Master 1 Ingénierie Statistique :
Supervised Study Project in Mathematics

Structural equation modeling and applications

May 23^{rd} 2023

Aëla JAGOT and Clément POUPELIN

Acknowledgement

We want to acknowledge Mr Jean-Michel Galharret for his accompaniment and availability all along this semester. He was both our advisor for this research work and our professor for the module about linear and logistic regression. Regular meetings and a well-defined guideline have permitted an approach to the notion of structural equation modeling in the best condition.

We also would like to acknowledge Mrs Anne Philippe who was our professor of linear and logistic regression as well as inferential statistics. She introduced to us these notions which were essential to carry out this research work.

Contents

1 Introduction

Structural equation modeling (SEM) can't be considered as one statistical technique like linear regression or logistic regression. It is more like a framework which integrate different multivariate techniques like for example path analysis, regression, factor analysis, and that into one model.

This project is made to be understood by a student who had been introduced to linear regression and generalized linear regression. So, for this research paper, we will presents structural equation modeling as a generalisation of linear regression with an insertion of latent variables.
More information about SEM will be provides in the resources in the section **Bibliography**.

The purpose here will be to see the similarities and differences between linear regression and SEM, understanding the new problems brought with this method and the importance in the specification and identification of the model.

Finally, to complete our researches, we will also propose an application of SEM in psychology using the language `R`. The study interrogates the connection between the notions of spirituality, social comparison, temporal comparison and self-esteem for individuals. The aim will be to compare different SEM with tests and criteria to conclude on connections between these notions.

2 Preamble on structural equation modeling

2.1 Context

Before talking about structural equation modeling (SEM), we need to make a quick reminder on regression.
Regression techniques are used to model the connection between a variable with one or more other explanatory variables which can be quantitative and qualitative. For the case of a multiple linear regression, we obtain this kind of equation :

$$Y = X\beta + \mathcal{E}$$

Here, the different variables have the following meanings :

- $Y \in \mathbb{R}^n$ with $n \in \mathbb{N}^*$ is the variable that we try to explain
- X is the matrix $n \times p$ composed of p explanatory variables (each variable is a column of the matrix)
- $\beta \in \mathbb{R}^p$ is the vector which contains the coefficient of regression that we try to estimate
- $\mathcal{E} \in \mathbb{R}^n$ is the vector of error which is uncorrelated with our explanatory variables.

For this kind of model and to estimate β, we need to establish hypothesis on the error. The most importantly is to suppose that $\mathbb{E}(\mathcal{E}) = 0$ and $Var(\mathcal{E}) = \sigma^2 I$ with I the identity matrix of size n. This second statement means that, for $i \in \{1, ..., n\}$, the errors $\mathcal{E}_i$ are uncorrelated and have the same variance.

Obviously, we can also make stronger hypothesis saying that $(\mathcal{E}_1, ..., \mathcal{E}_n)$ are independent and identically distributed random variables with $\mathbb{E}(\mathcal{E}) = 0$ and $Var(\mathcal{E}) < \infty$. We can even assume that $\mathcal{E}$ follow a normal distribution $\mathcal{N}(0, \sigma^2 I)$.

But, in some fields, the linear regression presents limitations. Indeed, in psychology or sociology, the variables are often not directly measurable. For example, if we want to use variables which represent intelligence or the trust we have in a government, we can't directly measure that properly.
It is there that SEM is going to be interesting. Indeed, that is easily to include this type of variables which are called latent variables and it also permit to propose a clear visualization of the direct and indirect effects of a variable in a multi-variable model.

2.2 Introduction to structural equation modeling

As explained during the introduction, structural equation modeling (SEM) refers to a family of statistical techniques and could be related to others notions such as covariance structure analysis, covariance structure modeling or analysis of covariance structures. As you may guess, with SEM we will focus especially on variance and covariance analysis. But before being able to study SEM, we need to introduce some vocubulary and representations.

First of all, the term **latent variable** refers to a concept which is not observable like intelligence or trust. That is the opposite to the term **observed/manifest variable**. We usually represent a latent variable in an oval and an observed variable in a square or rectangle.

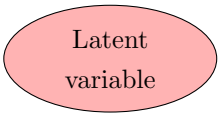


Figure 1 : Latent variable

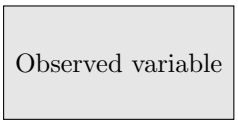


Figure 2 : Observed variable

Then, to build latent variables with observed variables, there exist two types of connections. Here, we are going the illustrate them using one latent variable and two observed variables. These last ones will be used as indicators for our latent variable. An indicator is considered as **formative** if the influence is exerted from the observed variable to the latent variable. Conversely, the indicator is considered as **reflective** if the influence is exerted from the latent variable to the observed variable.

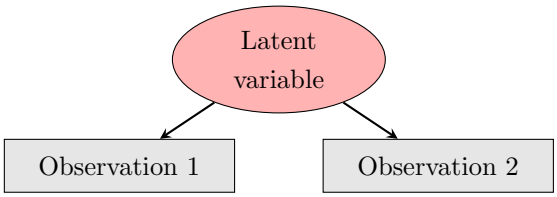


Figure 3 : The indicators are reflective

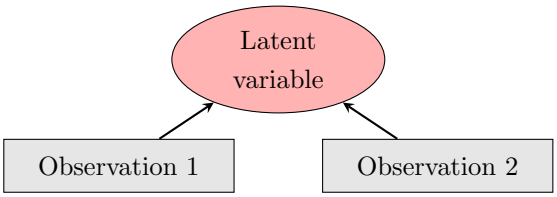


Figure 4 : The indicators are formative

It is also possible to have correlations between observed variables. This connection is often symbolized by a curve path.

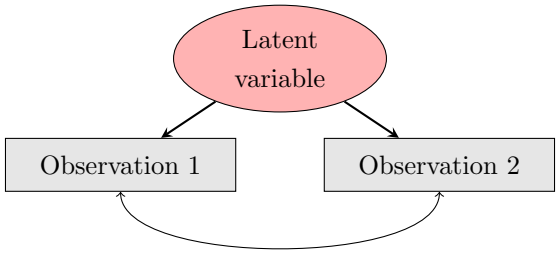


Figure 5 : Correlation between observed variables

Finally, a latent variable can have two roles. We can say that a variable is **exogenous** if it is not influenced by any other variable of the model and a variable is **endogenous** if it is influenced by at least one variable of the system.

