Algorithme et critères du PLS Path Modeling

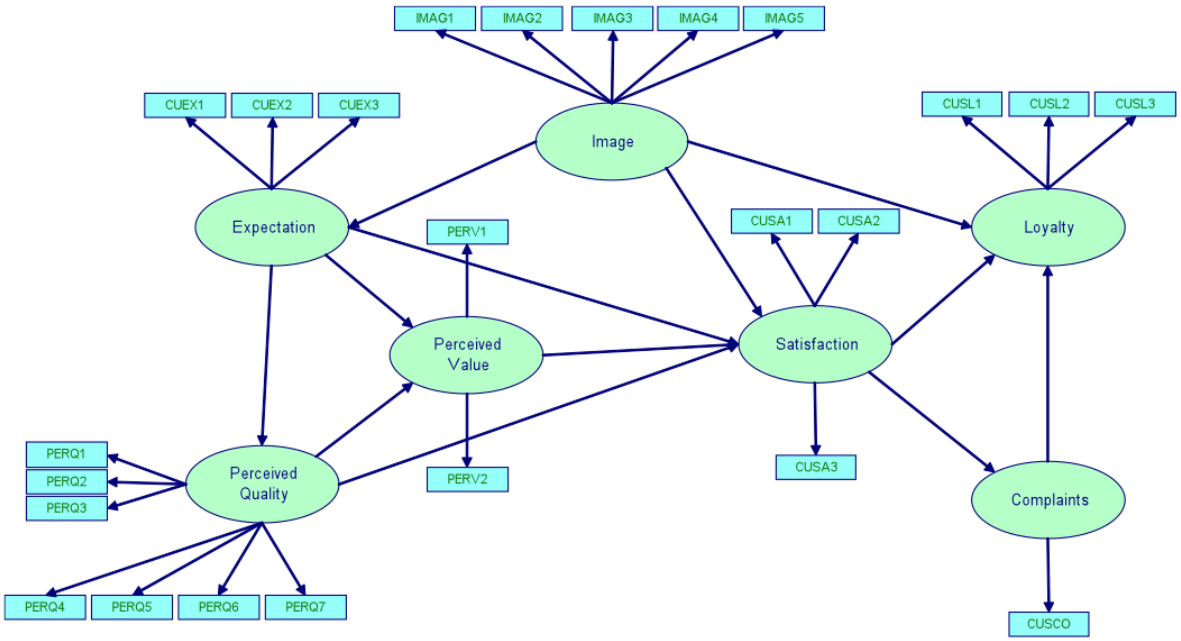
Giorgio Russolillo
giorgio.russolillo@cnam.fr

