EFMD EQUIS ACCREDITED

AACSB ACCREDITED

University of Ljubljana

FACULTY OF ECONOMICS

# Estimation of the covariance matrix based on two types of the forward search algorithm

**Aleš Toman**

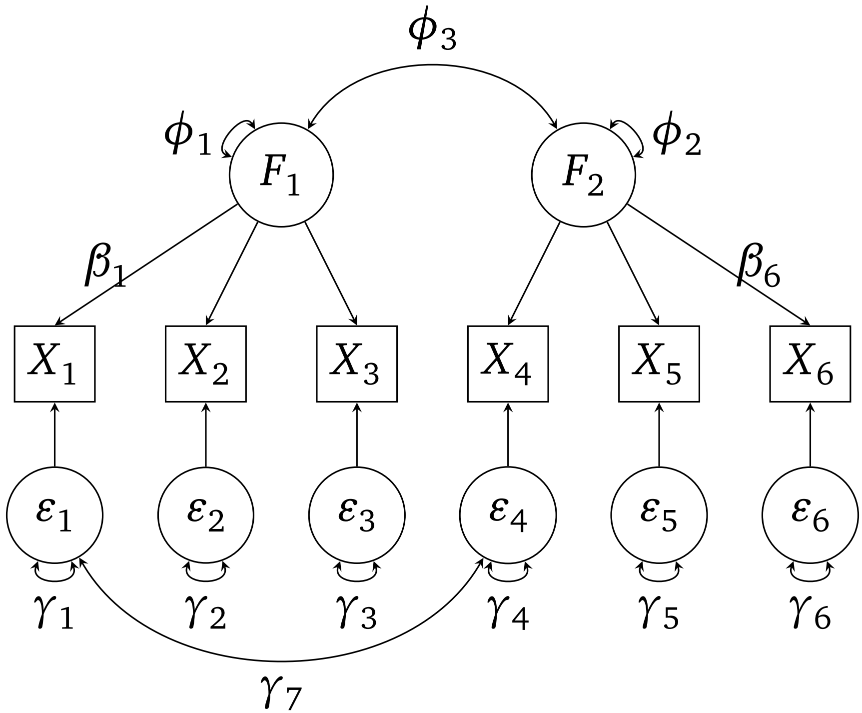
23rd International Workshop on Matrices and Statistics
Ljubljana, June 9, 2014

# Outline

- Confirmatory factor analysis
- Forward search algorithm
- Outlier identification
- Robust covariance matrix estimate
- Robust confirmatory factor analysis
- Conclusions

*The research was partially supported by Slovenian Research Agency and IMFM.*

# Confirmatory factor analysis [2]




Linear dependence: $X = \mu + \Lambda F + \varepsilon$.

Does the restricted model make ***reasonable fit*** to the data?

# Confirmatory factor analysis [2]

- $\mathrm{var}(X) = \Sigma \qquad p \times p$
- $\mathrm{var}(F) = \Phi \qquad q \times q$
- $\mathrm{var}(\varepsilon) = \Psi \qquad p \times p$

Model implied covariance matrix $\boxed{\widetilde{\Sigma} = \Lambda\Phi\Lambda^T + \Psi}$.

***Maximum likelihood estimates***

- Multivariate normal distribution.
- $\boxed{\widehat{\Sigma}}$ maximum likelihood estimate of the covariance matrix.
- Minimize $\boxed{F_{\mathrm{ML}} = \mathrm{trace}(\widetilde{\Sigma}^{-1}\widehat{\Sigma}) - \log\left(\det(\widetilde{\Sigma}^{-1}\widehat{\Sigma})\right) - p}$.

Model estimation and fit evaluation are ***based on the matrix*** $\widehat{\Sigma}$!

# Forward search algorithm [1]

The forward search algorithm is an ***iterative method***, that orders the data according to their ***distances from the proposed model***.

It helps us to identify observations with disproportionately ***high influence*** on statistical inference.

# Forward search algorithm [1]

The forward search algorithm is an ***iterative method***, that orders the data according to their ***distances from the proposed model***.

It helps us to identify observations with disproportionately ***high influence*** on statistical inference.

1. Split the sample into 2 subsets
   - outlier free ***basic set***,
   - non-basic set.
2. Add observations to the basic set.
3. Use ***forward plots*** to show the dynamics of estimates.

The algorithm enables
- data exploration,
- robust parameter estimation.