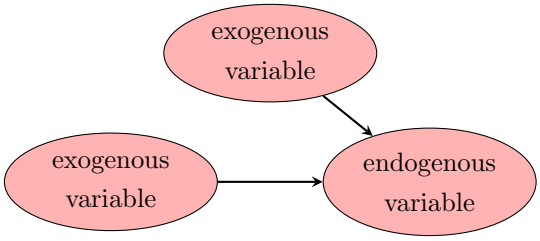


Figure 6 : Exogenous/endogenous variables (1)

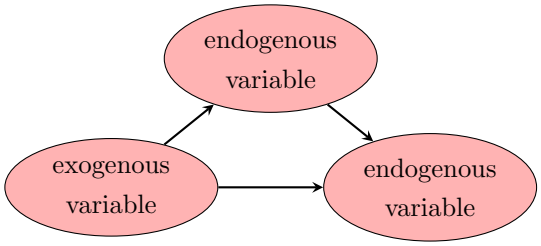


Figure 7 : Exogenous/endogenous variables (2)

Now, we are able to build a model of structural equation which will be composed of a measurement model and a structural model.
The measurement model is the place where are built the latent variables with the observed variables. The structural model represents the relations between the latent variables.

Here we have two types of models. The first one is called **recursive** because the connections between the variables are unidirectional and the second one is called **non-recursive** because the connections are not always unidirectional.

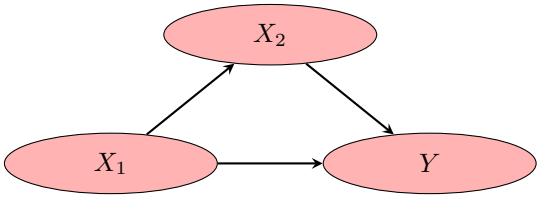


Figure 8 : Recursive model

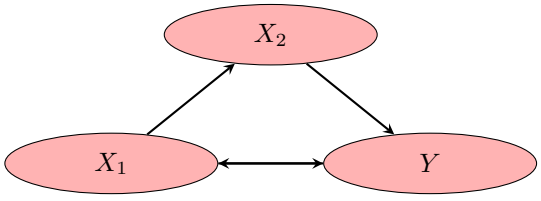


Figure 9 : Non-recursive model

2.3 Model with mediation

Now, we can propose an example of SEM.

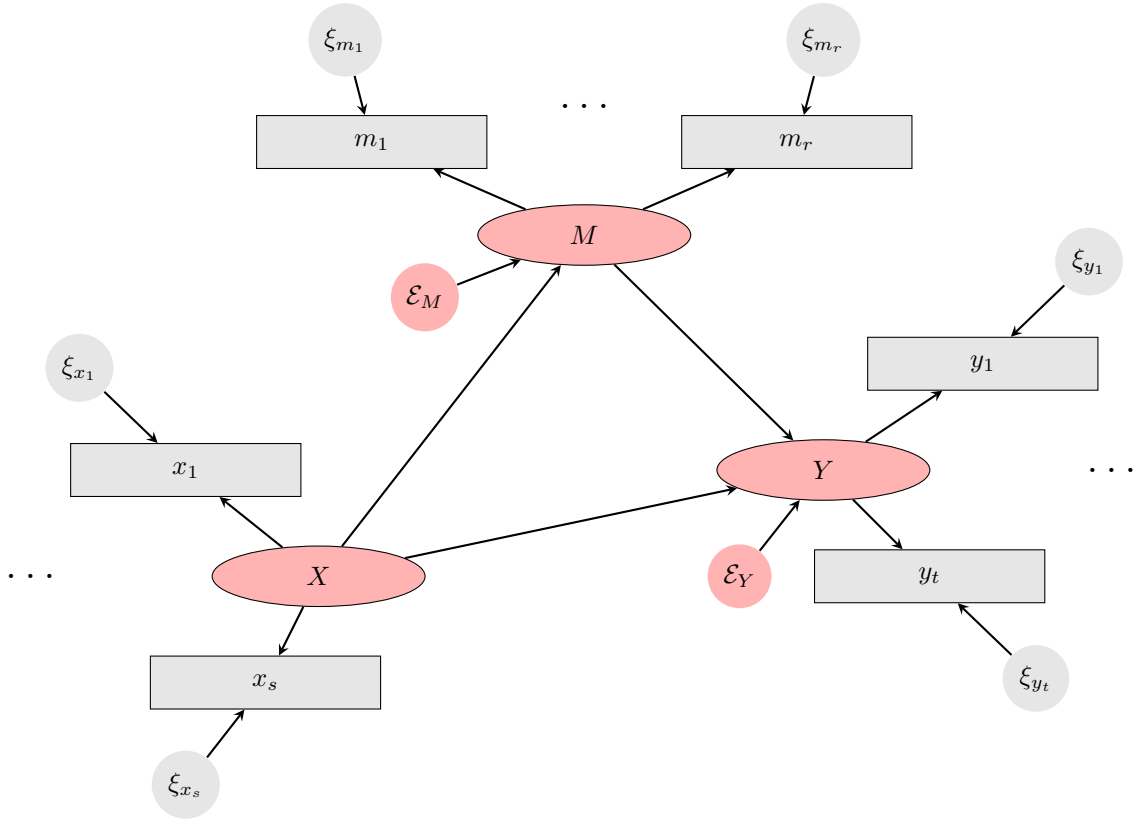


Figure 10 : Structural equation modeling called *Mediation Model*

This model can be interpreted as a study with the view to explain the direct effect of X in Y and the indirect effect of X in Y through M. The last variable is called **mediation variable**. In detail we have :

- Y and M are latent endogenous variables
- X is a latent exogenous variable
- m_i are observed endogenous variables, for $i \in \{1, ..., r\}$
- x_j are observed endogenous variables, for $j \in \{1, ..., s\}$
- y_t are observed endogenous variables, for $k \in \{1, ..., t\}$

We can rewrite the model into a system.

$$\begin{cases} \textbf{Measurement model} \\ m_i = \gamma_i M + \xi_{m_i}, \quad \text{for } i \in \{1, ..., r\} \\ x_j = \alpha_j X + \xi_{x_j}, \quad \text{for } j \in \{1, ..., s\} \\ y_k = \beta_k Y + \xi_{y_k}, \quad \text{for } k \in \{1, ..., t\} \\ \\ \textbf{Structural Model} \\ M = b_0 + b_1 X + \mathcal{E}_M \\ Y = a_0 + a_1 X + a_2 M + \mathcal{E}_Y \end{cases}$$

We can also rewrite the structural model in the form of a matrix equation.

$$\begin{pmatrix} M \\ Y \end{pmatrix} = \begin{pmatrix} b_0 \\ a_0 \end{pmatrix} + \begin{pmatrix} b_1 & 0 \\ a_1 & a_2 \end{pmatrix} \begin{pmatrix} X \\ M \end{pmatrix} + \begin{pmatrix} \mathcal{E}_M \\ \mathcal{E}_Y \end{pmatrix}$$

Now, thanks to the way we wrote the model, we can explicit the different effects. To do that we take the structural model and we use the first line into the second line.

