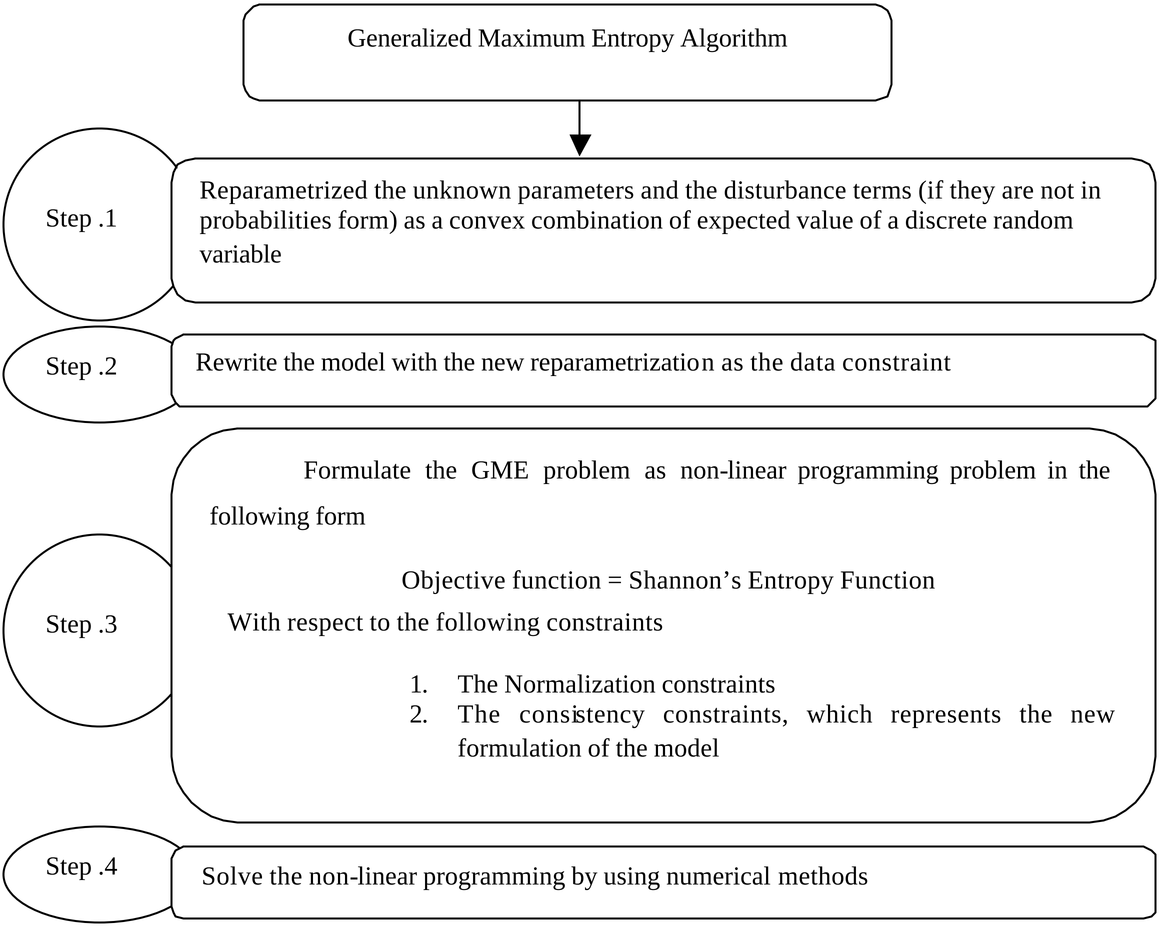
(vi) $\sum_{t=1}^{T} w_{jt} = 1$, j = 1,2,…,m

(vii) $\sum_{r=1}^{R} w^x_{qr} = 1$, q = 1,2,…, **q**

(viii) $\sum_{e=1}^{E} w^y_{pe} = 1$, p = 1,2,…,**p**

where $\psi(b,f,d^x,d^y,w^x,w^y,w)$ as given in (6).

