# Partial Least Squares (PLS): Path Modeling

Method Talk
Winter Term 2015/16
Pascal Stichler

# Outline

# Today's Lecture

## Objectives

1. Evaluate when to use PLS
2. Learn how PLS works and how to use it
3. Investigate how to evaluate a PLS model, interpret the results and adjust the model accordingly

# Outline

# PLS: A silver bullet?

Partial Least Squares Path Modeling is a statistical data analysis methodology that exists at the intersection of Regression Models, Structural Equation Models, and Multiple Table Analysis methods [9]

**Goal**: Use theoretical knowledge about structure of latent variables to predict indicators based on data

- Doing so with least possible distribution assumptions
- PLS-PM is known under several names: PLS-PM, PLS-SEM, component-based structural equation modeling, *projection to latent structures*, soft modeling etc.
- Developed by Herman Wold in the mid 1960s under the term of "soft modeling" [14]
- After initial introduction and discussions it received little attention until the late 1990s, however since then sharply rising interest

# Why use PLS?

PLS-PM is worth considering when ...

**Structural model**

- . . . you have a theoretical model that involves latent variables
- . . . the phenomenon you investigate is relatively new and measurement models need to be newly developed
- . . . the structural equation model is complex with a large number of latent variables and indicator variables [12]

**Observed variables**

- . . . you have small sample sets (e. g. more variables than observations) [7]
- . . . you have non-normal distributed data
- . . . you have multicollinearity problems
- . . . you have formative and reflective measures (to be discussed)
- . . . you need minimum requirements regarding measurement scales (e. g. ratio and nominal variables)
- . . . you need minimum requirements regarding residuals distribution [1]

# Outline

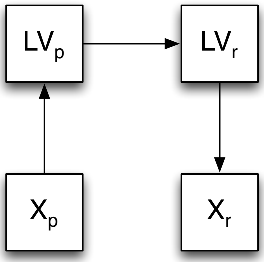
# General Overview

**Types of PLS:**

- **PLS-Path Modeling:**
  Component-based modeling based on theoretical structure model
  *Mainly used in: social sciences, econometrics, marketing and strategic management*
- **PLS-Regression:**
  Regression based approach investigating the linear relationship between multiple independent variables and dependent variable(s)
  *Mainly used in: chemometrics, bioinformatics, sensometrics, neuroscience and anthropology*
- **OPLS:** Orthognal projection improves interpretability
- **PLS-DA:** Used when $X_r$ is categorial

---

- **CB-SEM:** Covariance-based structural equation modelling




- Predictors $X_p \subset X$
- Responses $X_r \subset X$ with $X_p \cap X_r = \varnothing$
- Exogenous latent variables $LV_p \subset LV$
- Endogenous latent variables $LV_r \subset LV$ with $LV_p \cap LV_r = \varnothing$

# PLS-PM vs. CB-SEM

Both methods differ from statistical point of view. Hence, neither of the techniques is generally superior to the other and neither of them is appropriate for all situations. In general, the strenghts of PLS-SEM are CB-SEM's weaknesses, and visa versa. [3]

**PLS-PM (PLS-SEM)**

*Variance-based*

- The goal is prediction and theory development
- Formatively measured constructs are part of the structural model
- The structural model is complex
- The sample size is small and/or the data are non-normally distributed
- The plan is to use latent variable scores in subsequent analyses
- Available Software: SmartPLS, PLSGraph, R packages (plspm) etc.

