4. Causal structure in data

Suppose that (x_1, x_2, … , x_J) are random variables with a covariance matrix **Γ**. Suppose further that they depend on the variables (t_1, t_2, …) in a following way,

$$\begin{aligned} x_1 &= p_{11}\, t_1 + p_{12}\, t_2 + \ldots + p_{1A}\, t_A \\ x_2 &= p_{21}\, t_1 + p_{22}\, t_2 + \ldots + p_{2A}\, t_A \\ &\ldots \\ x_J &= p_{J1}\, t_1 + p_{J2}\, t_2 + \ldots + p_{JA}\, t_A \end{aligned} \tag{6}$$

In a vector/matrix notation (6) can be written as **x**=**Pt**. (When working with data it corresponds to the rows of $\mathbf{X}=\mathbf{TP}^T$). If the t-variables are mutually uncorrelated, the equations (6) imply that we can write $\mathbf{\Gamma}=\mathbf{PDP}^T$, where **D** is a diagonal matrix containing the variances of the t's. For the computed loadings we use the same notation. This situation corresponds to that the data are centred and loadings are generated such that that the matrix of score values has orthogonal columns. This amounts to stating that the score vectors ($\mathbf{t}_i$) in (1) are mutually orthogonal, $\mathbf{t}_i^T\mathbf{t}_j=0$ for $i\neq j$. (Note that this **D** may be different to the **D** in (1). **D** in (1) is a scaling matrix that depends on the procedure and the actual choice of scaling). The coefficients in **P** (J×A) tell us how the variables are related. Following table shows three examples.

	Ex. 1				**Ex. 2**				**Ex. 3**	
x	t_1	t_2	t_3		t_1	t_2	t_3		t_1	t_2
x_1	x	x			x	x			x	
x_2	x	x			x	x			x	
x_3	x	x				x	x		x	
x_4	x		x			x	x			x
x_5	x		x			x	x			x
x_6	x		x		x		x			x

Table 1. Schematic illustration of sizes of loading coefficients, **P**.

Cells in the table marked by *x* indicate that the corresponding loading coefficient is different from zero. The unfilled part thus consists of zeros. In the first example there are only non-zero loadings for the first latent variable. The variables x_1-x_3 have non-zero loadings on the second latent variable, while x_4-x_6 on the third. We see this type, when we work with psychometric data, see the discussion on the Spearman thesis in ref 10. The first latent variable explains all the variables, while the variables can be grouped in relation to the other latent variables. In the second example each of the ten variables can be explained by two latent variables. In the third examples the first latent variable explains x_1-x_3, while the second explains x_4-x_6. It indicates that the two sets of variables are mutually uncorrelated.

The non-zero coefficients can be generated as the ones that are not statistically significant. We can also prescribe certain coefficients to be non-zero and estimate the loading in relation to this pattern of non-zero loading values. In this case we study how well the loading pattern describes the present variables.

The latent variables are derived as linear combinations of the original variables as shown in the following set of equations,

$$
(7) \qquad \begin{aligned}
t_1 &= r_{11}\, x_1 + r_{21}\, x_2 + \ldots + r_{J1}\, x_J \\
t_2 &= r_{12}\, x_1 + r_{22}\, x_2 + \ldots + r_{J2}\, x_J \\
&\ldots \\
t_A &= r_{1A}\, x_1 + r_{2A}\, x_2 + \ldots + r_{JA}\, x_J
\end{aligned}
$$

If the t's are uncorrelated, the equations (7) imply that $\mathbf{D}=\mathbf{R}^T\mathbf{\Sigma}\mathbf{R}$, where **D** is the diagonal matrix of the variances of the t's. If **P** has full rank, **R** is defined as $\mathbf{R}=\mathbf{P}^{T-1}$. In the numerical computations there are different ways to define **R**. The equations (7) are useful, when you want to study the independence and conditional independence among the x-variables.

Sometimes we require a special structure in the coefficient matrices **P** or **R**. An example is when we require **P** to be lower triangular. In that case the equations (6) can be written as

$$
(8) \qquad \begin{aligned}
x_1 &= p_{11}\, t_1 \\
x_2 &= p_{21}\, t_1 + p_{22}\, t_2 \\
x_3 &= p_{31}\, t_1 + p_{32}\, t_2 + p_{33}\, t_3 \\
&\ldots \\
x_J &= p_{J1}\, t_1 + p_{J2}\, t_2 + \ldots
\end{aligned}
$$

This makes interpretation of the loadings simpler. E.g., x_1 is the same as t_1, x_2 depends on t_1 and t_2, and so on. Also, e.g., p_{33} can be interpreted as the part of x_3 that cannot be described by x_1 and x_2. When **P** is a lower triangular matrix, **R** will also be.