In this system we have (**p** + m^2 + nm + **q**n + **p**m + m + **q** + **p**) equations including (Sm^2 + nmL + **q**nA + **p**mC + mT + **q**R + **p**E) unknowns. However, to solve this non-linear programming system a numerical method should be used. The following diagram describes the GME algorithm in four steps,



4. A SIMULATION STUDY

To illustrate the GME estimation method, we conducted a simulation study using simplified model that is used for the Swedish customer satisfaction index, proposed by Claes .C et al (1999), that consists of three exogenous variables ξ_1, ξ_2, and ξ_3, and one endogenous variables η. The inner structure is defined as

$$\eta = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \gamma_3 \xi_3 + \zeta$$

where γ_1, γ_2 and γ_3 are regression coefficients, and ζ is disturbance term. The manifest variables are denoted as x for the ξ variables, and y for the η variable. The measurement models for ξ variables are formative (Bagozzi and Fornell (1982)) and given by:

$$\xi_1 = \pi_1 x_1 + \pi_2 x_2 + \pi_3 x_3 + \delta_1$$
$$\xi_2 = \pi_4 x_4 + \pi_5 x_5 + \pi_6 x_6 + \delta_2$$
$$\xi_3 = \pi_7 x_7 + \pi_8 x_8 + \pi_9 x_9 + \delta_3$$

where π are regression coefficients, and the δ are disturbances. The measurement model for the η variable is reflective and given by:

$$y_1 = \lambda_1 \eta + \varepsilon_1$$
$$y_2 = \lambda_2 \eta + \varepsilon_2$$
$$y_3 = \lambda_3 \eta + \varepsilon_3$$
$$y_4 = \lambda_4 \eta + \varepsilon_4$$

where λ are coefficients and ε are disturbance part. Given this structural model, the simulation study was done under the following conditions:

1- Generate 100 random samples each of size 15,20,25,30,40 from the given model.
2- For the formative model the x values were generated from symmetric Beta distribution with parameters (6,6).
3- All π coefficients are set to be 1/3.
4- The γ coefficients are initialled by (0.8, 0.1, 0.1).
5- The λ coefficients are initialled by (1.1, 1.0, 0.9, 0.8).
6- The error terms δ and ε are generated from the Uniform distribution U(0,1), while ζ generated from the standard Normal distribution.
7- Using the Fortran power station environment programs linked to IMSL library, all Normal varieties were generated from the subroutine ANORIN, the Beta varieties from RNBET and the GME system were solved by using successive quadratic programming method to solve a non-linear programming problem depending on NCONF based on the subroutine NLPQL.

Under these conditions the results for the MSE are given as shown in Table (1) for the GME approach and in Table (2) for the PLS method.

TABLE-1 MSE of The Estimated Coefficients By Using The GME

N	MSE($\bar{\hat{p}}$)	MSE($\hat{g}_1$)	MSE($\hat{g}_2$)	MSE($\hat{g}_3$)	MSE($\bar{\hat{l}}$)
15	7.406E-3	4.266E-2	6.679E-4	6.675E-4	7.406E-2
20	4.788E-3	2.081E-2	5.493E-4	4.970E-4	4.788E-2
25	4.046E-3	2.030E-2	5.111E-4	3.449E-4	4.606E-2
30	3.974E-3	1.965E-2	4.042E-4	2.577E-4	3.009E-2
40	3.915E-3	8.032E-3	3.827E-4	1.348E-4	1.470E-2

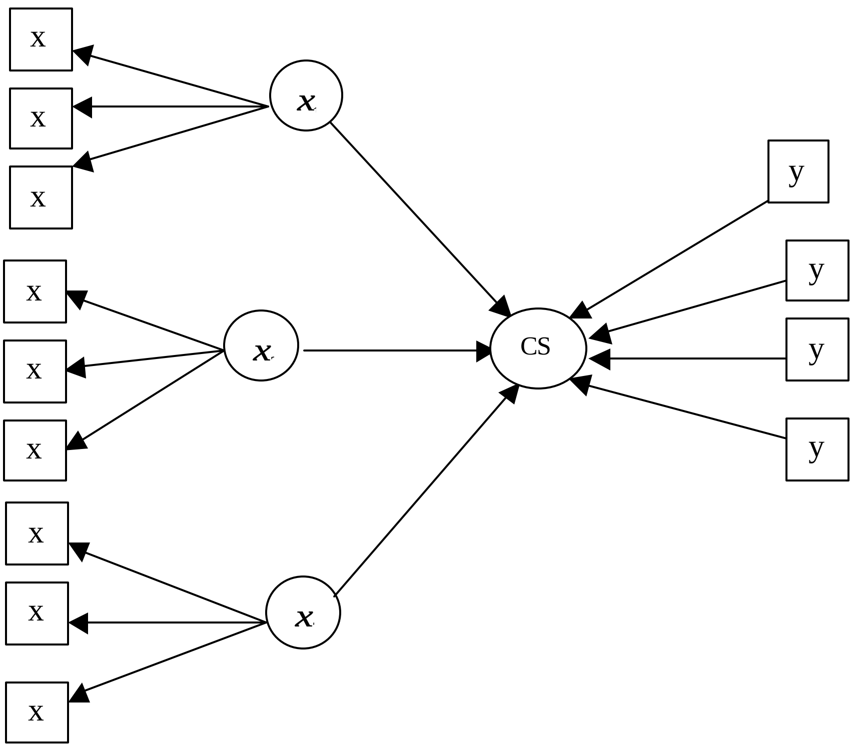
TABLE 2 MSE Of The Estimated Coefficients By Using The PLS