$$\begin{cases} M = b_0 + b_1 X + \mathcal{E}_M \\ Y = a_0 + a_1 X + a_2 M + \mathcal{E}_Y \end{cases}$$

$$\Leftrightarrow Y = a_0 + a_1 X + a_2(b_0 + b_1 X + \varepsilon_M) + \mathcal{E}_Y$$

$$\Leftrightarrow Y = a_0 + a_2 b_0 + a_1 X + a_2 b_1 X + \mathcal{E}'$$

$$\Leftrightarrow Y = (a_0 + a_2 b_0) + (a_1 + a_2 b_1) X + \mathcal{E}'$$

Here, $\mathcal{E}' = a_2 \mathcal{E}_M + \mathcal{E}_Y$.
We can notice that, because $\mathcal{E}_M \in [X]^\perp$ and $\mathcal{E}_Y \in [X, M]^\perp$, $\mathcal{E}' \in [X]^\perp$. We explicit $[X] = Vect\{X\}$ and $[X, M] = Vect\{X, M\}$. Therefore we can write the effects of X on Y.

$$\underbrace{\overbrace{a_2 b_1}^{\text{Indirect effect}} + \overbrace{a_1}^{\text{Direct effect}}}_{\text{Total effect}}$$

3 Estimation method

3.1 Maximum Likelihood Estimation (MLE)

Here S is the covariance matrix of the variables. The principle for the estimation is to minimize the function :

$$F_{ML} \;=\; \log|\Sigma| \;+\; \mathrm{Tr}(S\Sigma^{-1}) \;-\; \log|S| \;-\; (p+q) \tag{1}$$

In detail we have :

- p is the number of endogenous observed variables
- q is the number of exogenous observed variables
- Σ is the covariance matrix of the variables in the model

Obviously, we need to assumed that S and Σ are positive-definite. Otherwise, it would be possible for the undefined log of zero to appear in F_{ML}.

We looking for Σ which is minimizing the equation (1). For that, we are going to use the MLE on Σ to approach S. Indeed, we can see that (1) is minimized for $\Sigma = S$.
Actually, if $\Sigma = S$, we have $\mathrm{Tr}(S\Sigma^{-1}) = p + q$ and so $F_{ML} = 0$.

We can illustrate this result with a simple example which will be detailled on the **Annex**.
We take the following model :

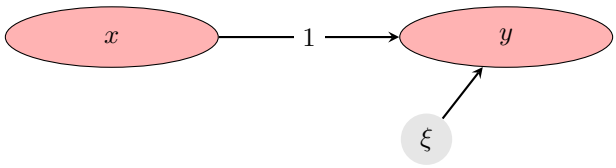


Figure 11 : Recursive model

We have the model without constant :

$$y = x + \xi$$

We pose that $\mathbb{E}(\xi) = 0$ and $\mathrm{Var}(\xi) = \Psi$ with Ψ a diagonal matrix.
We also write that $\mathrm{Var}(x) = \Phi$.

So we obtain

$$S = \begin{pmatrix} \mathrm{Var}(y) & \mathrm{Cov}(y,x) \\ \mathrm{Cov}(y,x) & \mathrm{Var}(x) \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Phi + \Psi & \Phi \\ \Phi & \Phi \end{pmatrix}$$

Then, we can build $\hat{\Sigma}$ which is the covariance matrix Σ where we have done a Maximum Likelihood Estimation (MLE).

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Phi} + \hat{\Psi} & \hat{\Phi} \\ \hat{\Phi} & \hat{\Phi} \end{pmatrix}$$

$$\begin{aligned} F_{ML} &= \log|\hat{\Sigma}| + \mathrm{Tr}(S\hat{\Sigma}^{-1}) - \log|S| - 2 \\ &= \log(\hat{\Phi}\hat{\Psi}) + \left(\frac{\mathrm{Var}(y) - 2\mathrm{Cov}(y,x) + \mathrm{Var}(x)}{\hat{\Psi}} + \frac{\mathrm{Var}(x)}{\hat{\Phi}} \right) - \log\left(\mathrm{Var}(y)\mathrm{Var}(x) - \mathrm{Cov}^2(y,x)\right) - 2 \end{aligned}$$

We find the critical points :

$$\begin{cases} \hat{\Phi} = \mathrm{Var}(x) \\ \hat{\Psi} = \mathrm{Var}(\xi) \end{cases}$$

These are the values which minimize F_{ML}.
Indeed the hessian matrix is positive-definite in these values.

So we have that F_{ML} is minimized for

$$\begin{pmatrix} \mathrm{Var}(y) & \mathrm{Cov}(x,y) \\ \mathrm{Cov}(x,y) & \mathrm{Var}(x) \end{pmatrix} = S$$

3.2 Specification and identification problems

During the building of a SEM, we have some steps to follow.

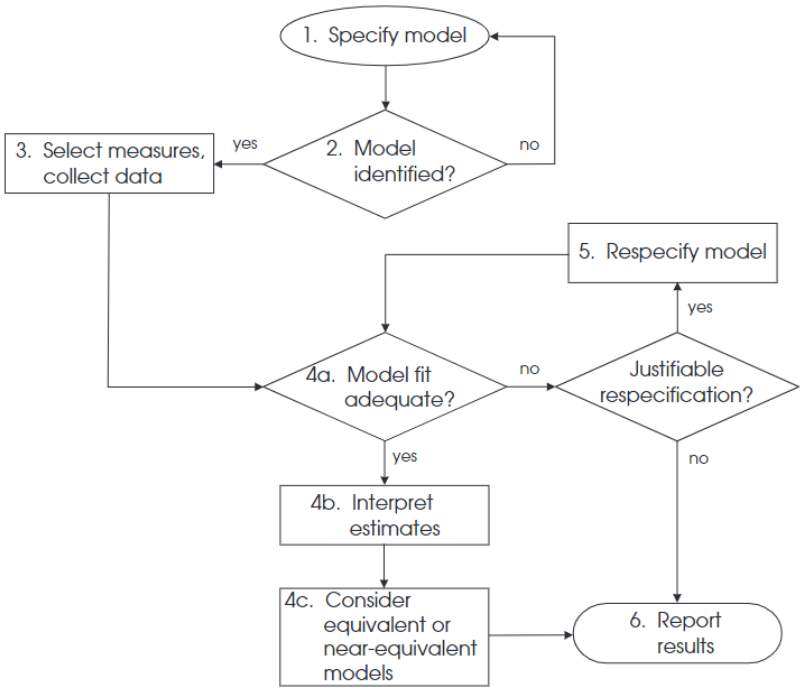


Figure 12 : Flowchart of the basic steps of structural equation modeling. Rex.B.Kline

The first step in the building of SEM is the specification. It is an important step made a researcher in psychology or sociology for example.

Nevertheless, the role of the statistician remains important to advise the researcher who propose the model. He can bring him to suggest variants to the initial model if that one have problems which implies to do another specification.

There is an important step where, in practice, we could encounter problems : the identification of the model. We introduce the notion degree of freedom DF defines as follows :

$$DF = \frac{(p+q)(p+q+1)}{2} - t$$

In detail we have :

- p is the number of endogenous observed variables
- q is the number of exogenous observed variables
- t is the number of parameter to estimate

We say that a model is **identifiable** if $DF \geq 0$.
We can even specify saying that if $DF = 0$, the model is **just-identified** and if $DF > 0$ the model is **over-identified**. On the contrary, if $DF < 0$ the model is **under-identified**.