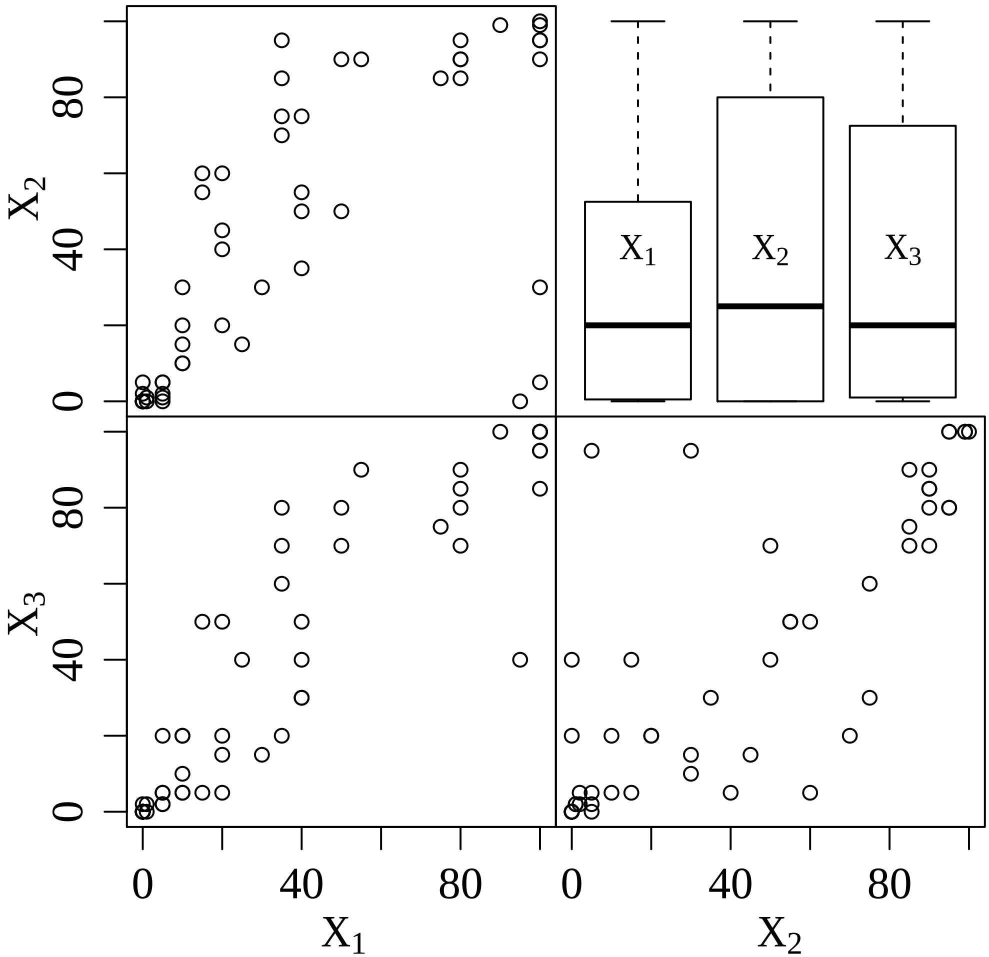
# Data

**G. W. Cermak, K. A. Bollen:** *Observer consistency in judging extent of cloud cover*, Atmospheric Environment, **17** (1983) 2109–2121.

- $n = 60$ slides (July 1980).
- $p = 3$ judges.
- Percent of the sky containing clouds.

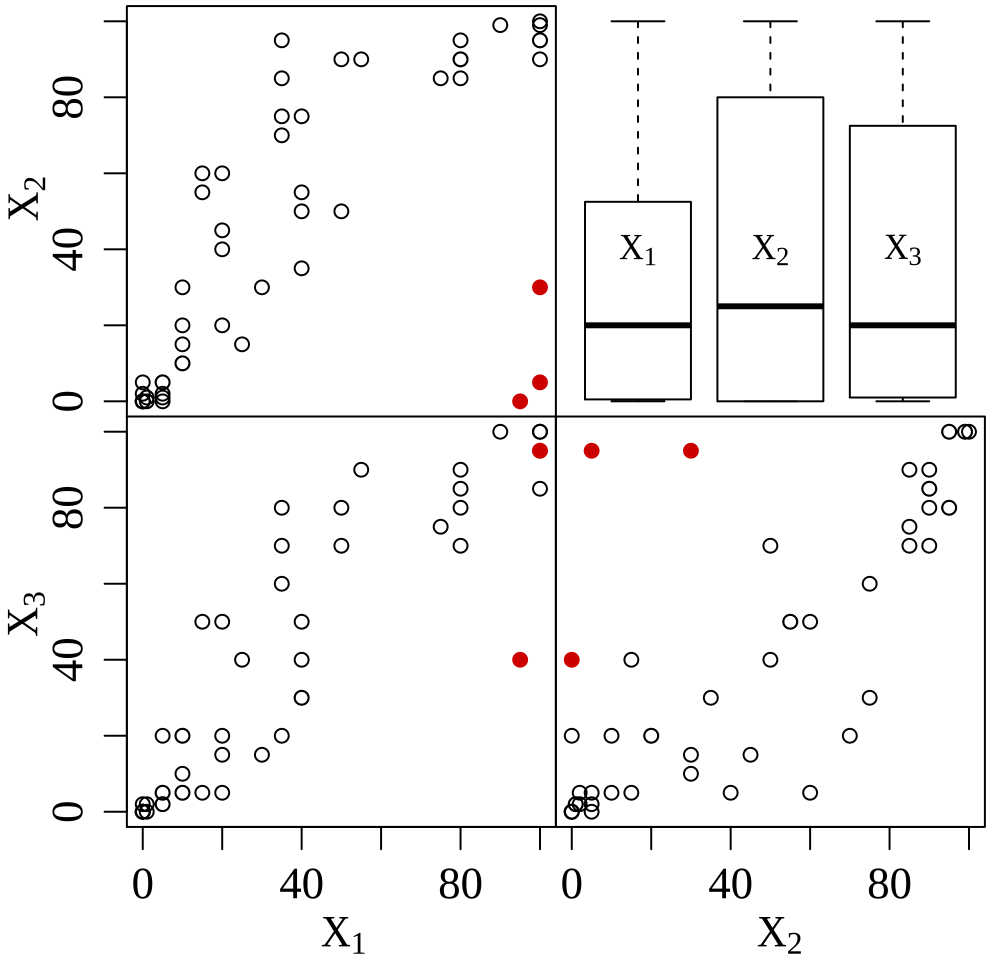
# Data

- Small sample?
- Asymmetric and non-normal distribution.
- Frequently used.



# Data

- Small sample?
- Asymmetric and non-normal distribution.
- Frequently used.
- Observations 40, 51, 52.



# Full sample estimates

$$\bar{x} = \begin{bmatrix} 32.95 \\ 37.65 \\ 35.55 \end{bmatrix}$$

$$S = \begin{bmatrix} 1301 & 1020 & 1237 \\ 1020 & 1463 & 1200 \\ 1237 & 1200 & 1404 \end{bmatrix}$$

# Robust covariance matrix estimate [3]

1. Split the sample into 2 subsets
   - outlier free basic set,
   - non-basic set.

$S$ sample covariance matrix

$S_{(-i)}$ sample covariance matrix ***with observation*** $i$ ***excluded***

$$S = \begin{bmatrix} \triangle_1 & \cdot & \cdot \\ \square_1 & \square_2 & \cdot \\ \pentagon_1 & \pentagon_2 & \pentagon_3 \end{bmatrix} \quad \Rightarrow \quad \operatorname{vecs}(S) = \begin{bmatrix} \triangle_1 \\ \square_1 \\ \square_2 \\ \pentagon_1 \\ \pentagon_2 \\ \pentagon_3 \end{bmatrix}$$