Laboratoire CEDRIC, Conservatiore National des Arts et Métiers, France

le cnam

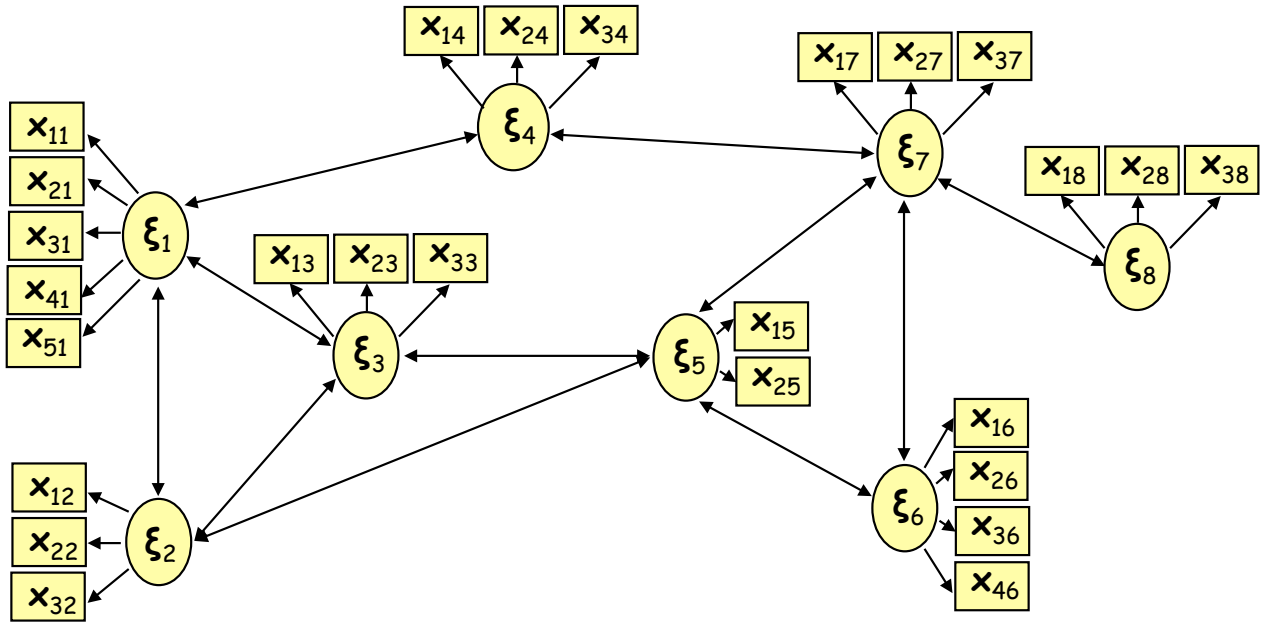
PLS Path Modeling

Le PLS-PM permet la modélisation d'un rseau de relations statistiques entre variables latentes indirectement mesurés au moyen d'ensembles d'indicateurs (les variables manifestes)



le cnam

PLS Path Modeling: notations

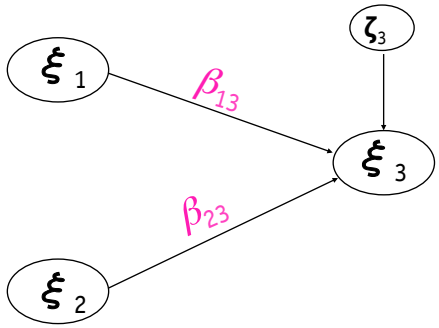


- P variables manifestes (MV) observées sur n individus $\Rightarrow x_{pq}$ MV générique
- Q blocs de variables manifestes. Le bloc q est composé par P_q variables manifestes $\Rightarrow \sum_{q=1}^{Q} P_q = P$
- Q variables latentes (LV) $\Rightarrow \xi_q$ LV générique

le cnam

Le modèle structurel

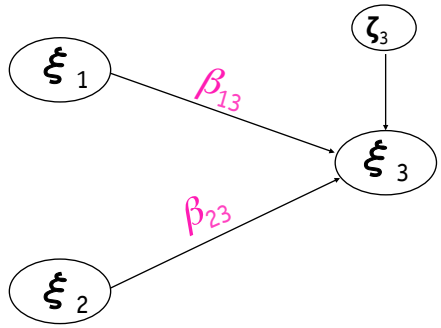
Le modèle structurel décrit les relations entre les variables latentes.



le cnam

Le modèle structurel

Le modèle structurel décrit les relations entre les variables latentes.



Pour chaque LV endogne le modle structurel peut être crit comme:

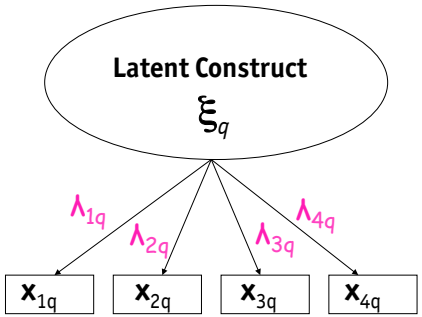
$$\boldsymbol{\xi}_j = \sum_{q:\boldsymbol{\xi}_q \to \boldsymbol{\xi}_j} \beta_{qj} \boldsymbol{\xi}_q + \boldsymbol{\zeta}_j$$

where β_{qj} est le path-coefficient qui relie la q-iéme LV á la j-éme LV endogne

le cnam

Le modèle de mesure

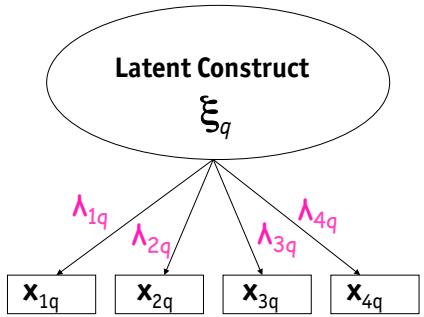
Le **modèle de mesure** décrit les relations entre **les variables manifestes et la variable latente correspondante**.



le cnam

Le modèle de mesure

Le **modèle de mesure** décrit les relations entre **les variables manifestes et la variable latente correspondante**.