Thereby, DF permit to the statistician to know which modification is possible to do or not on the model. If that one is over-identified, he will be able to test the model removing observed variables until it become just-identified. Then, these different models may be compared to determine the best one.

However, if the model is under-identified, he will not be able to use it. To overcome this model, it is possible to determine some constraints. Conventionally, the constraint that is fixed is to suppose that a coefficient equals 1.

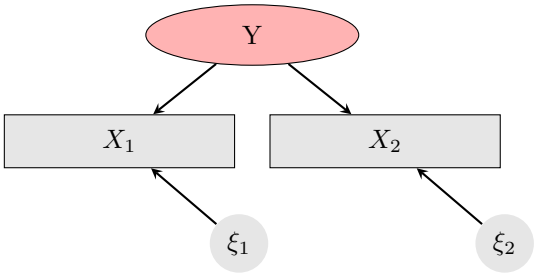


Figure 13 : Under-identified model

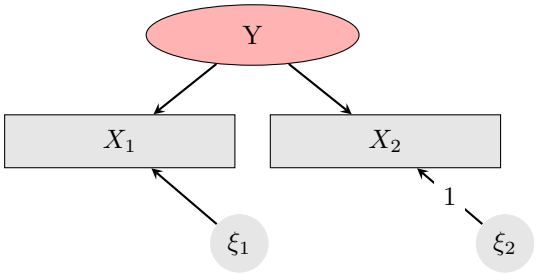


Figure 14 :Just-identified model

For the model under-identified, there are two observed variables and four parameters to estimate (two loading coefficients and two variance errors). Thereby, $DF = \frac{2(2+1)}{2} - 4 = -1 < 0$.
However, by introducing the constraint, we have three parameters to estimate (two loading coefficients and one variance error). So we obtain $DF = \frac{2(2+1)}{2} - 3 = 0$.

If we take back our *Mediation model*, it is possible to verify under which conditions the model would be unidentifiable.

In this model, we have $(r+s+t)$ observed variables, $(r+s+t)$ loading coefficients and variance errors for the observed variables, and two variance errors for the latent variables.

$$DF = \frac{(r+s+t)(r+s+t+1)}{2} - (2(r+s+t)+2)$$

Then we want to have $DF \geq 0$. So

$$\begin{aligned} DF \geq 0 &\Leftrightarrow \frac{(r+s+t)(r+s+t+1)}{2} - (2(r+s+t)+2) \geq 0 \\ &\Leftrightarrow \frac{K(K+1)-4K-4}{2} \geq 0, \quad \text{for } K = (r+s+t) \geq 0 \\ &\Leftrightarrow K^2 - 3K - 4 \geq 0, \quad \text{for } K = (r+s+t) \geq 0 \end{aligned}$$

We find that $K^2 - 3K - 4 \geq 0$ if $K \geq 4$. It means that, for our *Mediation model*, we need to have $r+s+t \geq 4$. Then, our model would be identifiable.

The condition of the $DF \geq 0$ is the most common verification for the identification of our model and the most used in practice.

3.3 Sample size and statistical power

For SEM, it is often said that we need at least 200 observations. However, we can nuance this pseudo-rule. Indeed, Mcquitty in 2004 suggest a table which provides an idea of the number of observations we need according to the degree of freedom of our model and the power that we want for it.

The power of a model is the probability to reject the null hypothesis when the alternative hypothesis is true. We can clearly see that, when the size of our sample decrease, the risk of a false validation increase.

	Minimal size of observations			
Degree of freedom	$\pi = 0.60$	$\pi = 0.70$	$\pi = 0.80$	$\pi = 0.90$
5	885	1132	1463	1994
10	486	613	782	1050
15	350	436	550	732
20	280	346	435	572
30	207	254	314	410
40	168	205	252	325
50	145	175	214	274
75	111	133	168	204
100	92	110	132	165
125	80	95	114	142
150	72	85	101	125
200	61	71	84	104
250	53	62	74	90
300	48	56	66	81
400	41	48	56	68

Tabular 1 : McQuitty 2004

As an example, if we want to have a statistical power of 0.80, which is the standard power expectation for a model with 125 degree of freedom, we need to have at least 114 observations.

3.4 Alternative of MLE

Before to move to the part which concerned the application in psychology, it is important to mention that it exists alternatives to the Maximum Likelihood Estimation (MLE). Indeed, the MLE is very responsive to the absence of normal distribution for the variables. That is why we can use other methods of estimation like the Generalized Least Squares (GLS).
For the GLS, the aim is to minimize :

$$F_{GLS} = \frac{1}{2} tr\left(\left(\left[S - \Sigma\right] W^{-1}\right)^2\right)$$

Here W^{-1} is a weight matrix. One of the most common choice is $W^{-1} = S^{-1}$.

There exists also the Unweighted Least Squares (ULS) and the aim is to minimize :

$$F_{ULS} = \frac{1}{2} tr\left(\left[S - \Sigma\right]^2\right)$$

There exists other methods of estimation that can be used but the main one remains the Maximum Likelihood Estimation.

4 Application in psychology

4.1 Presentation of the study

For the application in psychology, we chose a study realised by Justine Ollivaud, Jean-Michel Galharret and Nicolas Roussiau entitled "**Explicit spirituality, self-esteem and the mechanisms of social and temporal comparison**". The aim is to understand the connections between these notions through structural equation modeling and to conlude if spiritual individuals compare themself less than others and have an higher self-esteem.
The data frame used for the study includes 331 individuals. They are characterized by the 50 following variables :

- *Identifiant* : integer assigned to the individual
- *spirit1*, $\cdots$, *spirit16* : answers of the individual for the 16 questions which relate to spirituality
- *Score_Spirit* : total score of the individual for the questions which relate to spirituality
- *CompSoc1*, $\cdots$, *CompSoc8* : answers of the individual for the 8 questions which relate to social comparison
- *Score_CompS* : total score of the individual for the questions which relate to social comparison
- *CompTemp1*, $\cdots$, *CompTemp8* : answers of the individual for the 8 questions which relate to temporal comparison
- *Score_CompT* : total score of the individual for the questions which relate to temporal comparison
- *es1*, $\cdots$, *es10* : answers of the individual for the 10 questions which relate to self esteem
- *Score_ES* : total score of the individual for the questions which relate to self esteem
- *Sexe* : sexe of the individual
- *Age* : age of the individual
- *Croy_Rel* : belief or not of the individual to religion

To create the latent variables of our structural equation modeling, only will be used the results of the different questions. Their construction will be detailed in a further part.