- $s = \operatorname{vecs}(S)$
- $s_{(-i)} = \operatorname{vecs}(S_{(-i)})$

# Robust covariance matrix estimate [3]

1. Split the sample into 2 subsets
   - outlier free basic set,
   - non-basic set.

(Squared) ***Cook's distance*** of observation $i$

$$\mathrm{CD}_i^2 = (s_{(-i)} - s)^T (\widehat{\mathrm{cov}}\, s)^{-1} (s_{(-i)} - s)$$

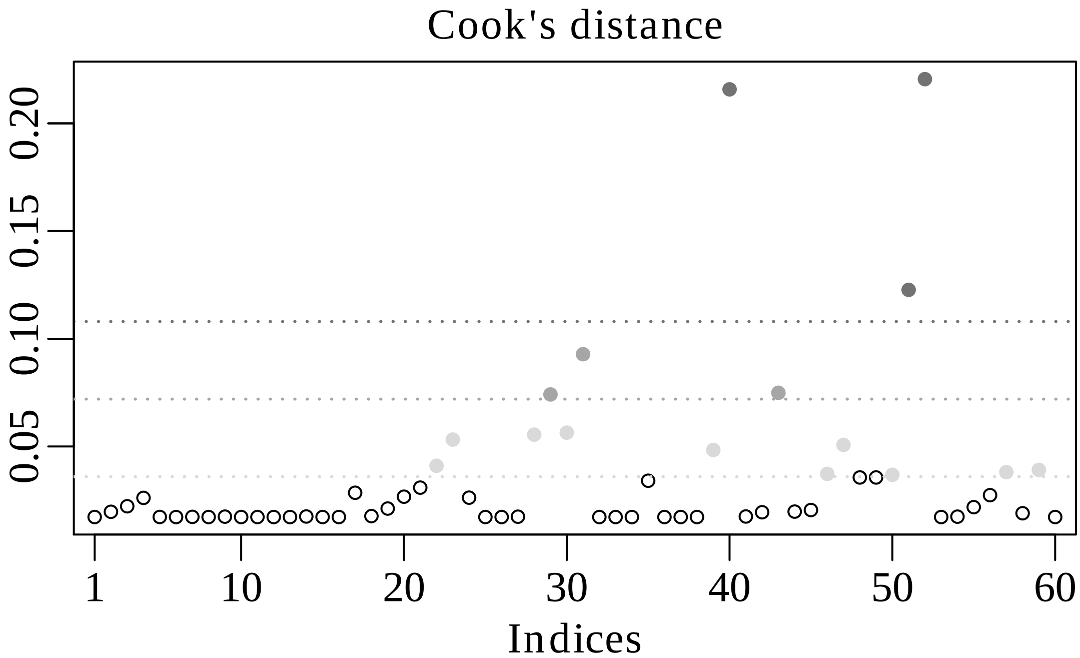
Multivariate normal distribution of $X$:

- $\mathrm{cov}(s_{gh}, s_{jk}) \propto \sigma_{gj}\sigma_{hk} + \sigma_{gk}\sigma_{hj}$
- $\widehat{\mathrm{cov}}(s_{gh}, s_{jk}) \propto s_{gj}s_{hk} + s_{gk}s_{hj}$

Take $m = 30$ observations with lowest Cook's distances.

# Robust covariance matrix estimate [3]

1. Split the sample into 2 subsets
   - ► outlier free basic set,
   - ► non-basic set.

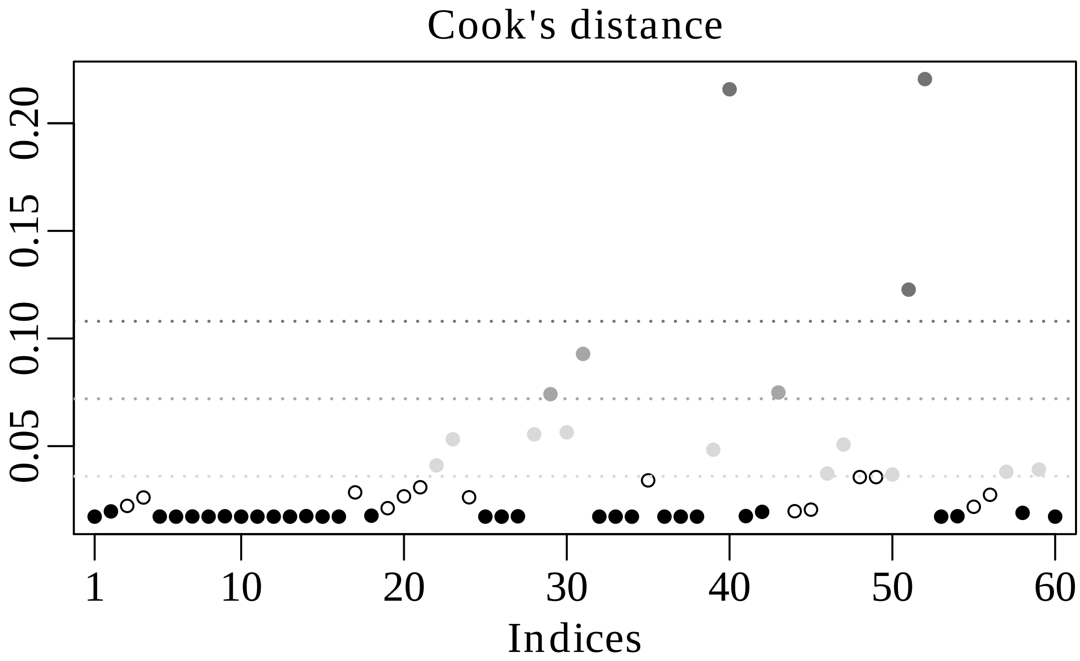



$$\mathrm{CD} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{CD}_i \text{ and levels } 1\cdot\mathrm{CD},\ 2\cdot\mathrm{CD} \text{ in } 3\cdot\mathrm{CD}$$

# Robust covariance matrix estimate [3]



1. Split the sample into 2 subsets ► outlier free basic set,
   ► non-basic set.




$CD = \frac{1}{n}\sum_{i=1}^{n} CD_i$ and levels $1 \cdot CD$, $2 \cdot CD$ in $3 \cdot CD$

# Robust covariance matrix estimate [3]



**2.** Add observations to the basic set.

$X^{(\ell)}$ the basic set with $\ell$ observations: we wish to include one more.