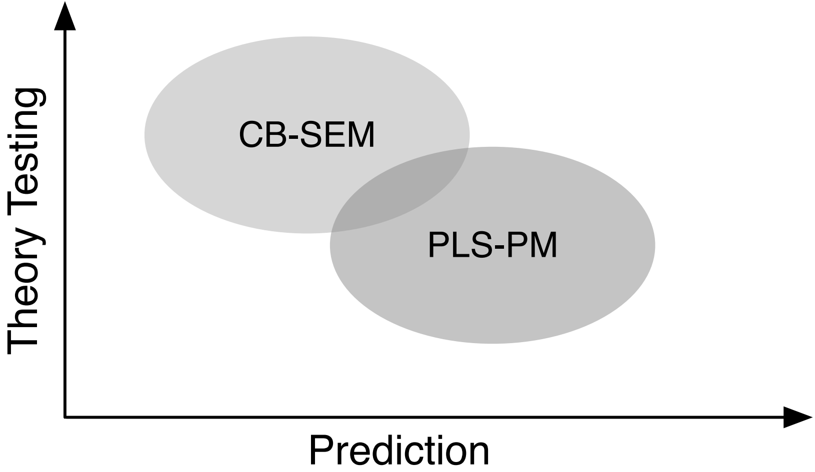
**CB-SEM**

*Covariance-based*

- The goal is theory testing, theory confirmation, or the comparison of alternative theories
- Error terms require additional specification, such as the covariation
- The structural model has non-recursive relationships
- The research requires a global goodness-of-fit criterion
- Available Software: LISREL, AMOS, EQS etc.

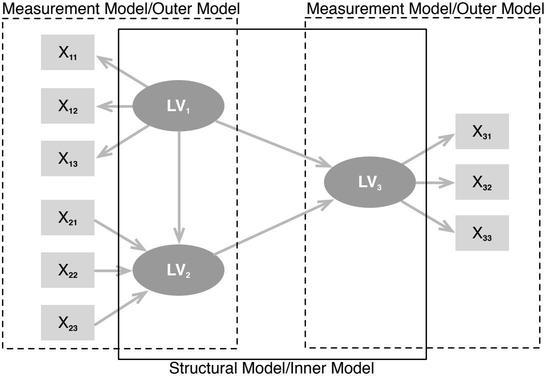
*Based on [8], [4], [11]*

# Comparison



# Outline

# Exemplary Model




**Formal definition:**

- $X$ data set with $n$ observations and $m$ variables
- $X$ can be divided into $J$ exclusive blocks with $K$ variables each $X_{1,1} \dots X_{J,K}$ etc.
- Each block $X_j$ associated with $LV_j$; estimation of variable ("score") denoted by $\widehat{LV_j} = Y_j$
- LV1 and LV3: reflective blocks; LV2: formative block [9]

# Structural Model (Inner Model)

1. **Linear Relationship**

   All relationships are considered linear relationships and can be noted as

   $$LV_j = \beta_0 + \sum_{i \to j} \beta_{ji} LV_i + \varepsilon_j$$

   The coefficients $\beta_{ji}$ represent the path coefficients

2. **Recursive Model mandatory**

   Causality flow must be unidirectional (no loops)

3. **Regression Specification (Predictor Specification)**

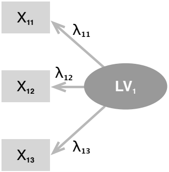
   $$E(LV_j | LV_i) = \beta_{0i} + \sum_{i \to j} \beta_{ji} LV_i$$

   Specifying that the regression has to be linear under the assumption that

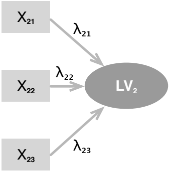
   $\mathrm{cov}(LV_j, \varepsilon_j) = 0$ and $\overline{\varepsilon_j} = 0$

# Measurement Model (Outer Model)

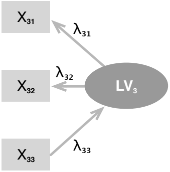
**Reflective Indicators**




**Formative Indicators**




**MIMIC***




| | Reflective Indicators | Formative Indicators | MIMIC* |
|---|---|---|---|
| ► **Linear relationships:** | $X_{jk} = \lambda_{0jk} + \lambda_{jk} LV_j + \varepsilon_{jk}$ ($\lambda_{jk}$ is called loading) | $LV_j = \lambda_{0j} + \lambda_{jk} X_{jk} + \varepsilon_j$ | equivalent to reflective and formative (depending on indicator) |
| ► **Regression Specification:** | $E(X_{jk} \| LV_j) = \lambda_{0jk} + \lambda_{jk} LV_j$ | $E(LV_j \| X_{jk}) = \lambda_{0j} + \lambda_{jk} X_{jk}$ | equivalent to reflective and formative (depending on indicator) |
| ► **Characteristics:** | - Unidimensional<br>- Correlated<br>- $X_{jk}$ "fully relevant" | - Multidimensional<br>- Uncorrelated<br>- $X_{jk}$ "partly relevant" | In R package *plspm* not possible |