4.2 Programming tools for SEM with `lavaan` on `R`

The library `lavaan` on `R` is usually used to estimate a large variety of multivariate statistical models, like structural equation modeling. For information, this package also permit to estimate path analysis, confirmatory factor analysis and growth curve models.
To programm a structural equation modeling on `R`, we proceed like this :

```
Model <- '
  # latent variables
    X =~ x1 + x2 + x3
    M =~ m1 + m2 + m3
    Y =~ y1 + y2 + y3
  # regressions
    M ~ X
    Y ~ M + X
  # residual covariances
    x1 ~~ x2   '
```

We can recognize different types of operators :

- =~ : indicate by which observed variables is measured the latent variable
- ~ : indicate by which linear regression is estimate the latent variable
- ~~ : indicate the covariance between residuals

Then, to fit the model we use the following function :

```
fit <- sem(model = Model, data)
```

To make a summary on *fit* permit to obtain some informations like the AIC criterion and the BIC criterion of the model but also an estimation of the differents coefficients.
To have a preview of the model *Model*, we use the package `semPlot` with the following function :

```
semPaths (fit)
```

For our application in psychology, we will use the packages `lavaan` and `semPlot` on `R`.

The package `lavaan` on `R` is not the only one which can be used to do a structural equation modeling. There exists other usefull packages like `sem`, `OpenMx`, `LISREL`, `EQS` and `Mplus`.
Furthermore, there is also possible to do a SEM on `Python` with the package `semopy`. The method used is similar to the one used with `lavaan` and the estimations are very closed.

4.3 Modeling

This part will concern the construction of different modeling by structural equation based on the data frame of our study.

4.3.1 Determination of the latent variables

Before to model the data frame, it is necessary to determine the four latent variables that will be used in the different models : Spirituality, Social comparison, Temporal comparison and Self-esteem. For the variable Spirituality, we preserve all the observed variables of the study. Nevertheless, for Social comparison, Temporal comparison and Self-esteem we only keep the observed variables which have positive coefficients. We can summarize the number of observed variables per latent variables in the following tabular.

Latent variables	Number of observed variables
Spirituality	16
Social comparison	4
Temporal comparison	4
Self-esteem	5

Tabular 2 : Number of observed variables per latent variables

We explicit the measurement model for the latent variables and with all the observed variables concerned under the form of equations.

$$\begin{cases} \text{Spirituality} =\sim spirit1 + spirit2 + spirit3 + spirit4 + spirit5 + spirit6 + spirit7 \\ \qquad\qquad + spirit8 + spirit9 + spirit10 + spirit11 + spirit12 + spirit13 \\ \qquad\qquad + spirit14 + spirit15 + spirit16 \\ \text{Social comparison} =\sim compSoc1 + compSoc3 + compSoc4 + compSoc6 \\ \text{Temporal comparison} =\sim compTemp1 + compTemp3 + compTemp4 + compTemp6 \\ \text{Self-esteem} =\sim es3 + es5 + es8 + es9 + es10 \end{cases}$$

Disposing of all our latent variables, it is possible to build different models of structural equations.

4.3.2 Construction of the model *mod1*

We build the model *mod1* which connects the latent variables Spirituality, Social comparison and Self esteem. Each of them is determined by different observed variables. We obtain the following graphic :

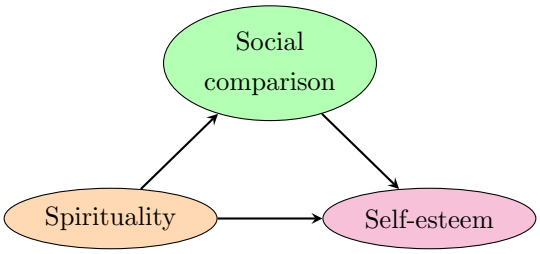


Figure 15 : Structural equation modeling called *mod1*

The aim is to analyse direct and indirect effects on Self-esteem for the model *mod1*. To reference to the first part of our project, we rename our latent variables.

$$\begin{cases} X & := \text{Spirituality} \\ M_1 & := \text{Social comparison} \\ Y & := \text{Self esteem} \end{cases}$$

This is a mediation model. It is possible to analyse the direct and indirect effects on Y. Before that, we explicit the structural model and we substitute in the expression of Y.

$$\begin{cases} M_1 = c_0 + c_1 X + \mathcal{E}_{M_1} \\ Y = a_0 + a_1 X + a_2 M_1 + \mathcal{E}_Y \end{cases}$$

$$\Leftrightarrow Y = (a_0 + a_2 c_0) + (a_1 + a_2 c_1) X + \mathcal{E}'$$

Finally, and with the summary of the fit of the model *mod1* on R, we can build the following tabular.

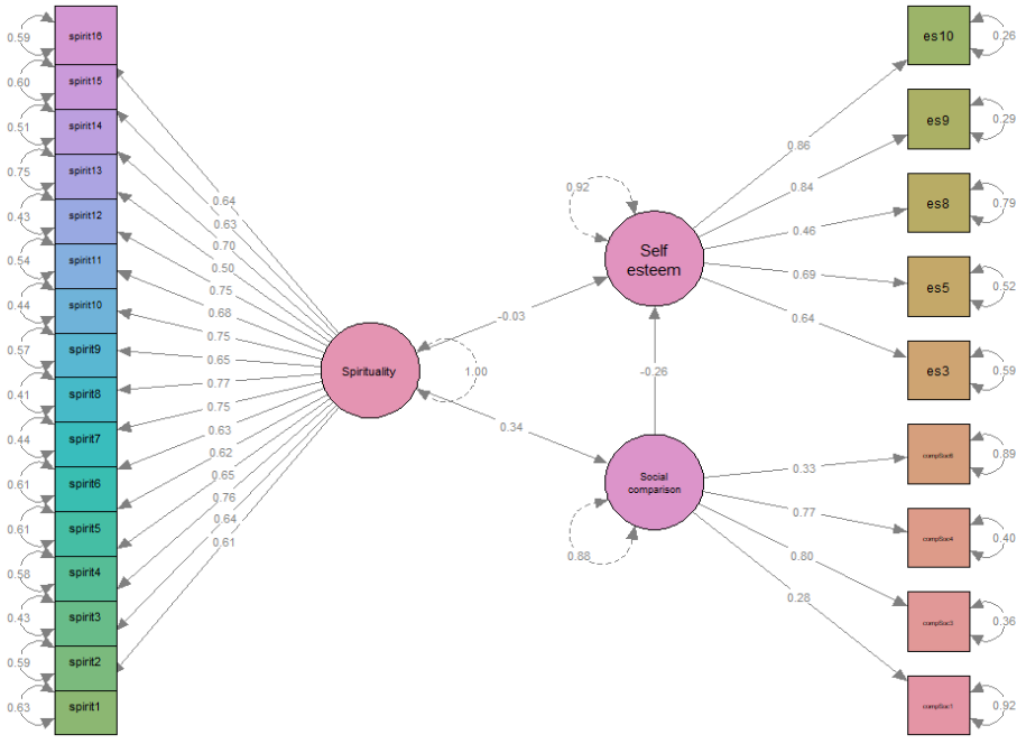
Effect	Expression	Estimate	Std.Err	P-value
Direct	a_1	-0.034	0.068	0.620
Indirect	$a_2 c_1$	-0.094	0.037	0.010
Total	$a_1 + a_2 c_1$	-0.128	0.059	0.032

Tabular 3 : Effects on Y for the model *mod1*

We can conclude that in the model *mod1*, the indirect effect is significant for the variable Self-esteem.