$S^{(\ell)}$ sample covariance matrix of the basic set.

- $S^{(\ell)}_{(-i)}$ covariance matrix ***with observation*** $i$ ***excluded***,
- $S^{(\ell)}_{(+i)}$ covariance matrix ***with observation*** $i$ ***added***.

$$s^{(\ell)} = \operatorname{vecs}\left(S^{(\ell)}\right) \qquad s^{(\ell)}_{(-i)} = \operatorname{vecs}\left(S^{(\ell)}_{(-i)}\right) \qquad s^{(\ell)}_{(+i)} = \operatorname{vecs}\left(S^{(\ell)}_{(+i)}\right)$$

# Robust covariance matrix estimate [3]

**2.** Add observations to the basic set.

(Squared) ***Cook's distance*** of observation $i$

$$\mathrm{CD}_i^{2(\ell)} = \begin{cases} \left(s_{(-i)}^{(\ell)} - s^{(\ell)}\right)^T \left(\widehat{\mathrm{cov}}\, s^{(\ell)}\right)^{-1} \left(s_{(-i)}^{(\ell)} - s^{(\ell)}\right); & i \in X^{(\ell)} \\ \left(s_{(+i)}^{(\ell)} - s^{(\ell)}\right)^T \left(\widehat{\mathrm{cov}}\, s^{(\ell)}\right)^{-1} \left(s_{(+i)}^{(\ell)} - s^{(\ell)}\right); & i \notin X^{(\ell)} \end{cases}$$

- $\widehat{\mathrm{cov}}\left(s_{gh}^{(\ell)}, s_{jk}^{(\ell)}\right) \propto s_{gj}^{(\ell)} s_{hk}^{(\ell)} + s_{gk}^{(\ell)} s_{hj}^{(\ell)}$

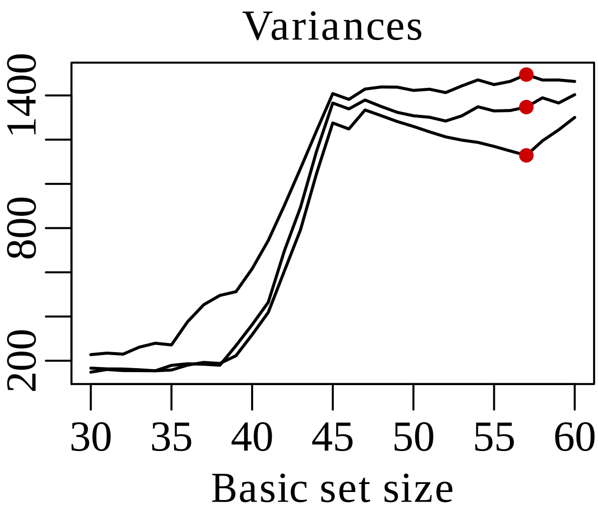
$\mathrm{CD}_i^{(\ell)}$ measures the ***the influence of observation*** $i$ ***on*** $S^\ell$.

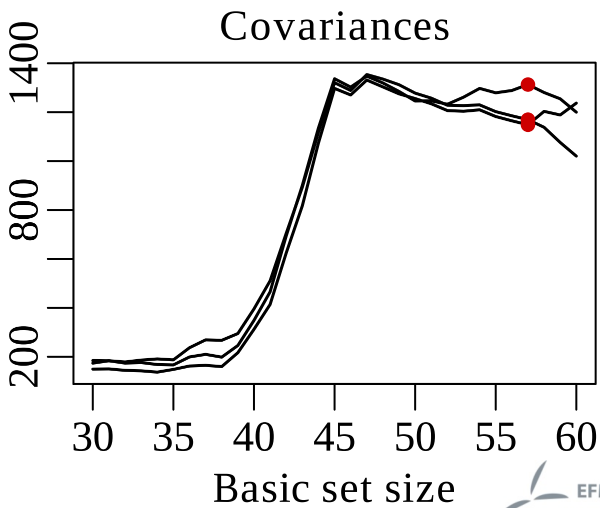
Take $\ell + 1$ observations with lowest Cook's distances.

