Latent variable model and identifiability

Brice Ozenne

November 2, 2017

1 A necessary condition for identifiability

One way to assess identifiability of a model is to count the number of parameters vs. the number of sufficient statistics brought by the data. If the number of parameters in the model exceed the number of sufficient statistics brought by the data the model is no more identifiable.

1.1 Example in univariate linear models

Assuming normality of my variable, the mean and variance are sufficient statistics (i.e. knowing the mean and the variance, I can simulate new data following the same law as my observed data). Then I need at least two observations to fit a linear model:

```
set.seed(10)
data <- data.frame(Y = rnorm(10),
          X = rnorm(10))
summary(lm(Y ∼ 1, data = data[1,, drop = FALSE]))
```

```
Call:
lm(formula = Y ~ 1, data = data[1, , drop = FALSE])

Residuals:
ALL 1 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01875         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
```

```
summary(lm(Y ∼ 1, data = data[1:2,, drop = FALSE]))
```

```
Call:
lm(formula = Y ~ 1, data = data[1:2, , drop = FALSE])
```

```
Residuals:
      1       2 
 0.1015 -0.1015 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08275    0.10150  -0.815    0.565

Residual standard error: 0.1435 on 1 degrees of freedom
```

If I also want to adjust on X, the new sufficient statistic to estimate is the covariance between X and Y so I need at least one more observation:

```
summary(lm(Y ~ X, data = data[1:2,, drop = FALSE]))
```

```

Call:
lm(formula = Y ~ X, data = data[1:2, , drop = FALSE])

Residuals:
ALL 2 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6277         NA      NA       NA
X             0.5867         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,Adjusted R-squared:    NaN 
F-statistic:   NaN on 1 and 0 DF,  p-value: NA
```

```
summary(lm(Y ~ X, data = data[1:3,, drop = FALSE]))
```

```

Call:
lm(formula = Y ~ X, data = data[1:3, , drop = FALSE])

Residuals:
       1        2        3 
-0.07153  0.09643 -0.02490 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -1.09101    0.09767 -11.171   0.0568 .
X            1.07216    0.12465   8.602   0.0737 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
Residual standard error: 0.1226 on 1 degrees of freedom
Multiple R-squared:  0.9867,Adjusted R-squared:  0.9733
F-statistic: 73.99 on 1 and 1 DF,  p-value: 0.07368
```

2 Application to latent variable models (lvm)

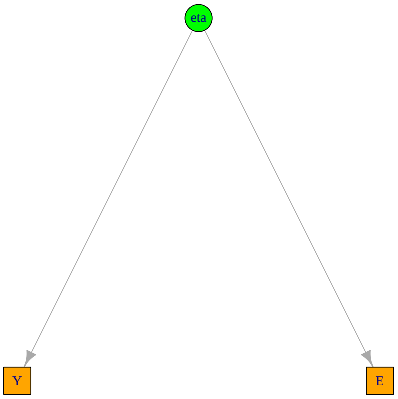
In a latent variable model things are simular. Because we are interested in modeling the relationship between variables, usually we focus on the covariance matrix: does the observed covariance matrix enable to identify the modeled covariance matrix?

2.1 Example: lvm with bivariate outcome

Consider the following model:

```r
library(lava)

lvm.2Y <- lvm(c(Y, E) ~ eta)
lvm.2Y <- latent(lvm.2Y, "eta")
```



This model involves 3 variables (2 observed and 1 latent). We can write the **full variance-covariance matrix** between all the variables:

$$\Sigma_{Y,\eta,E} = \begin{bmatrix} \mathbb{V}ar(Y) & \mathbb{C}ov(Y,\eta) & \mathbb{C}ov(Y,E) \\ & \mathbb{V}ar(\eta) & \mathbb{C}ov(\eta,E) \\ & & \mathbb{V}ar(E) \end{bmatrix}$$

The **empirical variance-covariance matrix** only contains three different parameters:

$$S = \begin{bmatrix} \mathbb{V}ar(Y) & \mathbb{C}ov(Y, E) \\ & \mathbb{V}ar(E) \end{bmatrix}$$

We can check that in R:

```
n <- 1e3
df.2Y <- sim(lvm.2Y, n, latent = FALSE)
cov(df.2Y)
```

```
         Y        E
Y 2.070255 1.112335
E 1.112335 2.138186
```

By default lava parametrize the **modeled variance-covariance matrix**, setting some covariance links to a fixed value. In lava notations, we have:

$$\theta = (Y \sim\sim Y, \eta \sim\sim \eta, E \sim \eta, E \sim\sim E)$$

We therefore need to estimate the 4 parameters from the empirical covariance matrix.