Visualization on `R`

As said before, all the structural equation modeling have been compiled on `R` with the package `lavaan`. For the visualization, we use the package `semPlot`. We have choosen the model *mod1* as an example to insert in the project a result of compilation.



`Figure 16 : Visualization of the model` *`mod1`* `on R`

This is the associated code, with *fit1* the fit associated to *mod1* :

```
semPaths (fit1,
          rotation = 2,
          what = "paths",
          whatLabels = "std",
          layout = 'tree',
          nodeLabels = labels,
          groups = "manlat",
          pastel = TRUE)
```

Identification of the model

The aim is to search if the model is identifiable or not. To do that, we use the notion of degree of freedom and we search DF. We refer to the calculation realised for the *Mediation model* and we substitute with the values. We have 25 observed variables, 25 loading coefficients and variance errors for the observed variables, and two variance errors for the latent variables. We obtain

$$DF = \frac{25 * (25+1)}{2} - (2 \times 25 + 2) = 325 - 52 = 273 > 0$$

The model *mod1* is identified, especially it is over-identified.
The method made here is similar for the models *mod2*, *mod3* and *mod4*.

Linear regression : another way to estimate the effects on the means of scores

For this study and to analyse connections between the notions of Spirituality, Social comparison and Self-esteem, realize a structural equation modeling is not the only solution. It is also possible to do several linear regressions. With the data frame, we have available different variables. Those who interest us there are *Score_Spirit*, *Score_compS* and *Score_ES*. They correspond, respectively in this ordre, to the total score of the individuals for the questions which relate to the latent variables early determined. Disposing of all these information, we are not obliged to do a structural equation modeling, we can also realize two linear regressions, one on the variable *Score_compS* and the other one on the variable *Score_ES*. We have the following equations :

$$\begin{cases} Score\_compS = c'_0 + c'_1\ Score\_Spirit + \mathcal{E}_{Score\_compS} \\ Score\_ES = a'_0 + a'_1\ Score\_Spirit + a'_2\ Score\_compS + \mathcal{E}_{Score\_ES} \end{cases}$$

We obtain similar equations to the model of structural equations.
It is possible to analyze direct and indirect effects on *Score_ES*.

Effect	**Expression**	**Estimate**	**Std.Err**	**P-value**
Direct	a'_1	0.103	0.052	0.047
Indirect	$a'_2 c'_1$	0.141	0.034	0.000
Total	$a'_1 + a'_2 c'_1$	0.244	0.055	0.000

Tabular 4 : Effects on *Score_ES* by linear regression

We can conclude that all the effects are significant. However, the estimations for the coefficients are very different to those obtained with SEM. We can't compare these two methods.

4.3.3 Construction of the model *mod2*

We build the model *mod2* which connects the latent variables Spirituality, Temporal comparison and Self esteem. Each of them is determined by different observed variables. We obtain the following graphic :

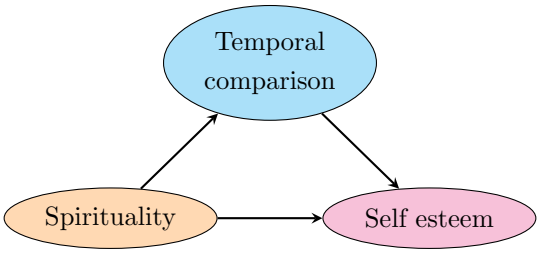


Figure 17 : Structural equation modeling called *mod2*

The aim is to analyse direct and indirect effects on Self-esteem for the model *mod2*. To reference to the first part of our project, we rename our latent variables.

$$\begin{cases} X & := \text{Spirituality} \\ M_2 & := \text{Temporal comparison} \\ Y & := \text{Self esteem} \end{cases}$$

This is a mediation model. It is possible to analyse the direct and indirect effects on Y. Before that, we explicit the structural model and we substitute in the expression of Y.

$$\begin{cases} M_2 = b_0 + b_1 X + \mathcal{E}_{M_2} \\ Y = a_0 + a_1 X + a_2 M_2 + \mathcal{E}_Y \end{cases}$$

$$\Leftrightarrow Y = (a_0 + a_2 b_0) + (a_1 + a_2 b_1) X + \mathcal{E}'$$

Finally, and with the summary of the fit of the model *mod2* on R, we can build the following tabular.

Effect	Expression	Estimate	Std.Err	P-value
Direct	a_1	-0.128	0.058	0.027
Indirect	$a_2 b_1$	0.004	0.012	0.702
Total	$a_1 + a_2 b_1$	-0.124	0.058	0.032

Tabular 5 : Effects on Y for the model *mod2*

We can conclude that in the model *mod2*, the direct effect is significant for the variable Self-esteem.

4.3.4 Construction of the model *mod3*

We build the model *mod3* which connects the latent variables Spirituality, Social comparison, Temporal comparison and Self esteem. Each of them is determined by different observed variables. We obtain the following graphic :

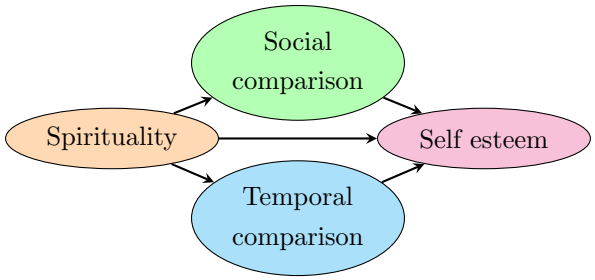


Figure 18 : Structural equation modeling called *mod3*

The aim is to analyse direct and indirect effects on Self-esteem for the model *mod3*. To reference to the first part of our project, we rename our latent variables as we already have done for *mod1* and *mod2* : $(\text{Spirituality}, \text{Social comparison}, \text{Temporal comparison}, \text{Self esteem}) = (X, M_1, M_2, Y)$.

This is a parallel mediation model. It is possible to analyse the direct and indirect effects on Y. Before that, we explicit the structural model and we substitute in the expression of Y.

$$\begin{cases} M_1 = c_0 + c_1 X + \mathcal{E}_{M_1} \\ M_2 = b_0 + b_1 X + \mathcal{E}_{M_2} \\ Y = a_0 + a_1 X + a_2 M_1 + a_3 M_2 + \mathcal{E}_Y \end{cases}$$

$$\Leftrightarrow Y = (a_0 + a_2 c_0 + a_3 b_0) + (a_1 + a_2 c_1 + a_3 b_1) X + \mathcal{E}'$$

Finally, and with the summary of the fit of the model *mod3* on R, we can build the following tabular.

