Paper 272

Assessing Predictive Power and Identifying Outliers in Covariance Structure Analysis Models

Joseph Retzer, Market Probe, Inc., Milwaukee WI
Kurt Pflughoeft, UW-Milwaukee, Milwaukee WI

Abstract

Predictive power in LISREL-like covariance structure analyses is arguably the most preferred measure to use when comparing competing models. This paper begins by briefly reviewing the implication of specified causal direction in SEM (Structural Equation Models) and the application of the PLS approach to facilitate prediction. It then proceeds to describe an alternative method for predicting manifest endogenous indicator levels while maintaining hypothesized causal directions. A simple application of conditional multivariate expectation is employed to construct this more robust measure. This predictive measure may be used to construct a statistic similar to PRESS (Predicted Residual Sum of Squares) for model evaluation and/or to identify outliers in the data. In addition, an example using real world survey data is presented and discussed.

The uniquely powerful capabilities of the SAS macro system are used in conjunction with estimated covariance matrices produced by PROC CALIS to straightforwardly construct the measure.

1 Theoretical Background

Predictive inference in confirmatory covariance structure analysis is often viewed as problematic and potentially a function of model complexity. In latent variable covariance structure models two approaches stand out as being most generally accepted. The first is the standard ML (maximum likelihood) technique most commonly associated with LISREL-type analysis and the second is PLS (Partial Least Squares). [1] Advantages of PLS include it's lack of distributional assumptions for measured variables and its avoidance of inadmissable solutions. PLS is often applied in situations where model prediction is of interest since it provides an exact definition of component scores (latent variables are represented as linear combinations of manifest variable measures) hence avoiding the problem of factor indeterminacy. Causal direction in PLS modeling therefore runs from indicator to construct unlike the case usually associated with LISREL-type models. Since theoretical concerns should be the primary determinant of model construction and subsequent technique selection, an alternative approach may be needed when prediction performance is desired. This alternative approach should not be viewed as a replacement for PLS, instead covariance fitting procedures (such as ML and GLS) and the variance based PLS approach should be viewed as complementary in nature rather than as competitors (Joreskög and Wold, 1982).

Once a predictive measure is derived, it may be employed in measuring predictive performance by jackknifing through the sample in order to predict individual cases of manifest endogenous variables. The squared difference between predicted and actual levels is then averaged creating a comparative index which will be referred to as the CVMSE (Cross Validation Mean Square Error). This measure is analogous to the PRESS (Predictive Residual Error Sum of Squares) statistic commonly used in regression analysis. These predictive errors may also be ordered allowing outlier identification.

Another predictive performance measure is the Q^2 statistic, commonly employed in PLS modeling.

[1] Originally suggested by Wold as the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm (Wold, 1980).

Computation of the Q^2 measure is illustrated in equation (1) below. [2]

$$Q^2 = 1.0 - \frac{\sum_n (Y_i - \hat{Y}_{(i)})^2}{\sum_n (Y_i - Y_{.(i)})^2}. \qquad (1)$$

Where:

Y_i = ith case of the dependent variable.
$Y_{.(i)}$ = mean of the dependent variable computed without the ith case.
$\hat{Y}_{(i)}$ = estimate of the dependent variable when the ith case is omitted.

In the case of PLS estimation, where OLS is employed to create estimates,

$$\hat{Y}_i = \sum_k X_{ki} b_{k(i)}.$$

Where:

$b_{k(i)}$ = set of regression coefficients computed when the ith case is omitted.
X_{ki} = ith case of the independent variable vector.

For the purposes of this paper our estimate of Y_i will be the condition MVN expected value. The Q^2 measure can be viewed as the jackknife analogue to the regression R^2 with certain important differences. These differences include,

- Q^2 may increase as variables are removed from the model implying a reduction in noise.
- Q^2 may be zero or negative implying that the dependent variable mean performs better in prediction than does the specified model.

2 Hypothesized Causal Direction

When specifying causal direction, constructs may be viewed as either "reflective" or "formative". Reflective constructs are those typically associated with LISREL-type models while "formative" constructs are properly represented with a PLS approach. In addition a model may contain constructs of both types. All three models are illustrated in figure 1 below, (Fornell and Bookstein, 1982).

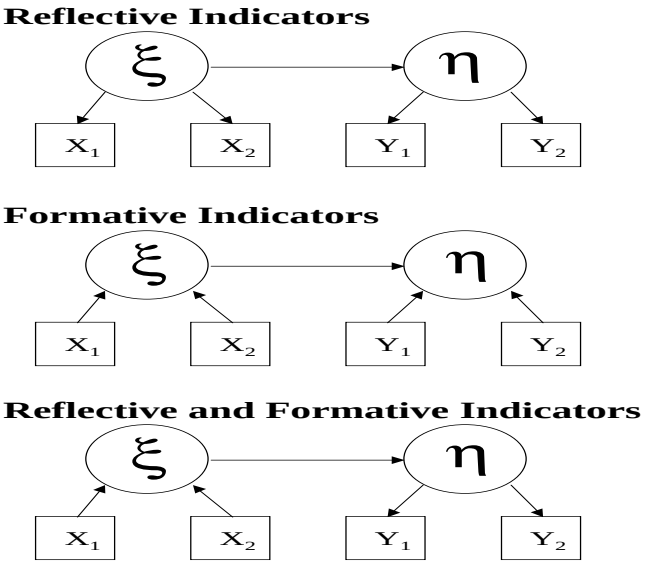

Figure 1.

