Comparison of Rebus PLS and Fimix PLS for Estimating Parameter in Structural Model

Tria Sofa Purnama¹, Erfiani², Indahwati³
¹,²,³Department of Statistics, Bogor Agricultural University, INDONESIA

ABSTRACT
Partial Least Square Path Modeling (PLS-PM) is simple method in modeling involved latent variables. It can apply to small of sample size and unspecified distribution of units. In heterogenous units exists, a segmentation approach that is the development of PLS needs to bias minimizing. The popular approaches are Fimix PLS and Rebus PLS. Fimix PLS detects heterogeneity from structural model and more Rebus PLS detects it both structural and measurement model. Both are compared in overlapping units in the segment and variance of residual. The performance was measured by mean bias absolute (MBA) for parameter in structural model. Rebus PLS was given more stable parameter estimator in all conditions. However, Fimix PLS showed the best performance in all conditions.

Keywords—Heterogeneity, Segmentation, Latent Variable, PLS-PM

I. INTRODUCTION

Regression analysis is a statistical model used to depict causal relationships between two or more variables. Cause variables are called independent variables and effect variables are called dependent variables. The variables are known a number, observed variables. Regression analysis cannot handle unobserved (latent) variables. A number of latent variables are constructed from several explanatory variables that serve as indicators. One of the statistical analyses used to analyze the causal relationships between latent variables is Structural Equation Modeling (SEM).

SEM is a model and its forming approach to the factor analysis, structural modeling, and path analysis. There are measurement models (factor analysis and path analysis) in SEM to specify the relationship between latent and indicator variables, while the structural model specifies causal relationships among latent variables. Partial least square (PLS) is a method that estimates the parameters in SEM (atau “Parameters in SEM are estimated using Partial Least Square (PLS)”). SEM estimate parameter component based [2] PLS has relatively loose assumptions, both on the distribution of observations and the sample size of variables (the size should not be large). In addition, PLS allows researchers to use different mode of indicators such as reflective (latent variables explain the explanatory variables) or formative (explanatory variables explain the latent variables). Parameter estimation and model validation bias caused by diversity of observations can not be solved by this method. Instead, it can be solve by grouping.

Grouping is used to produce homogeneous units of observation. Traditional grouping such as a priori classes with PLS with multiple group analysis [3] and segmentation tree algorithm path model [7] are alternative solutions to minimize heterogeneity of observations. Both assume that the source of heterogeneity of observation is observed in certain variables (eg, education level, geographic location). Moreover, a priori classes can not minimize heterogeneity in the model, particularly the heterogeneity in the score variables and relationships between variables in the model [4] It can be solved instead by PLS clustering approach.

PLS clustering approach assumes that the source of observation heterogeneity is unknown. However, PLS requires relationship between variables so this approach can not be used [4]. Another commonly used approach is a finite mixture approach and distance-based approach. Finite Mixture Model Partial Least Square (FIMIX-PLS) was developed by Hahn [2]. The method detects heterogeneity from structural model. Other methods that can be used in a PLS segmentation is Response Based Units Segmentations PLS (Rebus-PLS). Rebus-PLS is one of the distance-based approach introduced by Trinchera [6] and Esposito et al [1]. This method is more flexible and can detect heterogeneity from both structural and
measured model including exogenous and endogenous latent variables in the model.

The aim of this study was to compare Fimix PLS and Rebus-PLS. Both methods simulated in overlapping segments of the units and variance of residual.

II. PLS PM AND SEGMENTATION APPROACHES

Partial Least Square- Structural Equation Modeling

Partial Least Square Path Modeling (PLS PM) is a variance maximization method. The variance is in dependent variable explained independent variable and variance maximization gave a covariance matrix. The variables form linear combination: structural model (inner model) and measurement model (outer model). The structural model showed:

$$\eta = B\xi + \Gamma \zeta + \zeta$$  \hspace{1cm} (1)

where \(\eta\) is endogenous latent variable and \(\xi\) is exogenous latent variable. Path coefficient, \(B\) interrelate among endogenous latent variables and path coefficient, \(\Gamma\) do endogenous latent variable to exogenous latent variable.