For the mean parameters the full expectation vector would contain 3 parameters, one for each variable. We only observe two of them (Y and E) and by default lava fix the intercept of Y to be 0 so there are only two mean parameters.

In total we have 6 parameters to estimate (2 mean, 4 variance-covariance) but can only estimate 5 empirical moments (2 mean, 3 variance-covariance). The model is therefore not identifiable. This means that the lvm won't properly converge:

```
estimate(lvm.2Y, data = df.2Y)
```

```
                       Estimate  Std. Error     Z-value P-value
Measurements:
   E~eta             1.2791e+00  2.3246e+06  0.0000e+00       1
Intercepts:
   E                 1.4000e-03  2.9136e+04  0.0000e+00       1
   eta              -1.2530e-02  4.5480e-02 -2.7561e-01  0.7828
Residual Variances:
   Y                 1.1994e+00  1.5788e+06  0.0000e+00
   E                 7.1466e-01  2.5832e+06  0.0000e+00
   eta               8.6874e-01  1.5788e+06  0.0000e+00
```

The non identifiability come from the fact that the only equation defining the parameters $E \sim eta$ and $\eta \sim\sim \eta$ is:

$$\mathbb{C}ov\left[Y, E\right] = (E \sim eta) * (\eta \sim\sim \eta)$$

This is clearly not identifiable: $E \sim eta$ and $Y \sim eta$ need to be constrained to be 1.

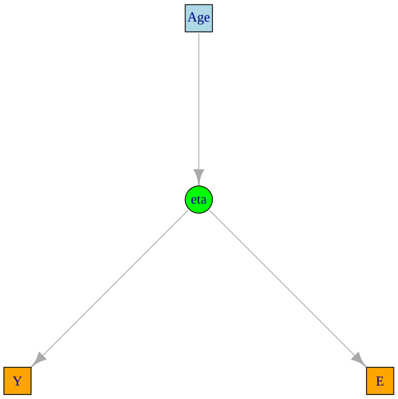
```
lvm.2Y.constrain <- lvm.2Y
regression(lvm.2Y.constrain,E~eta) <- 1
regression(lvm.2Y.constrain,Y~eta) <- 1
estimate(lvm.2Y.constrain, data = df.2Y)
```

```
                   Estimate Std. Error  Z-value P-value
Intercepts:
   E               -0.00210    0.04452 -0.04718  0.9624
   eta             -0.01253    0.04548 -0.27561  0.7828
Residual Variances:
   Y                0.95696    0.07081 13.51394
   E                1.02482    0.07269 14.09907
   eta              1.11122    0.07518 14.78015
```

2.2 Example 2: lvm with bivariate outcome with group effect

Let's modify the previous model by adding an exogenous variable affecting the latent variable:

```
lvm.2Y.Age <- lvm.2Y
regression(lvm.2Y.Age) <- eta~Age
```



The new **modeled variance-covariance matrix** contains 6 parameters:

$$\theta = (Y \sim\sim Y, \eta \sim\sim \eta, E \sim \eta, \eta \sim Age, E \sim\sim E, Age \sim\sim Age)$$

while the new **empirical variance-covariance matrix** contains 6 parameters:

$$S = \begin{bmatrix} \mathbb{V}ar(Y) & \mathbb{C}ov(Y,E) & \mathbb{C}ov(Y,Age) \\ & \mathbb{V}ar(E) & \mathbb{C}ov(E,Age) \\ & & \mathbb{V}ar(Age) \end{bmatrix}$$

So the model satisfy one necessary condition for being identifiable. This condition is however not sufficient to ensure identifiability but is easier to check than the NSC (nessary and sufficient condition). To check the NSC we need to write down the equations relating the empirical and the theoretical moments:

$$\begin{aligned}
\mathbb{V}ar(Y) &= Y \sim\sim Y + \eta \sim\sim \eta \\
\mathbb{V}ar(E) &= E \sim\sim E + (E \sim \eta) * (\eta \sim\sim \eta) \\
\mathbb{V}ar(Age) &= Age \sim\sim Age \\
\mathbb{C}ov(Y,E) &= (E \sim \eta) * (\eta \sim\sim \eta) \\
\mathbb{C}ov(Y,Age) &= (\eta \sim Age) * (Age \sim\sim Age) \\
\mathbb{C}ov(E,Age) &= (E \sim \eta) * (\eta \sim Age) * (Age \sim\sim Age)
\end{aligned}$$