The presentation in this section has been based on one block of variables (x_1,x_2, …) or one block of data **X** that is being described by latent variables (t_1,t_2, …) or one block of score data, **T**. But we will be working with several blocks of **X**-data. The same considerations (e.g. patterns of non-zeros) or special structure (e.g. lower triangular loading matrix) apply to each block.

5 Measures of fit and changes

We shall here discuss the concept of J-divergence. It is a useful concept, when we analyse a loading matrix that has been computed in path analysis. The importance of J-divergence is due that it can be applied in the analysis of the decompositions of data matrices that are of reduced rank. Thus it applies in the analysis of paths that has been carried out according to the H-principle. The theory and applications of J-divergence is analogous to the use of the maximum likelihood, ML, as an estimation method. Therefore, we shall briefly review the ML approach, and point out the similarity of J-divergence (see below) to the ML approach.

We shall consider the case that the random variable follows a multivariate normal distribution, X~N($\boldsymbol{\mu}$,$\mathbf{\Sigma}$), where we assume that $\boldsymbol{\mu}$ is zero. In the ML approach the log-likelihood function is often used to measure fit and changes in parameter estimates. It is given by

$$
(9) \qquad N(E) = c - I\, (\log\, (|\mathbf{\Sigma}| + \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S})\,)/2,
$$

where c is a constant and **S** the sample covariance matrix, **S**=**X**N**X**/I. From a theoretical point of view it is comfortable to work with the function (9). If $\mathbf{\Sigma}_1$ is the maximum likelihood estimate for $\mathbf{\Sigma}$ under null-hypothesis and $\mathbf{\Sigma}_2$ the estimate under a reduced (nested) hypothesis, the theory tells us that the difference

(10) $$2(N(\boldsymbol{\Sigma}_2) - N(\boldsymbol{\Sigma}_1)) = - I\ (\log(|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1|) + \mathrm{tr}\ ((\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{S}))$$

approximately will follow a Π^2-distribution with degrees of freedom equal to the difference between the number of parameters at the two hypothesis. (10) is often used to test the effect of adding the number of parameters by 1 and comparing (10) to the critical value of Π^2-distribution with one degree of freedom. The disadvantage of using this procedure is that the function (9) is unstable, when data has reduced rank. In the case of reduced rank the expression contains the ratio of two determinants that are both close to zero. It is possible to show that the determinants are related to the size of the score vectors in data. It is easy to show examples, where arbitrary small score vectors become significant, when using (9) or (10), although they have no predictive ability. It is instructive to look closer at the differential of (9),

(11) $$dN = N\ \mathrm{tr}(\ \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{S}\ \boldsymbol{\Sigma}^{-1})d\ \boldsymbol{\Sigma}$$

The equation shows that optimal $\boldsymbol{\Sigma}$ is one, where $\mathbf{S}$ is a generalised inverse of $\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1}\mathbf{S}\ \boldsymbol{\Sigma}^{-1}$.

From equation (11) we see that $\boldsymbol{\Sigma}^{-1}\mathbf{S}$ should be close to the identity matrix $\mathbf{I}$. When we work with reduced rank, we also need that $\boldsymbol{\Sigma}\mathbf{S}^{-1}$ is close to $\mathbf{I}$. This can be obtained by working with J-divergence,

(12) $$J = \mathrm{tr}(\ \mathbf{S}\ \boldsymbol{\Sigma}^{-1} +\ \mathbf{S}^{-1}\boldsymbol{\Sigma} - 2\ \mathbf{I})/2$$

The theory associated with (12) is the same or similar to (9) and (10). It is derived from testing equality of two covariance matrices, see ref 10. It is instructive to look at (12) in terms of the eigen values of $\mathbf{S}\ \boldsymbol{\Sigma}^{-1}$. If ($8_i$) are the eigen values, it is simple to show that

(13) $$J = ½\ \text{э}\ (1 - 8_i)^2/8_i$$

The equation (13) shows that the ratio of variances from $\mathbf{S}$ and $\boldsymbol{\Sigma}^{-1}$ should be both close to one and different from zero, if J in (13) is to be small. The differential of J is given by

(14) $$dJ = ½\ \mathrm{tr}(\ \mathbf{S}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{S}\ \boldsymbol{\Sigma}^{-1})d\ \boldsymbol{\Sigma}$$

If we compare (11) and (14) we see that (14) will be close to (11), when the inverse $\boldsymbol{\Sigma}^{-1}$ is close to $\mathbf{S}^{-1}$. Equations (11) and (14) give an intuitive background for that the theory of J-divergence is similar to the ML approach, although the equations (9) and (12) look different. But the advantage of working with (12) and (14) is that the expressions can be modified such that they also can be used in the case of reduced rank in data, see next section. Furthermore, we work with the expressions in the case of reduced rank in the same way as in case of full rank. Thus, the criteria we use are well motivated, when data are not of full rank.