N	MSE($\bar{\hat{p}}$)	MSE($\hat{g}_1$)	MSE($\hat{g}_2$)	MSE($\hat{g}_3$)	MSE($\bar{\hat{l}}$)
15	2.716E-1	6.456E-1	1.474E-1	1.570E-1	2.6287
20	2.037E-1	4.842E-1	1.105E-1	1.178E-1	1.9715
25	1.629E-1	3.874E-1	8.845E-2	9.425E-2	1.5772
30	1.086E-1	3.228E-1	7.370E-2	7.854E-2	1.3143
40	0.148E-1	2.421E-1	5.528E-2	5.890E-2	9.857E-1

Where $\bar{\hat{p}}$ (the estimate mean of the coefficients in the measurement models for ξ variables) and $\bar{\hat{l}}$ (the estimate mean of the coefficients the measurement model for the η variable). From the results it could be note that the GME outperform the PLS method, and it gives better estimate with a very small sample size.

4.1 APPLICATION TO SIMULATED DATA

In order to illustrate the GME algorithm in solving customer satisfaction models to compute CSI, the model described in this article for the Swedish customer satisfaction indexused under conditions (1-7) given in the last section to generate a hypothetical data of size 12. The GME estimated values are given in the following diagram:



CS represents the latent variable for customer satisfaction, then the CSI computed as follows (Bryant E. B (1995)):

$$CSI = \frac{E(CS) - Min(CS)}{Max(CS) - Min(CS)} \times 100$$

Where E(.),Min(.) and Max(.) denote the expected, the minimum and the maximum value of the variable, respectively. Those of corresponding manifest variables determine the minimum and the maximum values of CS latent variable:

$$Min(CS) = \sum_{i=1}^{4} w_i Min(y_i) \quad , Max(CS) = \sum_{i=1}^{4} w_i Max(y_i)$$

where, w_i are the weights, for this example a uniform weights were used . Therefore, the CSI using GME model is 82.03. The CSI results indicate that the service quality regarding to the simulated data is Excellent.

5. CONCLUDING REMARKS

In this article we proposed the generalised maximum entropy (GME) estimation approach to the customer satisfaction models, which provide a better approach as it is meant for situations with limited or incomplete data and it is more robust against departures from classical assumptions on statistical distributions. The performance of the GME approach investigated and compared with an existing technique from the literatures, partial least squares (PLS). It can be observed from the simulation results that PLS are unreliable when the sample size relatively small, and the GME approach outperform the PLS in terms of MSE. Therefore, the GME can be considered as an alternative to the conventional method PLS to measure customer satisfaction index.

ACKNOWLEDGMENTS

This research has been supported by a grant from Yarmouk University

REFERENCES

1. Abdullah, M. B., Al-Nasser, A. D. & Nooreha, H. 2000. Evaluating Functional Relationship Between Image, Customer Satisfaction and Loyalty using General maximum Entropy. *Total Quality Management* 11(6): 826-829.
2. Bagozzi, P. R. & Fornell, C. 1982. Theoritical Concept, Measurement and Meaning. In C. Fornell, (ed). *The Second Generation of Multivariate Analysis, Vol. II; Measurement and Evaluation*. Toronto: John Wiley. 42-60.
3. Bremerton G. R. 1990. *Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems*. England: Ellis Horwood limited.
4. Bryant, E. Barbara. 1995. American Customer Satisfaction Index: Methodology Report. National Quality Research Center. University of Michigan Business School. An Arbor, MI 48109-1234.