We can re-writte that in matricial form:

$$\begin{bmatrix} \mathbb{V}ar(Y) \\ \mathbb{V}ar(E) \\ \mathbb{V}ar(Age) \\ \mathbb{C}ov(Y,E) \\ \mathbb{C}ov(Y,Age) \\ \mathbb{C}ov(E,Age) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} Y \sim\sim Y \\ E \sim\sim E \\ \eta \sim\sim \eta \\ Age \sim\sim Age \\ E \sim \eta \\ \eta \sim Age \end{bmatrix}$$

$$b = X\theta$$

This problem is identifiable if and only if X has non-zero eigenvalues:

```
X <- rbind(c(1,0,1,0,0,0),
c(0,1,1,0,1,0),
c(0,0,0,1,0,0),
c(0,0,1,0,1,0),
c(0,0,0,1,0,1),
c(0,0,0,1,1,1))
svd(X)$d
```

```
[1] 2.4825297 2.0685860 1.2223636 0.7442566 0.6272682 0.3412366
```

Therefore the model is identifiable.
This is confirmed by the fact that lava is able to estimate the model:

```
df.2Y.Age <- sim(lvm.2Y.Age, n = n)
estimate(lvm.2Y.Age, data = df.2Y.Age)
```

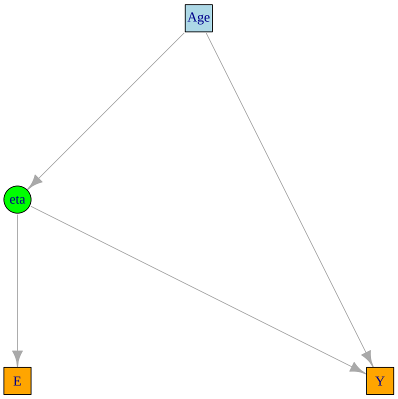
```
                    Estimate Std. Error  Z-value  P-value
Measurements:
   E~eta             0.98049    0.04851 20.21039   <1e-12
Regressions:
   eta~Age           0.99614    0.04703 21.18155   <1e-12
Intercepts:
   E                 0.00459    0.04627  0.09910   0.9211
   eta               0.01963    0.04505  0.43568   0.6631
Residual Variances:
   Y                 1.00969    0.08979 11.24480
   E                 1.16935    0.09111 12.83393
   eta               1.01691    0.08995 11.30476
```

Compared to the previous example two equations now define the parameters $E \sim eta$ and $\eta \sim\sim \eta$:

$$\begin{aligned} \mathbb{C}ov\left[Y,E\right] &= (E \sim eta) * (\eta \sim\sim \eta) \\ \mathbb{C}ov\left[E,Age\right] &= (E \sim eta) * (\eta \sim Age) * (Age \sim\sim Age) \end{aligned}$$

Note that the model is exactly identifiable in the sense that we have exactly the same number of parameters and moments. Adding an additional link between age and one outcome would make the model non identifiable since we would increase by one the number of parameters (p=7) while still having only 6 moments.

```
lvm.2Y.Age2 <- lvm.2Y.Age
regression(lvm.2Y.Age2) <- Y~Age
```



Indeed lava is not able to estimate the model:

```
df.2Y.Age2 <- sim(lvm.2Y.Age2, n = n)
estimate(lvm.2Y.Age2, data = df.2Y.Age2)
```

```
                       Estimate  Std. Error     Z-value P-value
Measurements:
   E~eta             1.4091e+00  2.1518e+06  0.0000e+00       1
Regressions:
   Y~Age             1.2775e+00  1.1396e+06  0.0000e+00       1
    eta~Age          7.4624e-01  1.1396e+06  0.0000e+00       1
Intercepts:
   E                -5.1720e-02  1.1079e+05  0.0000e+00       1
   eta               5.1490e-02  4.4290e-02  1.1624e+00  0.2451
Residual Variances:
   Y                 1.2724e+00  1.0436e+06  0.0000e+00
   E                 6.4440e-01  2.0722e+06  0.0000e+00
   eta               6.8340e-01  1.0436e+06  0.0000e+00
```