*multiple effect indicators for multiple causes

# Weight Relations (Scores)

- The latent variables are only virtual entities
- However, as all linear relations depend on the latent variables, they need a representation: Weight Relations

$$\textbf{Score: } \widehat{LV}_j = Y_j = \sum_k w_{jk} X_{jk}$$

- The score, as a representation of the latent variable, is calculated as the sum of its indicators (similar to the approach in principal component analysis)
- Because of this PLS is called a component-based approach

# Outline

# PLS-PM Algorithm Overview

1. Stage: Get the weights to compute latent variable scores
   $\rightarrow$ Most important and most difficult

2. Stage: Estimate the path coefficients (inner model)
   $\rightarrow$ Usually done via OLS

3. Stage: Obtain the loadings (outer model)
   $\rightarrow$ Calculation of correlations

# Stage 1: Latent Variable Scores

**Start:**
initial arbitrary outer weights (e.g. $w_{jk} = 1$)

**Step 1:**
Compute the external approximation of latent variables

$$Y_j = \sum_k w_{jk} X_{jk}$$

**Step 2:**
Obtain inner weights $e_{ij}$

**Inner weighting schemes:**

- Centroid scheme
- Factor scheme
- Path scheme

**Step 3:**
Compute the internal approximation of latent variables

$$Z_j = \sum_{i \leftrightarrow j} e_{ij} Y_i$$

**Step 4:**
Calculate new outer weights $w_{jk}$

**Mode A:** Simple regression
**Mode B:** Multiple regression
**(Mode C:)** Combination

Check for convergence of outer weights

# Stage 2 & 3

**2. Stage: Path Coefficients**

The path coefficient estimates $\widehat{\beta}_{ji} = B_{ji}$ are calculated usually using ordinary least squares in the multiple regression of $Y_i$ on the $Y_j$'s related with it

$$Y_j = \sum_{i \to j} \widehat{\beta}_{ji} Y_i$$

In case high multicollinearity occurs PLS regression can also be applied [11]

**3. Stage: Loadings**

For convenience and simplicity reasons, loadings are preferably calculated as correlations between a latent variable and its indicators:

$$\widehat{\lambda_{jk}} = cor(X_{jk}, Y_j)$$

# PLS-PM usage in R (package plspm)

## Parameters to define the PLS Path Model

| | |
|---|---|
| `Data` | Data for the model |
| `path_matrix` | Definition of inner model |
| `blocks` | List definitng the blocks of variables of the outer model |
| `scaling` | List defining the measurment scale of variables for non-metric data |
| `modes` | Vector defining the measuremnt mode of each block |

## Parameters related to the PLS-PM algorithm

| | |
|---|---|
| `scheme` | Inner path weighting scheme |
| `scaled` | Indicates whether the data should be standardized |
| `tol` | Tolerance threshold for checking convergence of the iterative stages |
| `maxiter` | maximum number of iterations |
| `plscomp` | Indicates the number of PLS components when handling non-metric data |

## Additional parameters

| | |
|---|---|
| `boot.val` | Indicates whether bootstrap validation must be performed |
| `br` | Number of bootstrap resamples |
| `dataset` | Indicates whether the data matrix should be retrieved |

# Outline

# Interpreting the Results

In PLS the real challenge is interpreting the results and making well-founded adjustments the model [9], p. 54

```
Partial Least Squares Path Modeling (PLS-PM)
---------------------------------------------
   NAME              DESCRIPTION
1  $outer_model      outer model
2  $inner_model      inner model
3  $path_coefs       path coefficients matrix
4  $scores           latent variable scores
5  $crossloadings    cross-loadings
6  $inner_summary    summary inner model
7  $effects          total effects
8  $unidim           unidimensionality
9  $gof              goodness-of-fit
10 $boot             bootstrap results
11 $data             data matrix
---------------------------------------------
You can also use the function 'summary'
```