5. Claes, C., Hackl, P. & Westlund, H., A. 1999. Robustness of partial Least Squares Method for Estimating Latent Variable Quality Structures. J. Applied Statistics. 26(4):435-446.
6. Fornell C. 1992. A National Customer Satisfaction Barometer: The Swedish Experience. J. Marketing, 56:6-21.
7. Garthwaite, H. P. 1994. An Interpretation of Partial Least Squares. *JASA*. 89(425): 122-127.
8. Geladi, P. & Kowalski, R. B. 1986. Partial least-Squares regression: A Tutorial. *Analytica Chimica Acta* 185: 1-17.
9. Golan, A., Judge, G. & Perloff, J. 1997. Estimation and Inference with Censored and Ordered Multinomial Response Data. *J. Econometrics* 79: 23-51.
10. Golan, A., Judge, G. & Karp, L.1996. A maximum entropy approach to estimation and inference in dynamic models or counting fish in the sea using maximum entropy. *J. of Economic Dynamics and Control* 20: 559-582.
11. Gronholdt, L., Martensen, A. & Kristensen, K. 2000. The Relationship Between Customer Satisfaction and loyalty: cross-industry differences. Total Quality Management. 11(6):509-514.
12. Helland, S. I. 1990. Partial Least Squares Regression and Statistical model. *Scand. J. Statist*. 17: 97-114.
13. Helland, S. I. 1988. On The Structure of Partial Least Squares Regression. *Comm. Statist. Simula*. 17(2): 581-607.
14. Joreskog, K. G. 1973. A general Method of Estimating a Linear Structural Equation system. In Goldberg, S. A. & Duncan, D. O (Eds). *Structural Equation Models in the Social Sciences*. New York: Seminar Press. 85-112.
15. Lohmoeller, J. B. 1989. *Latent Variable Path Modeling with Partial Least Squares*. New York: Springer-verlag.
16. Al-Nasser, A. D. Abdullah, M. B & Wan Endut, W. J. 2000. On Robust Estimation of Error In Variables models by using Generalized Maximum Entropy. Proceedings of international conference on mathematics and its applications in the new millennium. Dept. Mathematics. Uni Putra Malaysia. 18-19 July. 279-287.
17. Shannon C. E (1948). A mathematical Theory of Communications. Bell System Technical Journal, 27: 379-423.
18. Wold, H. 1973. Nonlinear Iteritive Partial Least Squares (NIPALS) Modeling: Some Current Development. In Krishnaiah, R. P (Ed). Multivariate Analysis III. New York: Academic. 383-407.
19. Wold, H. 1975. Soft Modeling by Latent Variables. The Nonlinear Itertive Partial Least Square (NIPALS). In Gani, J (ed). Perspectives in Probability and Statistitics: Papers in Honour of M. S. Barttlet. On the occation of his sixty-fifth birthday. Applied Probability Trust. London: Academic. 117-142.
20. Wold, H. 1980. Model Construction and Evaluation When Theoretical Knowledge is Scare: on the Theory and Application of Partial Least Squares. In Kmenta, J & Ramsey, B. J (Eds). *Model Evaluation in Econometrics*. New York: Academic. 47-74.

APPENDIX. A
PARTIAL LEAST SQUARE ALGORITHM

Suppose we have the following structural equations with the following relation

$$X = t_1' p_1 + t_2' p_2 + \ldots + t_A' p_A + E_A$$

$$y = t_1' q_1 + t_2' q_2 + \ldots + t_A' q_A + f_A$$

where X is a matrix of size (N x K), **y** is a vector of size N , $\mathbf{t_a}$ are N vectors of latent variables, $\mathbf{p_a}$ are k vectors of loading variables, q_a are scalars with same scores, E_a is the residual matrix and f_a the residual vector. The PLS algorithm has the following steps:

(i) Define the starting values for the X residuals (e_0) and **y** residual (f_0) as follows;

$$e_0 = x - \mu_x$$
$$f_0 = y - \mu_y$$

where

$$m_x = \left(\frac{\sum_{i=1}^{N} x_{ik}}{N} \right), k = 1,2,\ldots,K$$

$$m_y = \frac{\sum_{i}^{N} y_i}{N}$$

and **x** is k vectors of size N. For a = 1,2,… do steps (ii)-(vi) below:

(ii) Calculate the loading weight, $w_a = \mathrm{Cov}(e_{a-1}, f_{a-1})$

(iii) Estimate the score for the next PLS component by

$$t_a = e'_{a-1} w_a$$

(iv) determine x loading and y loading by Least Squares with

$$p_a = \mathrm{Cov}(e_{a-1}, t_a) / \mathrm{Var}(t_a)$$
$$q_a = \mathrm{Cov}(e_{a-1}, t_a) / \mathrm{Var}(t_a)$$

(v) Find the new residuals

$$e_a = e_{a-1} - p_a t_a$$
$$f_a = f_{a-1} - q_a t_a$$

(vi) Compare the t values with the one from the preceding iteration. If they are equal (with certain error, say 0.00001) then exit with the results or else go to (ii).

Deciding the number of components to include in regression model is a tricky problem (Garthwaite, 1994). However, Helland (1988) noted that the number of factors to retain in final equation is usually determined by a cross-validation procedure: The data set is divided into G parts, with calibration is done with one part and validation on the other part of the data. The number of factors is chosen so that the estimated error of prediction is minimised Wold (1978) discussed this method in context of PLS.