6 Sequential estimation of loading coefficients

When working with paths we may get many loading matrices. It is sometimes convenient to analyse a loading matrix closer. We often want e.g., revise the values in the loading matrix, when 'almost zero' elements are put to zero. We shall briefly describe closer how we can revise the estimates of the loading matrix. The procedure below has given a loading matrix $\mathbf{P}$=$\mathbf{P}_0$, such that the sample covariance matrix $\mathbf{S}$ is approximately $\mathbf{S}$□$\mathbf{P}_0\mathbf{P}_0$N. Here we suppose that the score vectors have been scaled to unit length. The algorithms associated with the H-principle in the path analysis always provide with a matrix $\mathbf{R}=\mathbf{R}_0$ such that $\mathbf{R}_0\mathrm{N}\mathbf{P}_0= \mathbf{I}$, where $\mathbf{I}$ is the identity matrix with number of diagonal elements equal to the number of score vectors. The task we consider here is to find a simple or a significant form of the loading matrix. The matrix **P** that we find may contain only significant non-zeros. It may also contain the best estimates according to some pattern of non-zeros, cf. Table 1. Or it may contain a combination of these two, i.e., significant loading values in a given pattern of non-zero loading values.

We shall use the following notation:

$$\mathbf{S}=\mathbf{P}_0\mathbf{P}_0^T, \qquad \mathbf{S}^+=\mathbf{R}_0\mathbf{R}_0^T, \quad \mathbf{\Sigma}=\mathbf{P}\mathbf{P}^T, \qquad \mathbf{\Sigma}^+=\mathbf{R}\mathbf{R}^T.$$

The matrix **R** in the procedure below is computed as $\mathbf{R}=\mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}$. At the start the matrix P is a zero matrix. We fill it out by non-zeros until the matrix $\mathbf{\Sigma}$ is sufficiently 'close' to **S** according to the J-divergence. Note that the number of columns in **P** can be smaller than the number of columns in $\mathbf{P}_0$. There are three steps in filling out **P**:

1. Find the next element in **P** that should be considered as non-zero.
2. Estimate the values of all non-zero elements in **P**.
3. Judge the improvement and closeness to $\mathbf{P}_0$.

There are many ways to handle each of the three steps. We shall here only consider one approach of each.

Step 1. We are to find a new loading matrix $\mathbf{P}_1$, $\mathbf{P}_1=\mathbf{P} + c\,\mathbf{E}_{ij}$, where $\mathbf{E}_{ij}$ has 1 as element (i,j) and zero for others. It is natural to choose the next non-zero element that gives the maximal increase in $[\mathrm{tr}(\mathbf{P}_1\mathbf{P}_1^T\mathbf{S}^+)-\mathrm{tr}(\mathbf{P}\mathbf{P}^T\mathbf{S}^+)]$. If we solve the equation $M[\mathrm{tr}(\mathbf{P}_1\mathbf{P}_1^T\mathbf{S}^+)-\mathrm{tr}(\mathbf{P}\mathbf{P}^T\mathbf{S}^+)]/Mc=0$, it is shown in Appendix 1 that a solution is given by

(15) $$c = - \mathbf{p}_j^T\mathbf{u}_i/(u_{ii})$$

Here $\mathbf{p}_j$ is the j-th column of **P** and $\mathbf{u}_i$ the i-th column of $\mathbf{U}=\mathbf{S}^+=(u_{ij})$. It gives an optimal increase as,

(16) $$[\mathrm{tr}(\mathbf{P}_1\mathbf{P}_1^T\mathbf{S}^+)-\mathrm{tr}(\mathbf{P}\mathbf{P}^T\mathbf{S}^+)] = (\mathbf{p}_j^T\mathbf{u}_i)^2/(u_{ii})$$

The task is thus to find the element in **P** that gives the maximal value of

(17) $\max\ (\mathbf{p}_j^T\mathbf{u}_i)^2/(u_{ii})$, for (i,j) within the pattern of actual or required non-zeros.