Pour chaque variable manifeste le modle de mesure peut être crit comme:

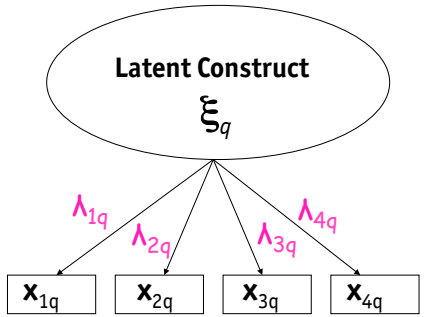
$$x_{pq} = \lambda_{pq}\boldsymbol{\xi}_q + \boldsymbol{\epsilon}_{pq}$$

o λ_{pq} est le loading (coefficient de rgression de la q-iéme LV sur la p-iéme MV)

le cnam

Le modèle de mesure

Le **modèle de mesure** décrit les relations entre **les variables manifestes et la variable latente correspondante**.



Pour chaque variable manifeste le modle de mesure peut être crit comme:

$$x_{pq} = \lambda_{pq}\boldsymbol{\xi}_q + \boldsymbol{\epsilon}_{pq}$$

o λ_{pq} est le loading (coefficient de rgression de la q-iéme LV sur la p-iéme MV)

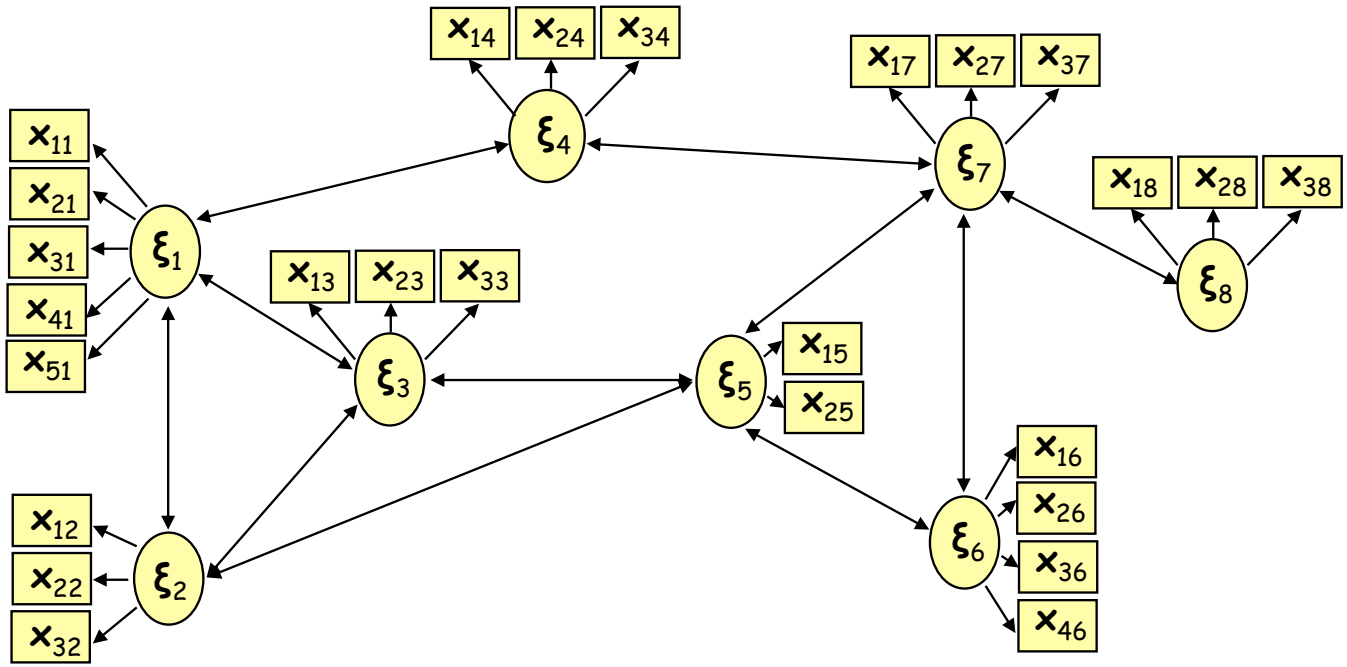
L'approche PLS est base sur les composantes: la LV est estimé sous forme d'une combinaison linéare des MV de son bloc (weight relation):