The PLS approach is often times chosen for predicting from covariance structure models. The research must bear in mind however that causal direction is assumed to be formative in nature when using this technique. If theory suggests our indicators are reflective, an alternative approach to prediction should be applied. Estimation of the multivariate normal conditional expected mean using the maximum likelihood derived covariance matrix of the models manifest variables relies solely on basic structural model assumptions. Specifically, the technique requires only that the manifest variables be jointly normal (a standard assumption in most covariance structure models). If we partition the manifest variable matrix into two parts,

- Y, variables to predict and
- X, variables which explain the Y's,

[2] See (Ball, 1963)

and we assume that

$$\begin{bmatrix} Y \\ X \end{bmatrix}$$

is distributed as $N(\mu, \Sigma)$ where

$$\mu = \begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{YY} \end{bmatrix}$$

and that $|\Sigma_{XX}| > 0$, then the conditional distribution of Y given $X = x$, is normal with mean

$$\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(x - \mu_X)$$

and covariance

$$\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

(Johnson and Wichern, 1992). The conditional mean serves as a basis for prediction of a single observation and may be extended to create model prediction performance indices and used to identify potential model outliers.

3 An Illustration

The data used to illustrate conditional mean prediction is taken from a global customer satisfaction survey which measured, among other things, stated likelihood to repurchase. A theoretical structure is initially specified and subsequently adjusted, through exploratory analysis, to arrive at a final model (a nested model of that originally specified). This model is illustrated graphically in Figure 2. [3]

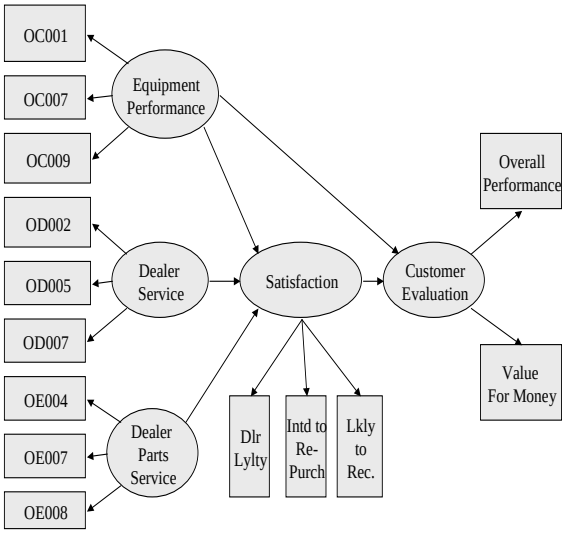

Figure 2.

Manifest indicators include,

OC001: "Reliability of Equipment."

OC007: "Ease of Operating the Equipment."

OC009: "Operator Comfort in Using the Equipment."

OD002: "Professionalism of the Service Personnel."

OD005: "Time Taken to Complete Repairs."

OD007: "Delivery of Equipment as Promised or Loaner Provided."

OE004: "Off the Shelf Parts Availability."

OE007: "Reliability of Parts Service."

OE008: "Layout and Appearance of Dealer's Facilities."

Dlr Lylty: "Dealer Loyalty."

Intd to Re-Purch: "Intend to Repurchase."

Lkly to Rec: "Likely to Recommend."

Overall Perf.: "Overall Performance Satisfaction."

Value for Money: "Perceived Value for the Money."

The data is segmented into two groups, European and American. For each group the

[3] SAS PROC CALIS code for this model along with the CVMSE iterative algorithm are given in Appendix I.

CVMSE is calculated and used to assess relative model predictive power across geographic segments. In addition, errors for each group are ordered and apparent outliers are identified.

4 Results

The analysis was run on two equal size samples (n=300) taken from US and European data respectively. Predictive perfomance measures for both groups are presented in Table I below.

Results Using All 300 Observations		
	United States	Europe
$Q^2_{overall}$	.60	.42
Q^2_{value}	.60	.41
CVMSE	1.79	4.22

Table I.

Table I suggests the hypothesized model results in superior prediction of US data for both dependent manifest variables. Not surprisingly therefor, the model employing that data has a relatively lower CVMSE score as well.

Outlier identification was next employed to remove the top 25 cases with the largest predictive errors. Predictive performance measures corresponding to the new data sets (n=275) are given in Table II below.