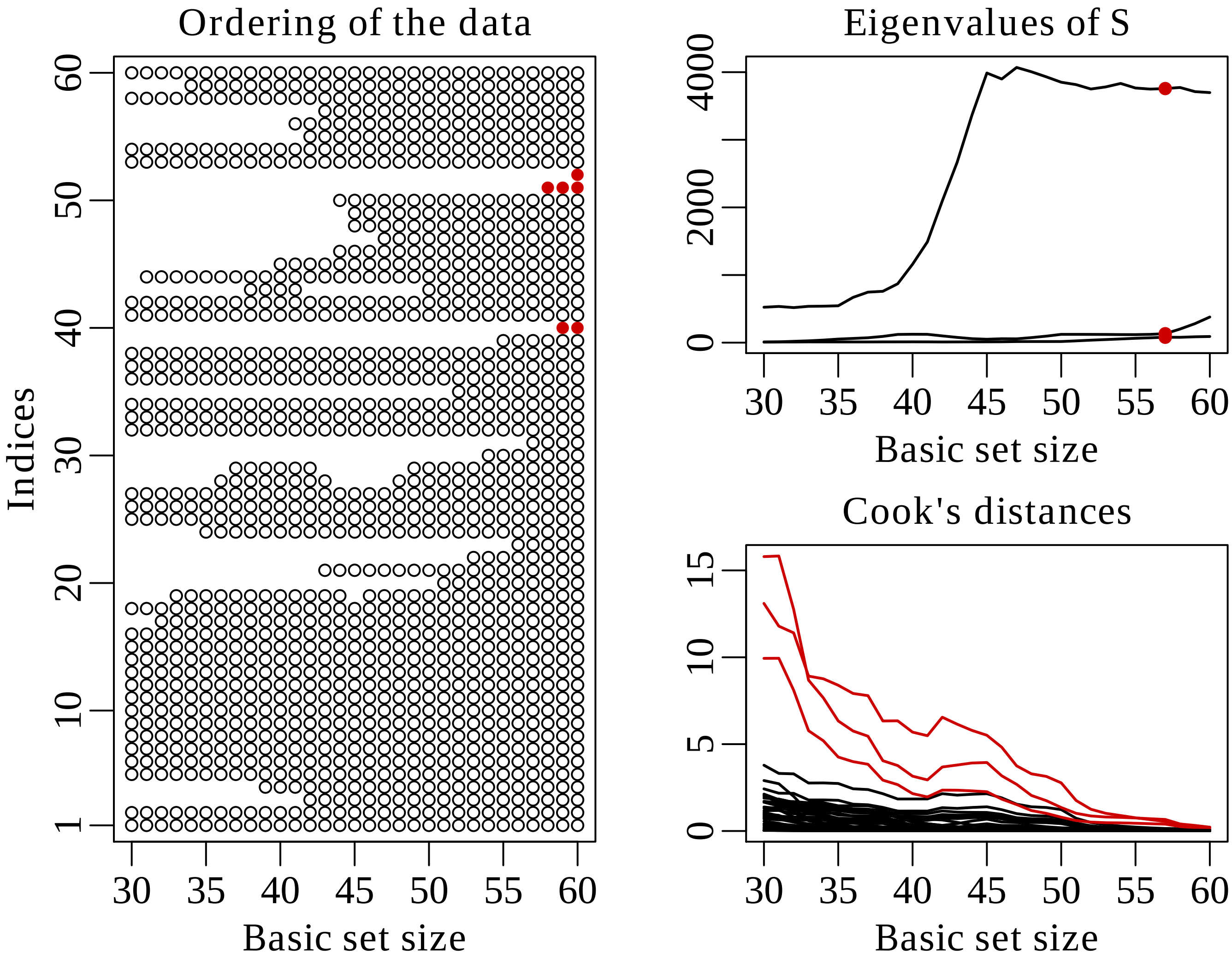
# Robust covariance matrix estimate [3]

**3.** Use forward plots to show the dynamics of estimates.

- Ordering of the data.
- Variances, covariances, and their functions.
- Cook's distances.





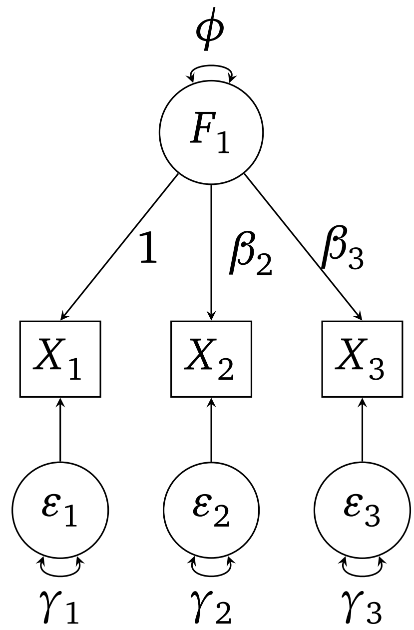
Ordering of the data
Indices
Basic set size
Eigenvalues of S
Basic set size
Cook's distances
Basic set size

# Data and the confirmatory factor model

$$\bar{x} = \begin{bmatrix} 32.95 \\ 37.65 \\ 35.55 \end{bmatrix}$$

$$\widehat{\Sigma} = \begin{bmatrix} 1279 & 1003 & 1216 \\ 1003 & 1439 & 1180 \\ 1216 & 1180 & 1380 \end{bmatrix}$$

- One-factor model.
- Independent errors.

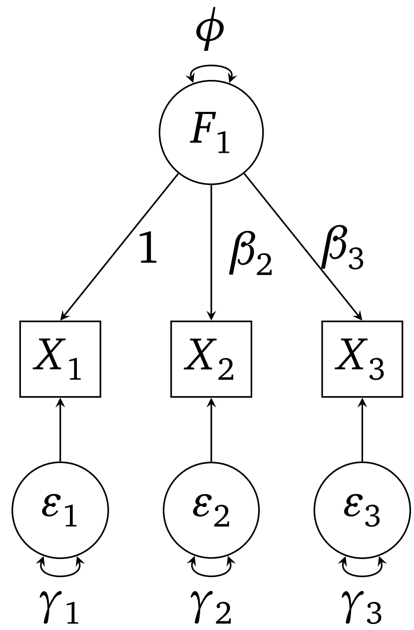


# Data and the confirmatory factor model

$$\bar{x} = \begin{bmatrix} 32.95 \\ 37.65 \\ 35.55 \end{bmatrix}$$

$$\widehat{\Sigma} = \begin{bmatrix} 1279 & 1003 & 1216 \\ 1003 & 1439 & 1180 \\ 1216 & 1180 & 1380 \end{bmatrix}$$

- One-factor model.
- Independent errors.
- ***Heywood case:*** $\widehat{\gamma}_3 = -50.584$



# Robust confirmatory factor analysis [4]

1. Split the sample into 2 subsets
   - outlier free basic set,
   - non-basic set.

Distance from the model is measured with ***observational residuals***

$$e_i = x_i - \hat{x} = x_i - \bar{x} - \widehat{\Lambda}\hat{f}_i.$$

Regression method of ***factor scores estimation***

$$\hat{f}_i = \widehat{\Phi}\widehat{\Lambda}^T\widehat{\Sigma}^{-1}(x_i - \bar{x}), \qquad \text{where} \qquad \widehat{\Sigma} = \widehat{\Lambda}\widehat{\Phi}\widehat{\Lambda}^T + \widehat{\Psi}.$$

$$\boxed{e_i = \left(I - \widehat{\Lambda}\widehat{\Phi}\widehat{\Lambda}^T\widehat{\Sigma}^{-1}\right)(x_i - \bar{x})}$$

***Summarized observational residuals*** are

$$\boxed{\left(e_i^S\right)^2 = e_i^T\left(\widehat{\Psi}\widehat{\Sigma}^{-1}\widehat{\Psi}\right)^{-1}e_i.}$$