$$\hat{\boldsymbol{\xi}}_q = \sum_{p=1}^{P_q} w_{pq}x_{pq}$$

le cnam

PLS Path Modeling: l'algorithme

1. Obtenir Q vecteurs de poids w_q travers la boucle itrative PLS
2. Calculer les scores des variables latentes
3. Calculer les loadings et les path coefficients



le cnam

PLS Path Modeling: l'algorithme

Le but de la boucle PLS est de définir une **système de pondérations** appliquer chaque bloc de MV en vue d'estimer les LV correspondantes, selon la relation:

$$\hat{\boldsymbol{\xi}}_q = \sum_{p=1}^{P_q} w_{pq} x_{pq}$$

le cnam

PLS Path Modeling: l'algorithme

Le but de la boucle PLS est de définir une **système de pondérations** appliquer chaque bloc de MV en vue d'estimer les LV correspondantes, selon la relation:

$$\hat{\xi}_q = \sum_{p=1}^{P_q} w_{pq} x_{pq}$$

L'algorithme itératif est basé sur deux étapes principales; dans chaque étape la variable latente est approximèe par une proxy.

- **Estimation externe**
  $\Rightarrow t_q =$ somme pondérée des ses VM
- **Estimation interne**
  $\Rightarrow z_q =$ somme pondérée des estimations externes des LV reliées

le cnam

L'estimation externe

estimation externe

La proxy t_q est une somme pondérée des ses VM

$$t_q \propto \pm \sum_{p=1}^{P_q} w_{pq} x_{pq} = \pm X_q w_q$$

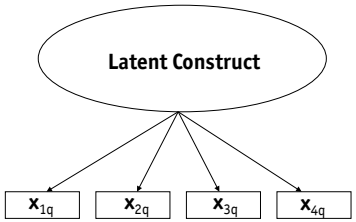
le cnam

L'estimation externe

estimation externe

La proxy t_q est une somme pondérée des ses VM

$$t_q \propto \pm \sum_{p=1}^{P_q} w_{pq} x_{pq} = \pm X_q w_q$$



Mode A (flèches oriéntées vers l'extérieur - **reflective**):

$$w_{pq} = X_q^T z_q / (z_q^T z_q)$$

- Les indicateurs devraient **covarier**
- Plusieurs régressions simples
- Variance expliquée (**AVE**, communality)

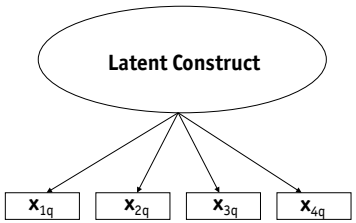
le cnam

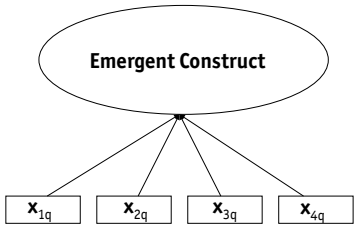
L'estimation externe

estimation externe

La proxy t_q est une somme pondérée des ses VM

$$t_q \propto \pm \sum_{p=1}^{P_q} w_{pq} x_{pq} = \pm X_q w_q$$





Mode A (flèches oriéntées vers l'extérieur - **reflective**):

$$w_{pq} = X_q^T z_q / (z_q^T z_q)$$

- Les indicateurs devraient **covarier**
- Plusieurs régressions simples
- Variance expliquée (**AVE**, communality)

Mode B (flèches oriéntées vers l'intérieur - **formative**):

$$w_{pq} = (X_q^T X_q)^{-1} X_q^T z_q$$

- Les indicateurs ne devraient pas **covarier**
- Une régression multiple (multicollinearité?)
- Structural Predictions (R^2 plus élevées pour les variables latentes endogènes)

le cnam

L'estimation interne

Inner Estimation

La proxy z_q est une somme pondérée des estimations externes des LV reliées

$$z_q \propto \sum_{q'=1}^{Q'} e_{qq'} t_{q'}$$

le cnam

L'estimation interne

Inner Estimation

La proxy z_q est une somme pondérée des estimations externes des LV reliées

$$z_q \propto \sum_{q'=1}^{Q'} e_{qq'} t_{q'}$$

1. **Schéma Centroide**

le cnam

L'estimation interne

Inner Estimation

La proxy z_q est une somme pondérée des estimations externes des LV reliées

$$z_q \propto \sum_{q'=1}^{Q'} e_{qq'} t_{q'}$$

1. **Schéma Centroide**
2. **Schéma Factoriel**

le cnam

L'estimation interne

Inner Estimation

La proxy z_q est une somme pondérée des estimations externes des LV reliées

$$z_q \propto \sum_{q'=1}^{Q'} e_{qq'} t_{q'}$$

1. **Schéma Centroide**
2. **Schéma Factoriel**
3. **Schéma Structurel (ou Path weighting)**