Effect	Expression	Estimate	Std.Err	P-value
Direct	a_1	-0.036	0.060	0.548
Indirect1	$a_2 c_1$	-0.116	0.048	0.016
Indirect2	$a_3 b_1$	0.023	0.019	0.221
Indirect	$a_2 c_1 + a_3 b_1$	-0.093	0.042	0.026
Total	$a_1 + a_2 c_1 + a_3 b_1$	-0.129	0.061	0.034

Tabular 6 : Effects on Y for the model *mod3*

We can conclude that in the model *mod3*, the indirect effect is significant for the variable Self-esteem. Especially, the most significant effect is given by the path :
Spirituality → Social comparison → Self-esteem.

4.3.5 Construction of the model *mod4*

We build the model *mod4* which connects the latent variables Spirituality, Social comparison, Temporal comparison and Self esteem. Each of them is determined by different observed variables. We obtain the following graphic :

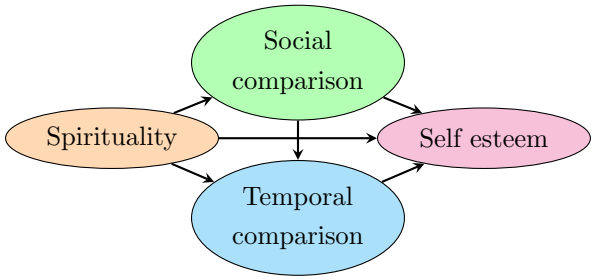


Figure 19 : Structural equation modeling called *mod4*

The aim is to analyse direct and indirect effects on Self-esteem for the model *mod4*. To reference to the first part of our project, we rename our latent variables as we already have done for *mod1* and *mod2* : (Spirituality, Social comparison, Temporal comparison, Self esteem) $= (X, M_1, M_2, Y)$.

This is a serial mediation model. It is possible to analyse the direct and indirect effects on Y. Before that, we explicit the structural model and we substitute in the expression of Y.

$$\begin{cases} M_1 = c_0 + c_1 X + \mathcal{E}_{M_1} \\ M_2 = b_0 + b_1 X + b_2 M_1 + \mathcal{E}_{M_2} \\ Y = a_0 + a_1 X + a_2 M_1 + a_3 M_2 + \mathcal{E}_Y \end{cases}$$

$$\Leftrightarrow Y = (a_0 + a_2 c_0 + a_3 b_0 + a_3 b_2 c_0) + (a_1 + a_2 c_1 + a_3 b_1 + a_3 b_2 c_1) X + \mathcal{E}'$$

Finally, and with the summary of the fit of the model *mod3* on R, we can build the following tabular.

Effect	Expression	Estimate	Std.Err	P-value
Direct	a_1	-0.033	0.062	0.599
Indirect1	$a_2 c_1$	-0.129	0.052	0.013
Indirect2	$a_3 b_1$	-0.002	0.011	0.867
Indirect3	$a_3 b_2 c_1$	0.034	0.021	0.098
Indirect	$a_2 c_1 + a_3 b_1 + a_3 b_2 c_1$	-0.097	0.041	0.018
Total	$a_1 + a_2 c_1 + a_3 b_1 + a_3 b_2 c_1$	-0.130	0.055	0.018

Tabular 7 : Effects on Y for the model *mod4*

We can conclude that in the model *mod4*, the indirect effect is significant for the variable Self-esteem. Especially, the most significant effect is given by the same path than *mod3*.

4.4 Validation

In this part, the objective will be to compare all the models we built to determine which one is the best for the study.

4.4.1 Criteria AIC and BIC

At first, the aim will be to compare models with the same size, that means *mod1* with *mod2* and *mod3* with *mod4*. To do that we use the criteria of AIC and BIC. All the values are given in the summaries of the different *fit*.

Model	AIC	BIC
mod1	18513.694	18715.206
mod2	18355.843	18557.356
mod3	22390.464	22629.997
mod4	22344.201	22587.537

`Tabular 8 : Criteria AIC and BIC for the different models`

In both cases, the best model is the model which minimize the criterion AIC and the criterion BIC. We retain the models *mod2* and *mod4*.

4.4.2 Test of nested models

Now stay available two models which are *mod2* and *mod4*. To decide between them, we realize a test of nested models using the function `anova` on `R`.

The Chi-Squared Difference Test give a p-value of 0,000104.
The test indicates that it is preferable to keep the additional variable of the model *mod4*.

We can conclude that the model *mod4* is the best model for the study.

4.5 Interpretation

The aim of the study was to understand the connections between spirituality, social comparison, temporal comparison and self-esteem for individuals. Especially, we wanted to analyze the effects of spirituality on self-esteem and determine if they pass through social comparison and/or temporal comparison.

Thanks to criteria and tests we used, we were able to define the best model for this study : *mod4*. It gather all the notions and take into account an influence of social comparison on temporal comparison. According to this model which highlights the significance of the indirect effect **Indirect1**, the main effect on self-esteem of individuals is social comparison due to spirituality.

To put it in a nutshell, we can conclude in favor of connection between the different notions. Especially, the effect of spirituality on self-esteem transits by social comparison.

5 Conclusion

To conclude this research paper, we can begin expressing the interest we had for that subject. Indeed, to study in parallel linear regression and structural equation modeling was a plus in the understanding of variable analysis. And it is now possible for us to understand how we can use latent variables.
Both being interested in psychology and social sciences, this research work is a first approach to essential techniques which need to be mastered.

Nevertheless, to begin to learn about linear regression as a beginner while making researches about SEM was a sort of brake for us. At first, not disposing of all necessary notions, it was hard to understand all the subtleties of SEM.

To finish this research paper, we would like to affirm the interest we had learning about structural equation modeling reading books and programming on `R`. We both preserve a good memory of this experience.

6 Bibliography

References

[1] Rex B. Kline, *Principles and Practice of Structural Equation Modeling (Fourth Edition)*, The Guilford Press, 2016.

[2] Kenneth A. Bollen, *Structural Equations with Latent Variables*, John Wiley & Sons, Inc., 1989.

[3] Arielle Bonneville-Roussy, Fabien Fenouillet, Yannick Morvan, *Introduction aux analyses par équations structurelles : applications avec Mplus en psychologie et sciences sociales*, Dunod, 2022.

[4] Patrice Roussel, François Durrieu, Eric Campoy, Assaâd El Akremi, *Méthodes d'équations structurelles : Recherche et Applications en Gestion*, Paris : Economica, 2002.

[5] Eva Delacroix, Alain Jolibert, Élisa Monnot, Philippe Jourdan, *Marketing Research : Méthodes de recherche et d'études en marketing*, Dunod, 2021.

[6] Pierre-Charles Pupion, *Statistiques pour la gestion : Applications avec Excel, SPSS, Amos et SmartPLS*, Dunod, 2012.

[7] Justine Ollivaud, Jean-Michel Galharret, Nicolas Roussiau, *Explicit spirituality, self-esteem and mechanisms of social and temporal comparison*, 2023.