# Robust confirmatory factor analysis [4]

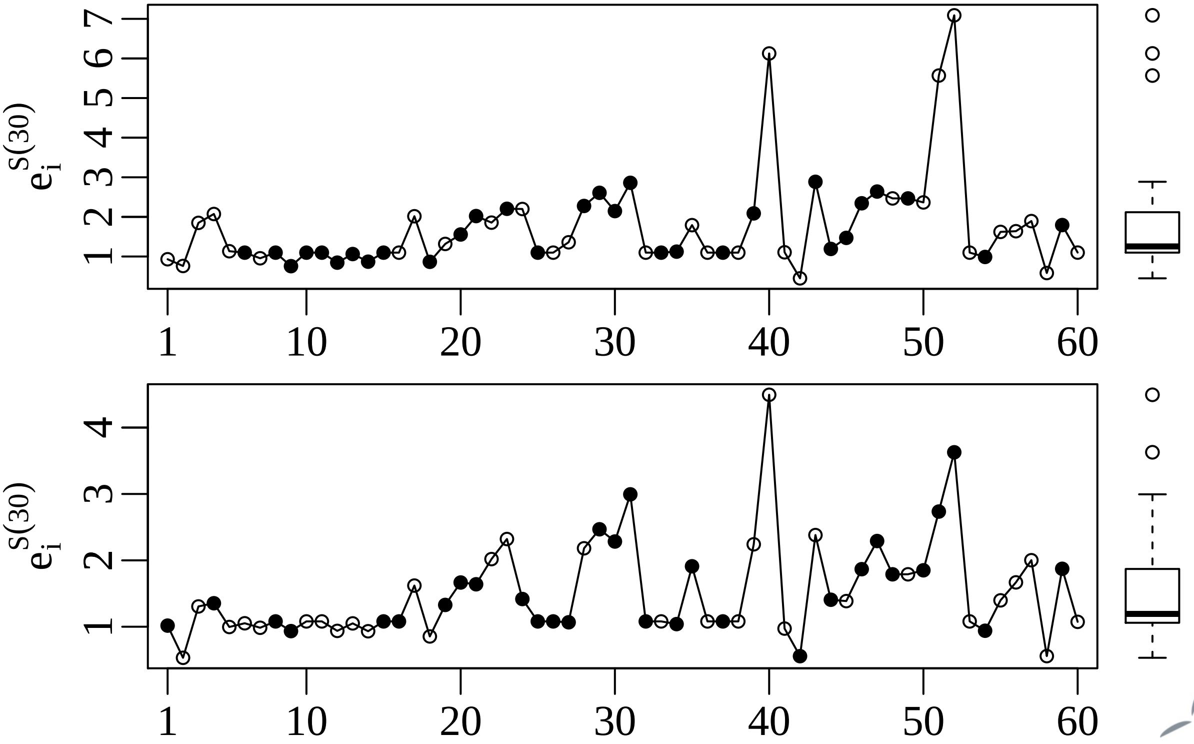
1. Split the sample into 2 subsets
   - outlier free basic set,
   - non-basic set.

1. Take a ***random subet*** $X^{(m)}$ of size $m = 30$.
2. Estimate the confirmatory factor model on the subset.
3. Compute $e_i^{(m)} = \left(I - \widehat{\Lambda}^{(m)}\widehat{\Phi}^{(m)}\left(\widehat{\Lambda}^{(m)}\right)^T\left(\widehat{\Sigma}^{(m)}\right)^{-1}\right)\left(x_i - \bar{x}^{(m)}\right)$.
4. Compute $\left(e_i^{S(m)}\right)^2 = \left(e_i^{(m)}\right)^T\left(\widehat{\Psi}^{(m)}\left(\widehat{\Sigma}^{(m)}\right)^{-1}\widehat{\Psi}^{(m)}\right)^{-1}e_i^{(m)}$.
5. Find the median of summarized observational residuals.

***Repeat*** 1000-***times*** and take the subset with the smallest median.

# Robust confirmatory factor analysis [4]

1. Split the sample into 2 subsets ► outlier free basic set,
   ► non-basic set.



# Robust confirmatory factor analysis [4]

**2.** Add observations to the basic set.

$X^{(\ell)}$ the basic set with $\ell$ observations: we wish to include one more.

1. Estimate the confirmatory factor model on the basic set.
2. Compute $e_i^{(\ell)} = \left(I - \widehat{\Lambda}^{(\ell)} \widehat{\Phi}^{(\ell)} \left(\widehat{\Lambda}^{(\ell)}\right)^T \left(\widehat{\Sigma}^{(\ell)}\right)^{-1}\right)\left(x_i - \bar{x}^{(\ell)}\right)$.
3. Compute $\left(e_i^{S(\ell)}\right)^2 = \left(e_i^{(\ell)}\right)^T \left(\widehat{\Psi}^{(\ell)} \left(\widehat{\Sigma}^{(\ell)}\right)^{-1} \widehat{\Psi}^{(\ell)}\right)^{-1} e_i^{(\ell)}$.
4. Take $\ell + 1$ observations with smallest summarized observational residuals.

