Small sample solutions for SEM

Yves Rosseel
Department of Data Analysis
Ghent University – Belgium

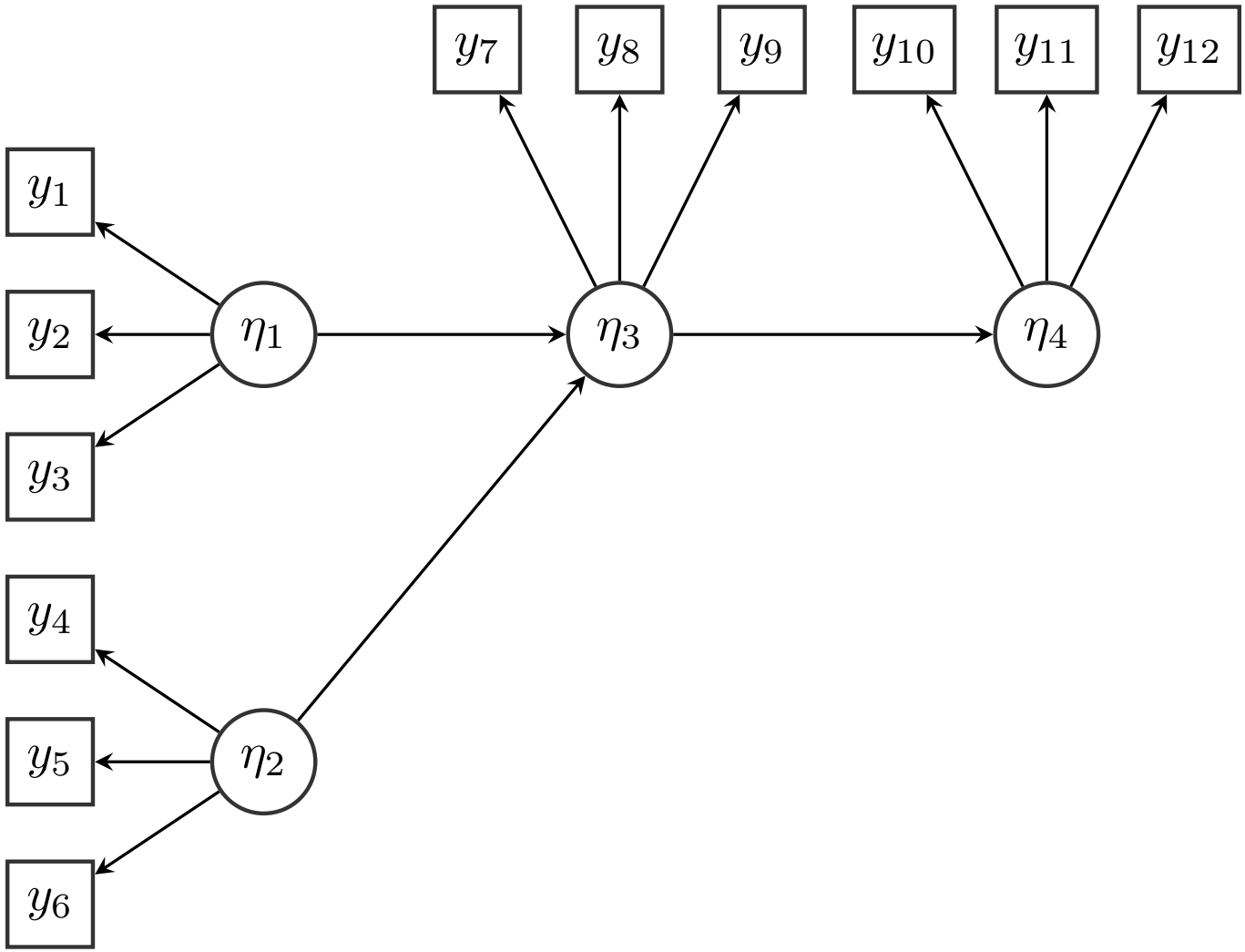
July 21, 2021
9th European Congress of Methodology
Valencia – Spain

structure of this talk

- small sample problems for structural equation modeling (SEM)
- small sample solution 1: bounded estimation (avoid convergence)
  - joint work with my PhD student Julie De Jonckere
- small sample solution 2: structural after measurement (SAM)
  - joint work with Wen Wei Loh (postdoc UGent)
- last slide

slides available at `http://lavaan.org`

context: structural equation modeling



small sample problems in SEM (in the frequentist framework)

- the statistical machinery behind SEM is based on large sample theory
- in practice, (very) small sample sizes ($N < 100$) are simply a reality
  - population is (very) small (i.e., children with severe facial burns)
  - high cost/effort per observation (i.e., fMRI data)
- this leads to many issues:
  - nonconvergence (the optimizer cannot find a solution)
  - extreme/nonsensical parameter values
  - parameter bias
  - standard errors, confidence intervals, test statistics cannot be trusted
- two problematic settings:
  - small models, (very) small sample sizes (e.g., $N = 20$)
  - large models, (relatively) small sample sizes (e.g., $N = 150$)

small sample solutions for SEM

- the most popular solution is to switch to a Bayesian approach
  - major advantage: no need for large sample asymptotics
  - correct standard errors and credible intervals even in small samples
  - but model evaluation requires a new set of skills
  - you need a prior distribution for each (free) parameter
  - when the sample size is (very) small (say, $N < 50$), Bayesian estimation only works with (highly) informative priors (otherwise, mode-switching issues may occur)
- in our SEM lab in Ghent, we try to find solutions for the frequentist framework
  - understanding and avoiding convergence issues (very small samples)
  - two-step approaches (factor score regression, SAM)
  - improving the quality of estimators (in small samples)

some background references

- new (edited) book on small samples:

  van de Schoot, R. (Ed.), Miočević, M. (Ed.). (2020). *Small Sample Size Solutions*. London: Routledge. (Free online download)

- two chapters:

  Rosseel, Y. (2020). Small sample solutions for structural equation modeling.