[8] Lavaan package on R [online]. available on : `https://lavaan.ugent.be/`.

[9] Lavaan package on R [online]. available on : `https://stats.oarc.ucla.edu/r/seminars/rsem/`.

[10] Lavaan package on R [online]. available on : `https://towardsdatascience.com/structural-equation-modeling-dca298798f4d`.

[11] Semopy package on Python [online]. available on : `https://semopy.com/`.

7 Annex

We are going to develop the example of the part of the maximum likelihood estimation.

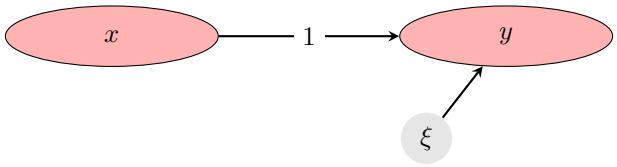


Figure 11 : Recursive model

We have the model without constant :

$$y = x + \xi$$

We pose that $\mathbb{E}(\xi) = 0$ and $\mathrm{Var}(\xi) = \Psi$ with Ψ a diagonal matrix.
We also write that $\mathrm{Var}(x) = \Phi$.
So we obtain

$$S = \begin{pmatrix} \mathrm{Var}(y) & \mathrm{Cov}(y,x) \\ \mathrm{Cov}(y,x) & \mathrm{Var}(x) \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Phi + \Psi & \Phi \\ \Phi & \Phi \end{pmatrix}$$

Then, we can build $\hat{\Sigma}$ which is the covariance matrix Σ where we have done a Maximum Likelihood Estimation (MLE).

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Phi} + \hat{\Psi} & \hat{\Phi} \\ \hat{\Phi} & \hat{\Phi} \end{pmatrix}$$

$$\begin{aligned} F_{ML} &= \log|\hat{\Sigma}| + \mathrm{Tr}(S\hat{\Sigma}^{-1}) - \log|S| - 2 \\ &= \log(\hat{\Phi}\hat{\Psi}) + \mathrm{Tr}\left(\begin{pmatrix} \mathrm{Var}(y) & \mathrm{Cov}(y,x) \\ \mathrm{Cov}(y,x) & \mathrm{Var}(x) \end{pmatrix}\begin{pmatrix} \frac{1}{\hat{\Psi}} & \frac{-1}{\hat{\Psi}} \\ \frac{-1}{\hat{\Psi}} & \frac{1}{\hat{\Phi}} - \frac{1}{\hat{\Psi}} \end{pmatrix}\right) - \log\left(\mathrm{Var}(y)\mathrm{Var}(x) - \mathrm{Cov}^2(y,x)\right) - 2 \\ &= \log(\hat{\Phi}\hat{\Psi}) + \left(\frac{\mathrm{Var}(y) - 2\mathrm{Cov}(y,x) + \mathrm{Var}(x)}{\hat{\Psi}} + \frac{\mathrm{Var}(x)}{\hat{\Phi}}\right) - \log\left(\mathrm{Var}(y)\mathrm{Var}(x) - \mathrm{Cov}^2(y,x)\right) - 2 \end{aligned}$$

Now, we are going to look at the critical points.

$$\begin{cases} \dfrac{\partial F_{ML}}{\partial \hat{\Phi}} = \dfrac{1}{\hat{\Phi}} - \dfrac{\mathrm{Var}(x)}{\hat{\Phi}^2} \\ \dfrac{\partial F_{ML}}{\partial \hat{\Psi}} = \dfrac{1}{\hat{\Psi}} - \dfrac{\mathrm{Var}(y) - 2\mathrm{Cov}(y,x) + \mathrm{Var}(x)}{\hat{\Psi}^2} \end{cases}$$

And,

1. $\dfrac{1}{\hat{\Phi}} - \dfrac{\mathrm{Var}(x)}{\hat{\Phi}^2} = 0 \Leftrightarrow \hat{\Phi} = \mathrm{Var}(x)$
2. $\dfrac{1}{\hat{\Psi}} - \dfrac{\mathrm{Var}(y) - 2\mathrm{Cov}(y,x) + \mathrm{Var}(x)}{\hat{\Psi}^2} \Leftrightarrow \hat{\Psi} = \mathrm{Var}(y) - 2\mathrm{Cov}(y,x) + \mathrm{Var}(x)$

Then, we can verify that the following hessian matrix is positive-definite.

$$\begin{pmatrix} \dfrac{-1}{\hat{\Phi}^2} + \dfrac{2\mathrm{Var}(x)}{\hat{\Phi}^3} & 0 \\ 0 & \dfrac{-1}{\hat{\Psi}^2} + \dfrac{2(\mathrm{Var}(y) - 2\mathrm{Cov}(y,x) + \mathrm{Var}(x))}{\hat{\Psi}^3} \end{pmatrix}$$

This matrix is positive-definite if :

- $\dfrac{-1}{\hat{\Phi}^2} + \dfrac{2\mathrm{Var}(x)}{\hat{\Phi}^3} \geq 0 \Leftrightarrow 2\mathrm{Var}(x) \geq \hat{\Phi}$.
  However $\hat{\Phi}$ is the MLE of $\mathrm{Var}(x)$ so we can agree that the inequality is true.

- $\dfrac{-1}{\hat{\psi}^2} + \dfrac{2(\mathrm{Var}(y) - 2\mathrm{Cov}(y,x) + \mathrm{Var}(x))}{\hat{\Psi}^3} \geq 0 \Leftrightarrow 2(\mathrm{Var}(y) - 2\mathrm{Cov}(y,x) + \mathrm{Var}(x)) \geq \hat{\Psi}$
  $\Leftrightarrow 2\mathrm{Var}(y-x) \geq \hat{\Psi} \Leftrightarrow 2\mathrm{Var}(\xi) \geq \hat{\Psi}$.
  However $\hat{\Psi}$ is the MLE of $\mathrm{Var}(\xi)$ so we can agree that the inequality is true.

Then, we see that $\mathrm{Var}(y) - 2\mathrm{Cov}(y,x) + \mathrm{Var}(x) = \mathrm{Var}(y-x) = \mathrm{Var}(\xi)$.
Because of the fact that ξ and x are uncorrelated, we have that $\mathrm{Cov}(x,y) = \mathrm{Cov}(x, x+\xi) = \mathrm{Var}(x)$ and $\mathrm{Var}(x) + \mathrm{Var}(\xi) = \mathrm{Var}(x+\xi) = \mathrm{Var}(y)$.
So, we have that F_{ML} is minimized for

$$\begin{pmatrix} \mathrm{Var}(x) + \mathrm{Var}(\xi) & \mathrm{Var}(x) \\ \mathrm{Var}(x) & \mathrm{Var}(x) \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(y) & \mathrm{Cov}(x,y) \\ \mathrm{Cov}(x,y) & \mathrm{Var}(x) \end{pmatrix} = S$$