# Robust confirmatory factor analysis [4]

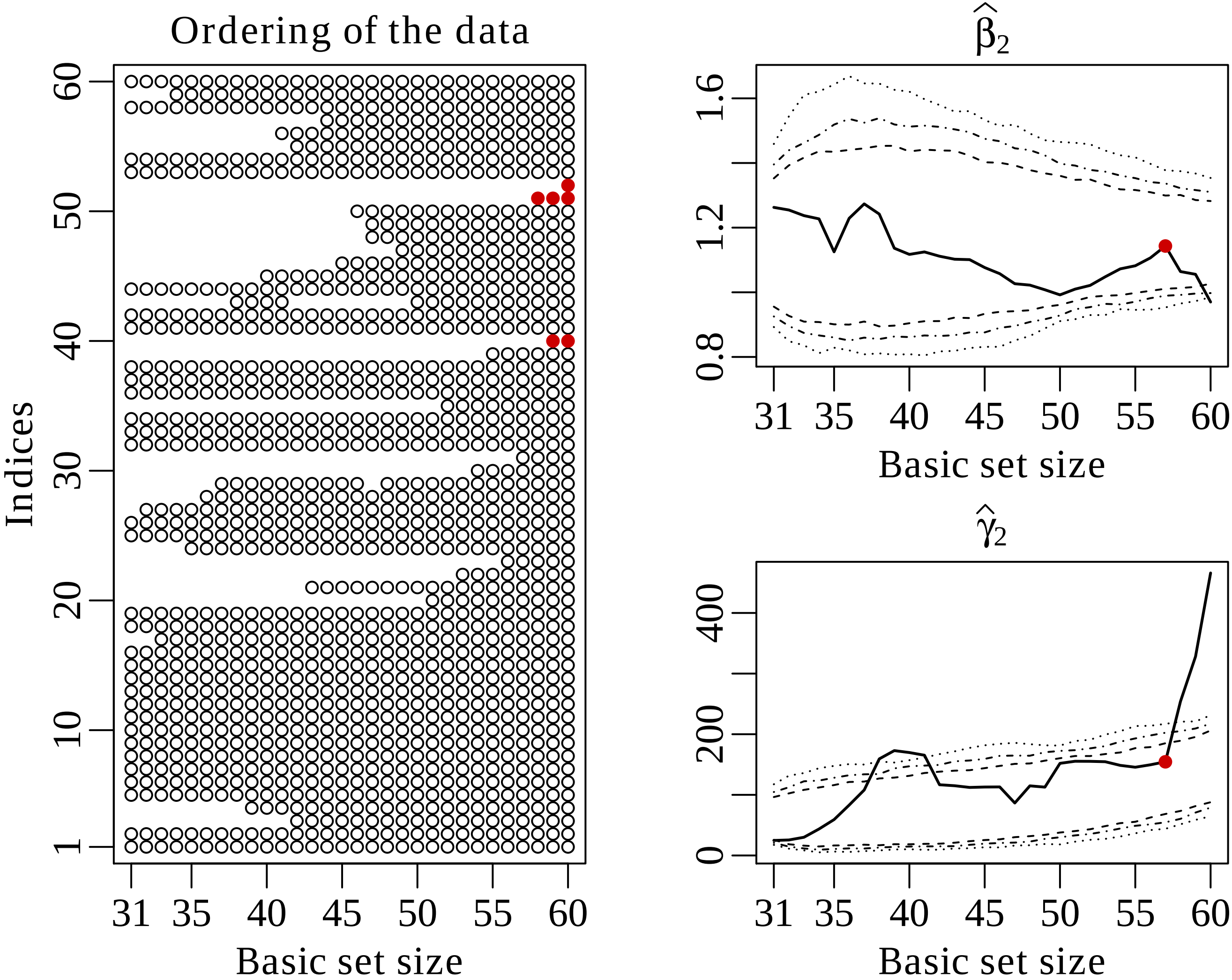
**2.** Add observations to the basic set.

$X^{(\ell)}$ the basic set with $\ell$ observations: we wish to include one more.

1. Estimate the confirmatory factor model on the basic set.
2. Compute $e_i^{(\ell)} = \left(I - \widehat{\Lambda}^{(\ell)}\widehat{\Phi}^{(\ell)}\left(\widehat{\Lambda}^{(\ell)}\right)^T\left(\widehat{\Sigma}^{(\ell)}\right)^{-1}\right)\left(x_i - \bar{x}^{(\ell)}\right)$.
3. Compute $\left(e_i^{S(\ell)}\right)^2 = \left(e_i^{(\ell)}\right)^T\left(\widehat{\Psi}^{(\ell)}\left(\widehat{\Sigma}^{(\ell)}\right)^{-1}\widehat{\Psi}^{(\ell)}\right)^{-1} e_i^{(\ell)}$.
4. Take $\ell + 1$ observations with smallest summarized observational residuals.

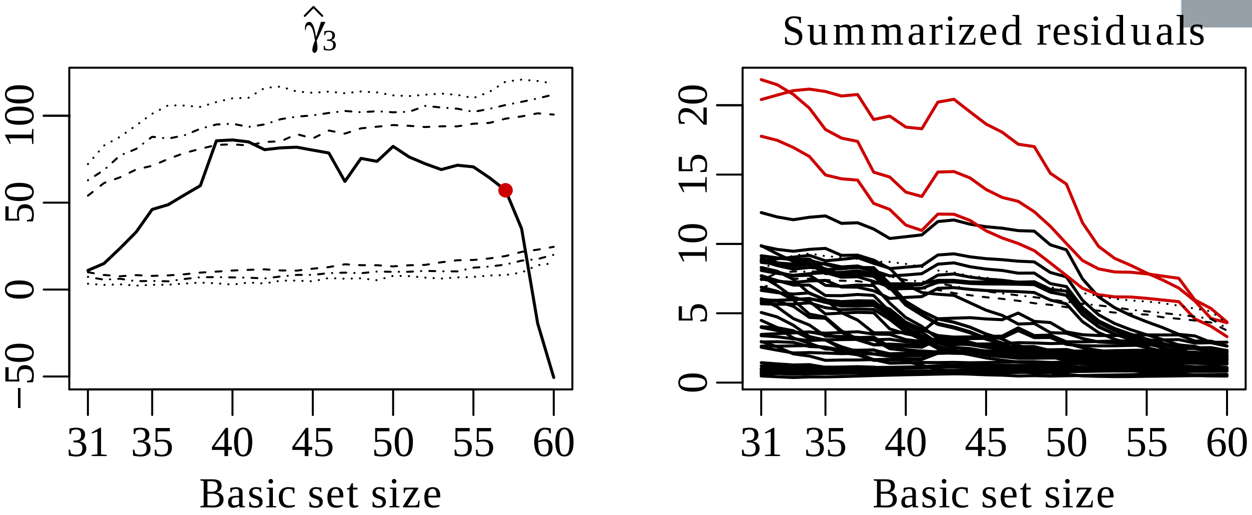
**3.** Use forward plots to show the dynamics of estimates.

- Ordering of the data.
- Parameter estimates, fit indices, summarized obs. residuals.