**Steps of Model Assessment:**

1. Assessment Measurement Model (Outer Model)
2. Assessment Structural Model (Inner Model)

(It is important to keep this order due to model dependencies)

# 1. Measurement Model Assessment (Outer Model)

- **Formative Blocks:** Evaluation relatively straightforward
- **Reflective Blocks:** Evaluation rather complex $\Longrightarrow$ *Test theory* applied

**Formative Blocks:**
*Variables are considered as causing the latent variable*

- They do not necessarily measure the same underlying construct
- Not supposed to be correlated
- Compare outer weights to check which indicator contributes most efficiently
- Elimination of variables should be based on multicollinearity

**Reflective Blocks:**
*Variables are considered as measuring the same underlying construct*

- Hence they need a strong mutual association
- Further they should be strongly related to its latent variable

1. Unidimensionality of indicators
2. Indicators well explained
3. Constructs differ from each other

# Deep Dive: Reflective Indicators

1. **Unidimensionality of indicators:** All for one and one for all
   - (a) **Cronbach's alpha**
     Measures the average inter-variable correlation
     *(considered good if > 0.7)*
   - (b) **Dillon-Goldstein's rho**
     Focus on the variance of the sum of variables (considered a better indicator than Cronbach's alpha ([1], p.320)
     *(considered good if > 0.7)*
     (see [11], [13] p. 50 for formal definition)
   - (c) **First eigenvalue**
     First eigenvalue of correlation matrix should be larger than one and second one significantly smaller *(preferably smaller than 1)*

2. **Loadings & Communalities:** Indicators well explained
   - Loadings are considered for each indicator *(considered good if > 0.7)*
   - Communalities (squared loadings): amount of indicator variance explained by its corresponding LV

# Deep Dive: Reflective Indicators

3. **Cross-loadings:** Constructs differ from each other
   cross-loadings $\hat{=}$ loadings of an indicator with the rest of the latent variables
   **Goal:** Ensure that shared variance between construct and its indicators is higher than for other constructs (no "traitor" indicators)
   $\Longrightarrow$ Loadings should always be highest for the respective block

[...]$crossloadings

| | name | block | sentiment | market_data | eco_data | index |
|---|---|---|---|---|---|---|
| 1 | AllNetOptimismHEOfWords | sentiment | 0.1328359 | -0.05301580 | 0.08488984 | 0.0680479 |
| 2 | IfoGK | sentiment | 0.9623578 | 0.09879567 | 0.86490826 | 0.7715654 |
| 3 | ZEWEconomicSituation | sentiment | 0.9690857 | 0.33325094 | 0.90054235 | 0.8756996 |
| 4 | CV.Price.DollarEuroExchangeData | market_data | 0.2277527 | 1.00000000 | 0.38186338 | 0.4356771 |
| 5 | CV.Price.GdpUsaData | eco_data | 0.3850477 | -0.19944865 | 0.30162050 | 0.2575881 |
| 6 | IndustryProduction | eco_data | 0.8419315 | 0.46058092 | 0.95782799 | 0.8547013 |
| 7 | CV.Price.CdaxData | index | 0.8541450 | 0.43567709 | 0.88901138 | 1.0000000 |

# 2. Structural Model Assessment (Inner Model)

## Standard OLS regression output:

```
$index
                  Estimate  Std. Error      t value      Pr(>|t|)
Intercept      -2.394262e-15 0.002646805 -9.045858e-13  1.000000e+00
eco_data        6.296528e-01 0.007855595  8.015342e+01  0.000000e+00
sentiment      -2.169772e-02 0.002665119 -8.141370e+00  4.226202e-16
eco_indicators  1.502800e-01 0.006732013  2.232319e+01 1.504294e-108
UkUsMarkets     1.916292e-01 0.007661909  2.501063e+01 3.492772e-135
```