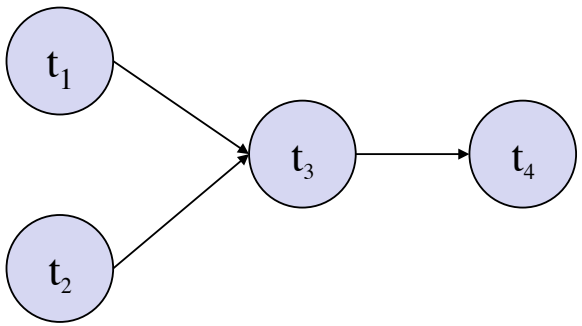
le cnam

L'estimation interne

Inner Estimation

La proxy z_q est une somme pondérée des estimations externes des LV reliées

$$z_q \propto \sum_{q'=1}^{Q'} e_{qq'} t_{q'}$$



1. **Schéma Centroide**
2. **Schéma Factoriel**
3. **Schéma Structurel (ou Path weighting)**

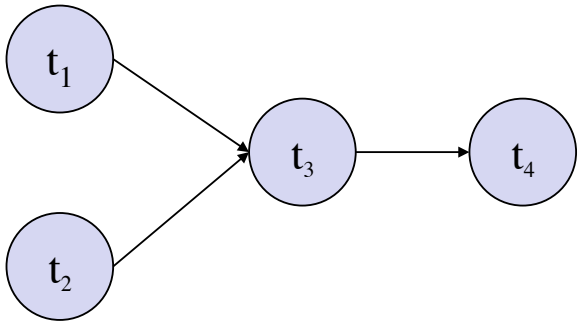
le cnam

L'estimation interne

Inner Estimation

La proxy z_q est une somme pondérée des estimations externes des LV reliées

$$z_q \propto \sum_{q'=1}^{Q'} e_{qq'} t_{q'}$$



1. **Schéma Centroide**
   $z_3 = e_{31} t_1 + e_{32} t_2 + e_{34} t_4$ where $e_{qq'} = sign(cor(t_q, t_{q'}))$
2. **Schéma Factoriel**
3. **Schéma Structurel (ou Path weighting)**

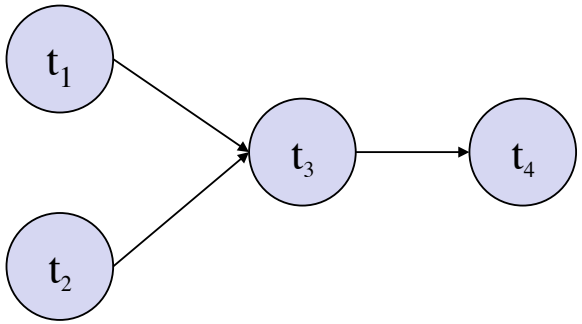
le cnam

L'estimation interne

Inner Estimation

La proxy z_q est une somme pondérée des estimations externes des LV reliées

$$z_q \propto \sum_{q'=1}^{Q'} e_{qq'} t_{q'}$$



1. **Schéma Centroide**
   $z_3 = e_{31} t_1 + e_{32} t_2 + e_{34} t_4$ where $e_{qq'} = sign(cor(t_q, t_{q'}))$
2. **Schéma Factoriel**
   $z_3 = cor(t_3, t_1) t_1 + cor(t_3, t_2) t_2 + cor(t_{q3}, t_4) t_4$
3. **Schéma Structurel (ou Path weighting)**