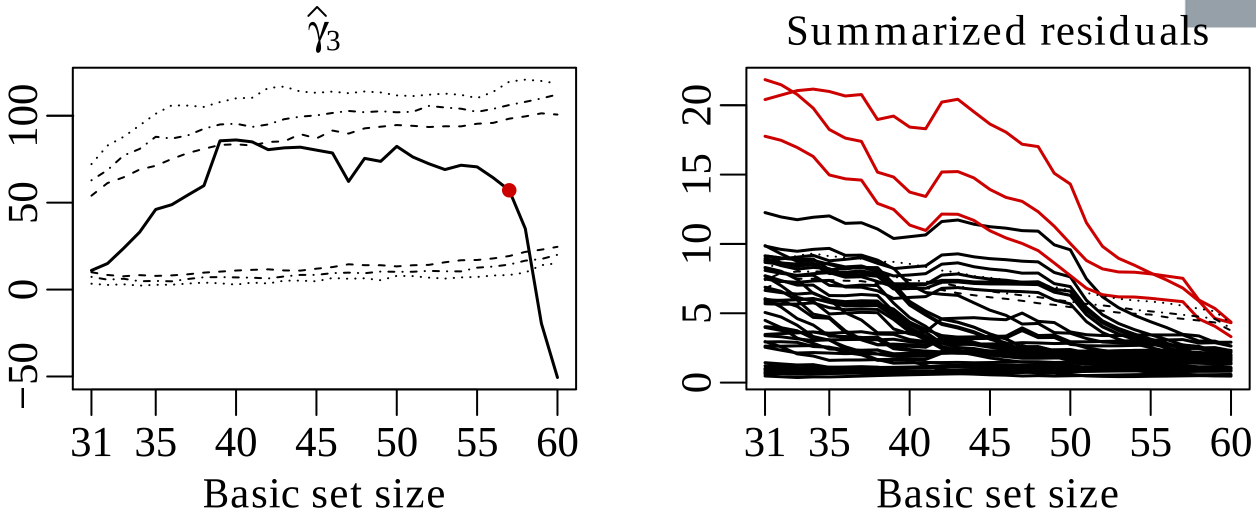


Ordering of the data
Indices
Basic set size
$\widehat{\beta}_2$
Basic set size
$\widehat{\gamma}_2$
Basic set size

# Robust confirmatory factor analysis [4]



# Robust confirmatory factor analysis [4]

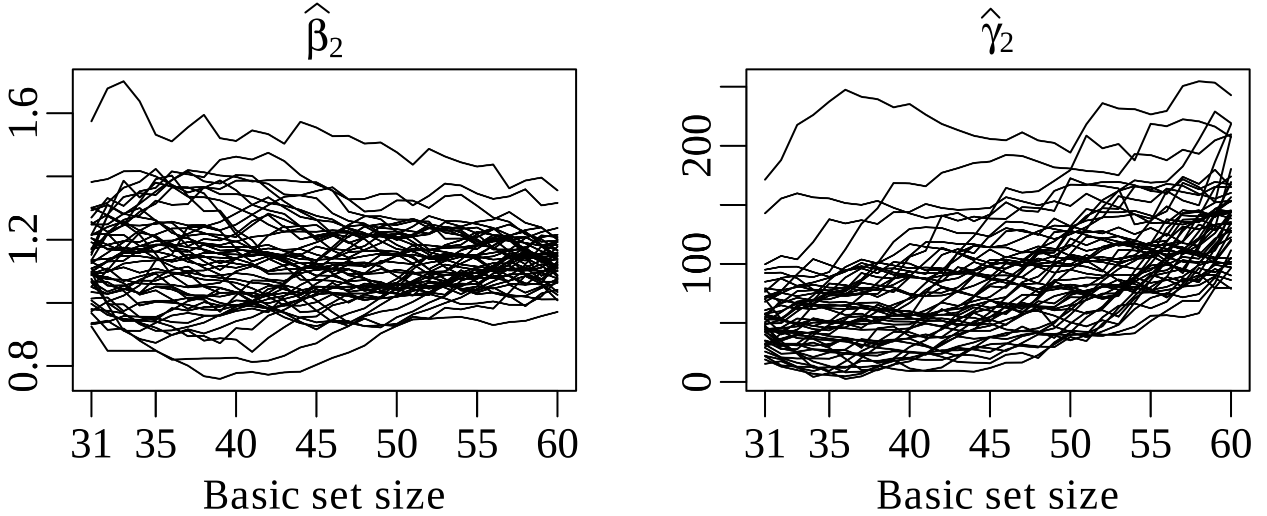



## Conclusions

- ▶ Important to consider the model in robust inference.
- ▶ The algorithm enables us to explore the data.
- ▷ Rarely used in applied research.

# Simulation envelopes

- Identify the step $\ell^*$ of the first major changes.
- Take $\widehat{\Lambda}^{(\ell^*)}$, $\widehat{\Phi}^{(\ell^*)}$, $\widehat{\Psi}^{(\ell^*)}$ and compute $\widehat{\Sigma}^{(\ell^*)}$.
- Simulate 1000 samples from $N_p(0, \widehat{\Sigma}^{(\ell^*)})$.
- Repeat forward search analysis on each of the samples.




- Find pointwise confidence interval.

# References

[1] **A. C. Atkinson, M. Riani, A. Cerioli:** *Exploring multivariate data with the forward search*, Springer-Verlag, New York, 2004.

[2] **T. A. Brown:** *Confirmatory factor analysis for applied research*, The Guilford Press, New York, 2006.

[3] **W.-Y. Poon, Y.-K. Wong:** *A forward search procedure for identifying influential observations in the estimation of a covariance matrix*, Structural Equation Modeling **11** (2004) 357–374.

[4] **A. Toman:** *Robust confirmatory factor analysis based on the forward search algorithm*, Statistical Papers, **55** (2014) 233–252.

University *of Ljubljana*

FACULTY OF ECONOMICS

# Thank you for your attention!

**ales.toman@ef.uni-lj.si**

23rd International Workshop on Matrices and Statistics
Ljubljana, June 9, 2014