The error in structural model (\(\zeta\)) assumes that having a zero mean \(E(\zeta) = 0\) and uncorrelated with exogenous \(E(\xi) = 0\).

The measurement model has correlation between latent variable and their corresponding manifest variable (\(x\) and \(y\)). The model has causal correlation of latent variable to its manifest variables, reflective measurement model. The reflective measurement model are also expected each manifest correlated. The manifest variable in reflective mode showed:

$$x_i = \lambda \xi_i + e_i$$

$$y_i = \lambda \eta_i + \delta_i$$  \hspace{1cm} (2)

where \(\lambda\) is simple regression coefficient of manifest variable on its latent variable and \(e\) and \(\delta\) are residual term, each assume having a zero mean and uncorrelated with its manifest variables.

Finite Mixture Model PLS (Fimix-PLS)

Fimix-PLS is composed of PLS and finite mixture approach [3]. The approach give maximum likelihood method per segments. The segmentation means to minimize unobservable heterogeneity considering correlation each latent variables in structural model. The \(\ln L\) function maximize estimator:

$$\ln L = \sum_{i=1}^{I} \sum_{k=1}^{K} \pi_{ik} \ln (f(\eta_i | \xi_i, \xi_j - \beta_k, -\Gamma_k, \Psi_k)) + \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \ln (\rho_k)$$  \hspace{1cm} (3)

where

\(\rho\) is mixing proportion of each \(k\) segment in finite mixture model and \(\xi_i, \beta_k, \Gamma_k, \Psi_k\) show vector in any segments which unknown parameter of each \(k\) segments. Expectation Maximization (EM) algorithm is used to estimate parameter.

Response Based Units Segmentation PLS (Rebus-PLS)

Rebus-PLS was introduced by Trinchera [8] and Esposito et al [1] to detect heterogeneity in structural and measurement model. The approach is segmenting the uncorrelated variable in structural and measurement. Both take into account the number of segment formed by clustering technique. It gives each unit is in certain segment. Thus, all units in the segment modeled PLS-PM. The final models consist of convergent units. The units updated the allocation among segments which take into account statistics, Closeness Measurement (CM).

The update is allocating units based on the closest CM. It reflects distance between certain unit to its model. The measurement of distance showed [1].

$$CM_{ji} = \sum_{p=1}^{P} \sum_{k=1}^{K} \frac{\pi_{ik} (x_{jp} - \hat{x}_{jk}) (x_{jp} - \hat{x}_{jk})}{\pi_{ik} (x_{jp} - \hat{x}_{jk}) (x_{jp} - \hat{x}_{jk})}$$

where \(\hat{x}_{jk}\) is communality index for \(p\)th variable, \(q\)th block and \(k\)th latent variable segment, \(e_{ijk}^2\) is residual of the measurement model for \(i\)th units, \(f_{ijk}\) is residual of structural for \(i\)th endogenous variable, \(I\) is the number of units, and \(K\) is the number of dimension.

III. DATA AND METHODOLOGY

In this paper, the data were proposed as simulation model and the generation study based on Structural Equation Modeling approach. The model was formed by 100 units and numeric manifest variables. The steps showed [6]:

1. Determine path coefficient in measurement model \(\lambda_{ij}^{(x)} = 0.7\) where \(i = 1,2,3\) and \(j = 1,2\). The units (n=100) divided into two equal segments \(n_1 = n_2 = n/2\) and the path coefficient in structural model \(\gamma_1, \gamma_2\) based on simulation data characteristics (Table 1) and also data simulation combination (Table 2).