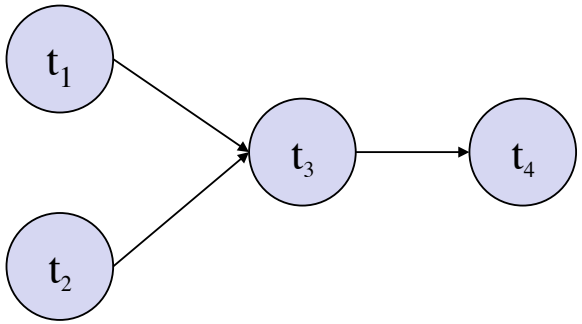
le cnam

L'estimation interne

Inner Estimation

La proxy z_q est une somme pondérée des estimations externes des LV reliées

$$z_q \propto \sum_{q'=1}^{Q'} e_{qq'} t_{q'}$$

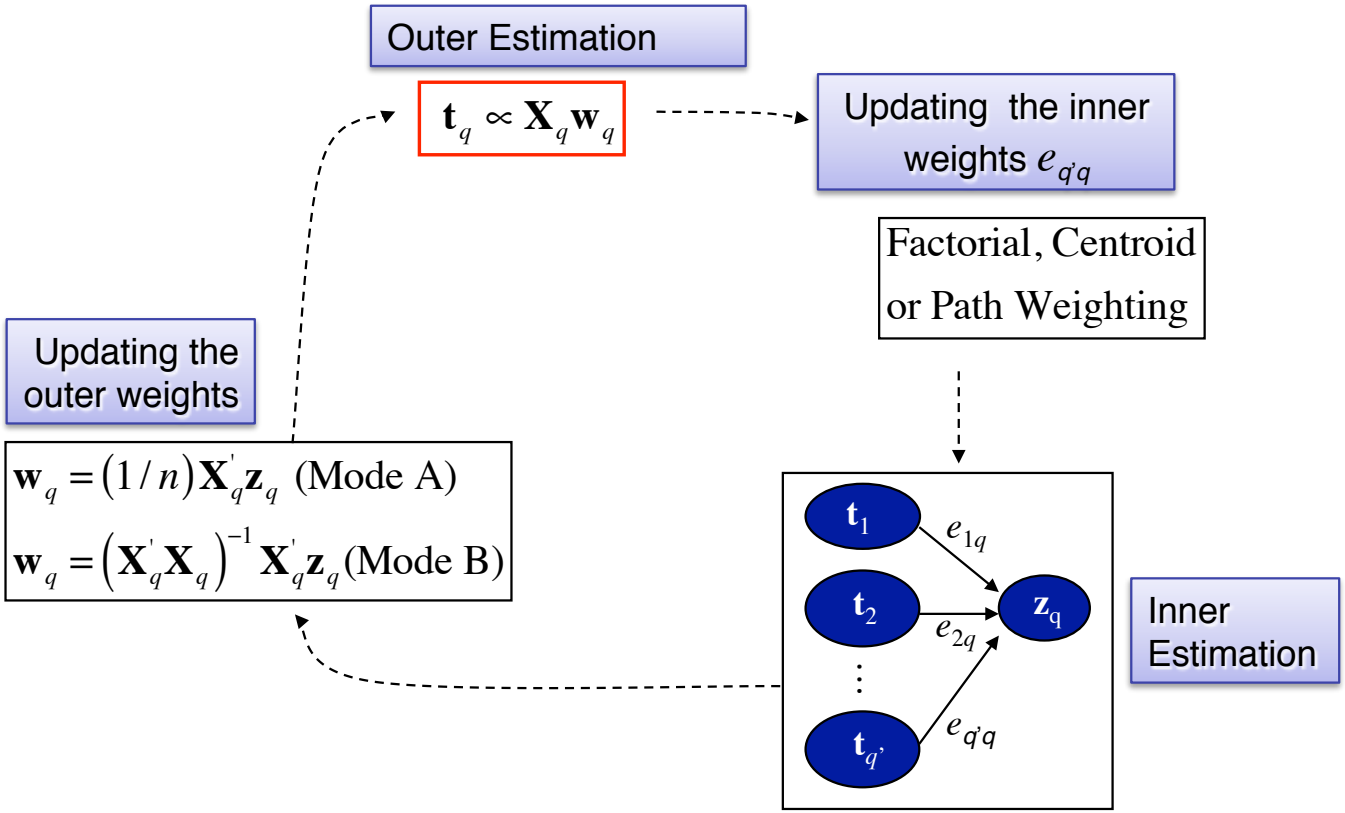


1. **Schéma Centroide**
   $z_3 = e_{31} t_1 + e_{32} t_2 + e_{34} t_4$ where $e_{qq'} = sign(cor(t_q, t_{q'}))$
2. **Schéma Factoriel**
   $z_3 = cor(t_3, t_1) t_1 + cor(t_3, t_2) t_2 + cor(t_{q3}, t_4) t_4$
3. **Schéma Structurel (ou Path weighting)**
   $z_3 = \beta_{31} t_1 + \beta_{32} t_2 + cor(t_{q3}, t_4) t_4$

   où les betas sont le coefficients de régression du modèle:
   $t_3 = \beta_{31} t_1 + \beta_{32} t_2 + \zeta_3$

le cnam

PLS-PM Iteration



Critères d'optimisation de l'algorithme

Glang (1988) et Mathes (1993) ont montré que l'équation stationnaire du PLS-PM Mode B résoudre le critre:

PLS-PM Mode B

$$\arg\max_{\forall w_q} \sum_q c_{qq'} g(\text{cov}(X_q w_q, X_{q'} w_{q'})) \qquad \text{s.t. } \|X_q w_q\| = 1$$

$$\begin{cases} c_{qq'} = 1 & \text{si } X_q \text{ and } X_{q'} \text{ sont relis} \\ c_{qq'} = 0 & \text{sinon} \end{cases} \qquad \begin{cases} g() = & \text{carré (Schéma Factoriel)} \\ g() = & \text{valeur absolue (Schéma Centroide)} \end{cases}$$

Hanafi (2007) a prouvé que l'algorithme PLS dans ce cas est monotone convergente à ces critères.

le cnam

Critères d'optimisation de l'algorithme

Dans sa thèse, Kramer (2007)

1. montre que l'algorithme PLS, quand le Mode A est utilisé pour tous les blocs, n'est pas basé sur une équation stationnaire liée l'optimisation d'une fonction dérivable deux fois.
2. étend les résultats de Hanafi à un mode A modifié (connu dans la littrature comme **new mode A**), dans lequel une contrainte de normalisation est mise sur les poids extérieurs plutôt que sur les scores des variables latentes

Full New Mode A PLS-PM

$$\arg\max_{\forall w_q} \sum_q c_{qq'} g(\text{cov}(X_q w_q, X_{q'} w_{q'})) \qquad \text{s.t. } \|w_q\| = 1$$

le cnam

Critères d'optimisation de l'algorithme

Dans sa thèse, Kramer (2007)

1. montre que l'algorithme PLS, quand le Mode A est utilisé pour tous les blocs, n'est pas basé sur une équation stationnaire liée l'optimisation d'une fonction dérivable deux fois.
2. étend les résultats de Hanafi à un mode A modifié (connu dans la littrature comme **new mode A**), dans lequel une contrainte de normalisation est mise sur les poids extérieurs plutôt que sur les scores des variables latentes

Full New Mode A PLS-PM

$$\arg\max_{\forall w_q} \sum_q c_{qq'} g(\text{cov}(X_q w_q, X_{q'} w_{q'})) \qquad \text{s.t. } \|w_q\| = 1$$

Il a été montré de facon empirique que le mode A se rapproche de ce critère.

le cnam

Critères d'optimisation de l'algorithme

Tenenhaus et Tenenhaus (2011) ont proposé ce critère qui unifie les deux critères précédents:

critère général

$$\underset{\forall X_q, w_q}{\arg\max} \sum_{q \neq q'} c_{qq'} g\{\text{cov}(X_q w_q, X_{q'} w_{q'})\}$$
$$\text{s.t.} \quad \tau_q \|w_q\| + (1 - \tau_q)\text{var}(x_{pq} w_q) = 1$$

$\begin{cases} c_{qq'} = 1 & \text{si } X_q \text{ et } X_{q'} \text{ sont reliés} \\ c_{qq'} = 0 & \text{sinon} \end{cases}$

$\begin{cases} g\{\} = & \text{carré (Schéma Factoriel)} \\ g\{\} = & \text{valeur absolue (Schéma Centroide)} \end{cases}$

$\begin{cases} \tau_q = 1 & \text{si Mode A} \\ \tau_q = 0 & \text{si Mode B} \end{cases}$

le cnam

Conclusions

- Le PLS-PM est utilisé comme une méthode alternative pour estimer un SEM
- il est (presque) jamais sous-identifié (règles habituelles de régression)
- c'est une technique distribution-free
- En tant que technique d'estimation, PLS promet que la consistence "at large".
- La prédiction (entendue comme reproduction de données) est considérée plus importante que l'estimation des paramètres
- Les scores des variables latentes son directement obtenus.
- Les propriétés de l'algorithme dépendent du schéma et du mode utilisés.

le cnam