## 3 further indicators of model quality:

- $R^2$ **determination coeffcient:** Amount of variance of endogenous LVs explained by its independent LVs *(considered low below 0.3 and high above 0.6)*
- **Redundancy Index:** Amount of variance in the endogenous block that explained by its independent LVs (defined as $Rd(LV_j, x_{jk}) = loading^2_{jk} R^2_j$)
- **Goodness-of-Fit (GoF):** No single criterion exists for overall quality of a model. GoF as a pseudo criterion:

$$GoF = \sqrt{\overline{communality} \times \overline{R^2}}$$ *(considered good if >0.7)* [10] [11]

- **Validation:** Resampling (bootstrapping, jackknifing) possible; more traditional approaches are not (as there are no assumptions made on the distribution)

# Outline

# Summary: PLS

## Advisable for the following conditions (based on [8])

| | |
|---|---|
| Focus | Prediction and theory development |
| Distribution | Minimum assumptions made regarding indicator distribution |
| Sample size | Small sample size possible (however questioned in literature [2], [6], [5]) |

## Model definition

| | |
|---|---|
| Indicators | Define blocks of variables and respective latent variables |
| Measurement Model | Define relations (formative/reflective) |
| Structural Model | Define internal model |

## Interpreting the results

| | |
|---|---|
| Measurement Model (formative) | Eliminate multicollinearity |
| Measurement Model (reflective) | Unidimensionality, loadings & communalities and cross-loadings |
| Structural Model | Consider $R^2$, redundancy index and GoF |
| Validation | Apply resampling (bootstrapping, jackknifing) |

# Bibliography I


- W. W. CHIN. The partial least squares approach to structural equation modeling. In: Modern methods for business research, Vol. 295, No. 2 (1998), pp. 295–336.
- D. GOODHUE, W. LEWIS, and R. THOMPSON. PLS, small sample size, and statistical power in MIS research. In: System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on. Vol. 8. IEEE. 2006, 202b–202b.
- J. F. HAIR JR et al. A primer on partial least squares structural equation modeling (PLS-SEM). Sage Publications, 2013.
- J. F. HAIR, C. M. RINGLE, and M. SARSTEDT. PLS-SEM: Indeed a silver bullet. In: Journal of Marketing Theory and Practice, Vol. 19, No. 2 (2011), pp. 139–152.
- G. A. MARCOULIDES, W. W. CHIN, and C. SAUNDERS. A critical look at partial least squares modeling. In: Mis Quarterly (2009), pp. 171–175.

# Bibliography II


- G. A. MARCOULIDES and C. SAUNDERS. Editor's comments: PLS: a silver bullet? In: MIS quarterly, Vol. 30, No. 2 (2006), pp. iii–ix.
- B.-H. MEVIK and R. WEHRENS. The pls package: principal component and partial least squares regression in R. In: Journal of Statistical Software, Vol. 18, No. 2 (2007), pp. 1–24.
- W. REINARTZ, M. HAENLEIN, and J. HENSELER. An empirical comparison of the efficacy of covariance-based and variance-based SEM. In: International Journal of research in Marketing, Vol. 26, No. 4 (2009), pp. 332–344.
- G. SANCHEZ. PLS path modeling with R. In: Online, January (2013).
- M. TENENHAUS, S. AMATO, and V ESPOSITO VINZI. A global goodness-of-fit index for PLS structural equation modelling. In: Proceedings of the XLII SIS scientific meeting. Vol. 1. CLEUP Padova. 2004, pp. 739–742.

# Bibliography III


- M. TENENHAUS et al. PLS path modeling. In: Computational statistics & data analysis, Vol. 48, No. 1 (2005), pp. 159–205.
- N. URBACH and F. AHLEMANN. Structural equation modeling in information systems research using partial least squares. In: Journal of Information Technology Theory and Application, Vol. 11, No. 2 (2010), pp. 5–40.
- V. E. VINZI, L. TRINCHERA, and S. AMATO. PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement. In: Handbook of partial least squares. Springer, 2010, pp. 47–82.
- H. WOLD et al. Estimation of principal components and related models by iterative least squares. In: Multivariate analysis, Vol. 1 (1966), pp. 391–420.