<table>
<thead>
<tr>
<th>(I)</th>
<th>Segment 1</th>
<th>Segment 2</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\gamma_1)</td>
<td>0.9</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>(\gamma_2)</td>
<td>0.1</td>
<td>0.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>
2. Generate a sample of size 100 from a standard normal distribution for exogenous latent variables \((\xi_1, \xi_2)\) which was uncorrelated. Generate number of size 100 from normal distribution (zero mean and uncorrelated) for residual \(\left(\delta_1, \delta_2, \delta_3\right)\) in structural model and residual \(\left(\epsilon_{11}, \epsilon_{12}, \epsilon_{13}, \epsilon_{21}, \epsilon_{22}, \epsilon_{23}\right)\) in measurement model. Another condition proposed to Table 2.

3. Measure manifest and latent variables for measurement and structural model according to Structural Equation Modeling (Figure 1).

```markdown
Figure 1: Structural Equation Modeling for Simulation
```

The combinations of data simulation in this paper considered overlap and variance of residual. The first condition simulated 0%, 6%, and 26% overlap between the two equaled sample size. This 0% overlap showed none of units formed an overlap model \(n = 100, n_1 = n_2 = 50\). Otherwise, the 6% overlap showed 47 units for each segment one and segment two, there was also 6 units in overlap. The similar approachment were in other combinations.

### IV. RESULT AND DISCUSSION

**Estimator Path Coefficient in Structural Model**

Performance of segmentation PLS approaches used to capture the best estimator for some conditions. This paper focused on parameter estimator in structural model to determine the best method capturing segmentation in heterogeneous units. The range of the estimator given Global Model and the segmentation approaches were compared especially for Scenario 1.

The variations of estimator given in Scenario 1 (overlap 0%) and variance of residual 0.1. Figure 4 showed that the average of \(\hat{Y}_1\) in FIMIX-PLS closer to path coefficient simulation \((\gamma_1^{(1)} = 0.9)\) dan \(\gamma_1^{(2)} = 0.1\) and also it is smaller variation not only in path coefficient of segment 1 but also the coefficient of segment 2.

```markdown
Figure 4: Boxplots of the estimator \(\gamma_1^{(1)}\) and \(\gamma_1^{(2)}\) in condition Scenario 1 and variance of residual 0.1
```

The estimator of \(\gamma_2\) in segment 1 in Fimix PLS showed the sample mean estimator closest to path coefficient in structural model (Table 1) and smaller its variation. In segment 1, path coefficient in structural model for Fimix PLS and Rebus PLS were in parameter \(\gamma_2^{(1)}\) but in segment 2, both were less that its simulation path coefficient \(\gamma_2^{(2)} = 0.9\).

```markdown
Figure 5: Boxplots of the estimator \(\gamma_2^{(1)}\) and \(\gamma_2^{(2)}\) in condition Scenario 1 and variance of residual 0.1
```

**Comparison Performance Segmentation Approaches**

Segmentation was simulated in grouping some segments according to path coefficient in structural model differences. The performances in each method focused on
the differences. In this paper, mean bias absolute (MBA) depicted differences between path coefficient and the characteristic given (Table 1). The best method had the less MBA.

![Figure 6: Comparison of MBAs for each variance of residual in every Schene](image)

Figure 6 showed comparison of MABs of Fimix PLS and Rebus PLS in variance of residual conditions. Rebus PLS was indicated decreased MBAs otherwise increased constantly. In other ways, Fimix PLS gave increasing MBAs as high as variance of residual.

MBA of Fimix PLS and Rebus PLS in all condition were compare to determine the best method. Table 3 showed Fimix PLS had fewer of MBA than Rebus PLS. It meant the best method to estimate path coefficient in structural model was Fimix PLS.

<table>
<thead>
<tr>
<th>TABLE 3</th>
<th>MBA IN STRUCTURAL MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario</td>
<td>Overlap</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>Scenario 1</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>6%</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 3</td>
<td>26%</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V. CONCLUSION

Combination of data simulation sized 100 and it replicated for 10 times. The units divided equally into two groups and modeling in its characteristic path coefficient in structural model. The simulation were also in any condition of overlapping and variance of residual. The performance based on mean bias absolute (MBA) in structural model. Rebus PLS is more stable in decreasing and increasing of MBA for both conditions. Nevertheless, Fimix PLS given the best performance in both conditions.

REFERENCES