Results With Outliers Removed		
	United States	Europe
$Q^2_{overall}$	.65	.53
Q^2_{value}	.74	.43
CVMSE	1.04	2.35

Table II.

While both predictive performance measures ($Q^2_{overall}$ and Q^2_{value}) increase for both data sets, predictive accuracy involving "Overall Performance" improves markedly more in Europe while in the US data, predictive accuracy improvement is greater for "Value for the Money". Still, overall predictive performance is better using US data as noted by CVMSE.

5 Conclusion

For both pairs of analysis' (n=300 and n=275) the model appears to predict more accurately when using US data. This may imply that the theoretical model is a better reflection of US customer attributes rather than those of their European counterparts. As should be expected, general predictive performance improves as outliers are dropped for both the US and Europe. Interestingly however, the improvement of dependent mainifest variable prediction differs markedly across continents. This suggests that the presence of outliers has greater adverse effect on prediction of "Value for the Money" in the US and, conversly, on prediction of "Overall Performance" in Europe.

Further research in this area could focus on SEM group difference testing in order to suggest more a more appropriate model specification for the European data. In conjunction with this analysis, additional confirmatory theoretical model specification should be employed to improve the model for this data.

References

Ball, R. (1963). The significance of simultaneous methods of parameter estimation in econometric models. *Applied Statistics*, 12:14–25.

Fornell, C. and Bookstein, L. (1982). Two structural equation models: Lisrel and pls applied to consumer exit-voice theory. *Journal of Marketing Research*, 19:440–452.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*. Prentice Hall Inc.

Joreskög, K. and Wold, H., editors (1982). *Systems Under Indirect Observation: Causality, Structure, Prediction*, volume 1, chapter The ML and PLS Techniques for Modeling with Latent Variables: Historical and Comparative Aspects, pages 263–270. Amsterdam: North-Holland.

Wold, H. (1980). *Evaluation of Econometric Models*, chapter Model Construction and Evaluation When Theoretical Knowledge is Scarce–Theory and Application of Partial Least Squares. New York: Academic Press.

Contact Author

Joseph Retzer, Ph.D.
Market Probe, Inc.
Milwaukee WI, 53226
414-778-6000 (voice)
jjr@execpc.com

Appendix I

CVMSE Algorithm and SAS PROC CALIS Code

The following SAS PROC CALIS code illustrates the creation of an output data set containing the maximum likelihood estimated covariance matrix used in calculating the CVMSE for the model illustrated in this paper. [4]

```
proc calis data=model_dt cov
             outstat=_pcovs_ noprint;
   lineqs
     oc001    = p_c1_feq f_eqprf + err1,
     oc007    =          f_eqprf + err2,
     oc009    = p_c9_feq f_eqprf + err3,
     od002    = p_d2_fds f_dlrsrv + err4,
     od005    =          f_dlrsrv + err5,
     od007    = p_d7_fds f_dlrsrv + err6,
     oe004    = p_e4_fdp f_dlrprt + err7,
     oe007    =          f_dlrprt + err8,
     oe008    = p_e8_fdp f_dlrprt + err9,
     f_eval = pfevfsat f_sat   +
              pfevfep  f_eqprf  + err10,
     f_sat  = pfsatfep f_eqprf +
              pfsatfds f_dlrsrv +
              pfsatfdp f_dlrprt + err11,
     dlrlyty  = pdl_fsat f_sat + err12,
     re_purch =          f_sat + err13,
     rec      = prc_fsat f_sat + err14,
     overall  = p_fev_ov f_eval + err15,
     value    =          f_eval + err16;
   std
     err1-err16 = vare1-vare16,
     f_eqprf    = var_feq,
     f_dlrsrv   = var_fds,
     f_dlrprt   = var_fdp;
   cov
     f_eqprf f_dlrsrv = cfeqfds,
     f_eqprf f_dlrprt = cfeqfprt,
     f_dlrsrv f_dlrprt = cfdsfprt;
   var
     oc001 oc007 oc009 od002 od005 od007
       oe004 oe007 oe008 dlrlyty re_purch
       rec overall value;
   run;

quit;
```

This covariance matrix, along with means and cross products, are used to estimate the conditional expectation of the dependent manifest variable "Likelihood to Repurchase". This code is inserted in a SAS macro which jackknifes through the sample. On each iteration the following steps occur,

1. One observation is held out. The remaining data is used to estimate the associated covariance matrix.

2. The estimated covariance matrix in (1) is used in conjunction with the hold out observation to predict the value of the hold out's dependent manifest variables ("Likelihood to Repurchase" and "Dealer Loyalty").

3. The differences (errors) between predicted and actual endogenous manifest variables are calculated and retained.

Once the process above iterates through the entire data set, the errors are squared, summed and an average is calculated. This measure (CVMSE) can then be used to compare model predictive performance.

[4] All necessary information will exist in the output data set, "_pcovs_".