Step 2. Previous step gave a candidate for the next non-zero element. We need to revise the estimates of present non-zero elements in the light of the new one. If we differentiate **J** with respect to an element in **P**, we get from (14)

$$MJ/Mp_{ij} = ½\ \mathrm{tr}(\ \mathbf{S}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{S}\ \mathbf{\Sigma}^{-1})(\mathbf{e}_i\ \mathbf{p}_j^T + \mathbf{p}_j\ \mathbf{e}_i^T)$$

Here we need an expression for $M\mathbf{\Sigma}/Mp_{ij}$. From $\mathbf{\Sigma} =\mathbf{P}\mathbf{P}^T=\mathbf{p}_1\ \mathbf{p}_1^T + \mathbf{p}_2\ \mathbf{p}_2^T + \ldots$, we derive $M\mathbf{\Sigma}/Mp_{ij} =\mathbf{e}_i\ \mathbf{p}_j^T + \mathbf{p}_j\ \mathbf{e}_i^T$. Here $\mathbf{e}_i=(0,.,0,1,0,.)$, where the 1 is at the i-th element. If we require $MJ/Mp_{ij}=0$, we get

$$\mathrm{tr}(\mathbf{S}^{-1}\, \mathbf{p}_j\, \mathbf{e}_i^T) = \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}\, \mathbf{\Sigma}^{-1}\, \mathbf{p}_j\, \mathbf{e}_i^T) = \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}\, \mathbf{R}\mathbf{R}^T\, \mathbf{p}_j\, \mathbf{e}_i^T) = \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}\, \mathbf{R}\, \mathbf{e}_j\, \mathbf{e}_i^T)$$

This gives

$$\mathbf{u}_i^T\, \mathbf{p}_j = (\mathbf{\Sigma}^{-1}\mathbf{S}\, \mathbf{R})_{ij} \qquad \text{(the (i,j)-th element of } \mathbf{\Sigma}^{-1}\mathbf{S}\, \mathbf{R}\text{).}$$

If we write out the linear equations, we get

$$\text{(18)} \qquad u_{1i}\, p_{1j} + u_{2i}\, p_{2j} + \ldots + u_{Ji}\, p_{Jj} = (\mathbf{\Sigma}^{-1}\mathbf{S}\, \mathbf{R})_{ij}$$

The equation (18) is used to find $\mathbf{p}_j$. Some of the p-values in (18) can be zero and therefore are not in the equations. There is one equation for each non-zero element in $\mathbf{p}_j$. Therefore (18) will give us a quadratic coefficient matrix that is positive definite. The part of the coefficient matrix $\mathbf{U}$=$\mathbf{S}^+$ that is used in (18) will typically be of reduced rank. Therefore, some care must be given in finding the solution $\mathbf{p}_j$. Note that both $\mathbf{R}=\mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}$ and $\mathbf{\Sigma}^{-1}=\mathbf{R}\, \mathbf{R}^T$ depend on $\mathbf{P}$. Thus, some few iterations are necessary to get convergence. The starting values are the previous values of non-zero elements in $\mathbf{P}$ and the value c in (15) for the estimate of the last non-zero one. The task of (18) is to provide with a small adjustment of the values of the non-zero elements.

Step 3. The starting point is the expression for the J-divergence, (14), that we need to truncate to fit data with reduced rank. The products $\mathbf{S}\, \mathbf{\Sigma}^{-1}$ and $\mathbf{S}^{-1}\mathbf{\Sigma}$ are both positive definite. They can therefore be written as $\mathbf{C}\mathbf{C}^T$ and $\mathbf{C}^{T-1}\mathbf{C}^{-1}$, resp. It gives

$$J = \mathrm{tr}(\mathbf{S}\, \mathbf{\Sigma}^{-1} + \mathbf{S}^{-1}\mathbf{\Sigma} - 2\, \mathbf{I})/2 = \mathrm{tr}((\mathbf{C} - \mathbf{C}^{T-1})(\mathbf{C} - \mathbf{C}^{T-1})^T)/2$$

When we truncate, we use $\mathbf{C}$ as $\mathbf{C}\cong\mathbf{R}^T\mathbf{P}_0$ og $\mathbf{C}^{T-1}\cong (\mathbf{R}_0 N\mathbf{P})N$. It gives

$$\text{(19)} \qquad J \cong \mathrm{tr}((\mathbf{R}N\mathbf{P}_0 - (\mathbf{R}_0^T\mathbf{P}^T)\, (\mathbf{R}^T\mathbf{P}_0 - (\mathbf{R}_0^T\mathbf{P})^T)^T)/2 = \text{э}\, f_{ij}^2/2,$$

where $\mathbf{F}=\mathbf{R}^T\mathbf{P}_0 - (\mathbf{R}_0^T\mathbf{P})^T$. Both $\mathbf{R}^T\mathbf{P}_0$ and $(\mathbf{R}_0^T\mathbf{P})^T= \mathbf{P}^T\mathbf{R}_0$ approach the unity matrix of appropriate size, when $\mathbf{P}$ is filled out (and $\mathbf{R}$ computed as $\mathbf{R}=\mathbf{P}(\mathbf{P}^T\mathbf{P})^{-1}$). The filling out of $\mathbf{P}$ stops, when J is sufficiently small or $\mathbf{F}$ can be judged as a random matrix.