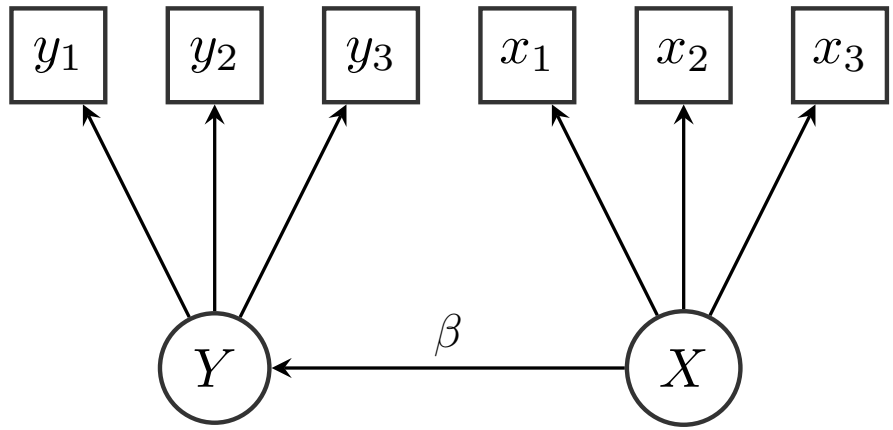
  Smid, S.C., & Rosseel, Y. (2020). SEM with small samples: Two-step modeling and factor score regression versus Bayesian estimation with informative priors.

- choice of priors in SEM:

  Van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. Psychological Methods, 23(2), 363–388.

problematic setting 1: very small samples

- consider the following SEM:



- this is a small model, with only 13 free parameters:
  - the factor loadings are set to 1, 0.8 and 0.6; all variances are set to 1.0
  - the regression coefficient is set to $\beta = 0.25$
  - (very) low reliabilities for the indicators: 0.265–0.515
- from this population model, we will generate a small sample ($N = 20$)

data generation ($N = 20$)

```
> library(lavaan)
> pop.model <- '
+     # factor loadings
+     Y =~ 1*y1 + 0.8*y2 + 0.6*y3
+     X =~ 1*x1 + 0.8*x2 + 0.6*x3
+
+     # regression part
+     Y ~ 0.25*X
+ '
> set.seed(8)
> Data <- simulateData(pop.model, sample.nobs = 20L)
```

fitting the model using ML

```
> model <- '
+     # factor loadings
+     Y =~ y1 + y2 + y3
+     X =~ x1 + x2 + x3
+
+     # regression part
+     Y ~ X
+ '
> fit.sem <- sem(model, data = Data, estimator = "ML")

lavaan WARNING: the optimizer warns that a solution has NOT been found!
```

output SEM

```
> parameterEstimates(fit.sem, header = FALSE, ci = FALSE, output = "text")[1:13,]

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y =~
    y1                1.000
    y2                1.683       NA
    y3                1.051       NA
  X =~
    x1                1.000
    x2              268.924       NA
    x3                0.428       NA

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y ~
    X                -0.159       NA

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .y1                1.706       NA
   .y2                0.763       NA
   .y3                1.066       NA
   .x1                1.408       NA
   .x2             -368.917       NA
   .x3                1.551       NA
```

R = 10000 replications: number of nonconverged solutions

sample size	no-bounds
10	5113
15	3919
20	2971
25	2237
30	1668
40	929
50	521
60	264
70	166
80	73
90	37
100	21

solution: 'bounded estimation'

- by default, lavaan (and most SEM software) uses unconstrained estimation (using quasi-Newton methods)
- given the data, we can determine data–driven lower and upper bounds for a subset of the model parameters (variances, covariances, and factor loadings)
- for more details, see

  > De Jonckere & Rosseel (submitted). Using bounded estimation to avoid nonconvergence in small sample structural equation modeling. Available from `https://osf.io/f7z6j/`

- works for all estimators, missing values, categorical, multigroup, ...
- we assume $N > P$ (P is the number of observed variables) and the model is identified
- limitation: does not work (out of the box) in the presence of linear equality constraints

solution: 'bounded estimation' (2)

- in scalar notation, we can write the one-factor model as

$$y_p = \lambda_p \, f + \epsilon_p$$

- we assume $\text{Cov}(f, \epsilon_p) = 0$ and write $\text{Var}(\epsilon_p) = \theta_p$ and $\text{Var}(f) = \psi$, and therefore

$$\text{Var}(y_p) = \lambda_p^2 \, \psi + \theta_p$$

- example: the 'standard' lower and upper bounds for θ_p are 0 and $\text{Var}(y_p)$
- simuation – four settings:
  - no bounds
  - ov.var: lower zero bounds for (residual) variances of observed indicators only
  - standard bounds: lower and upper bounds for variances and factor loadings
  - wide bounds: somewhat wider bounds (allowing for negative variances, more 'wiggle room' for the optimizer)

R = 10000 replications: number of nonconverged solutions

sample size	no-bounds	ov.var	standard	wide
10	5113	921	13	0
15	3919	414	13	1
20	2971	233	14	0
25	2237	135	11	0
30	1668	66	5	0
40	929	27	4	0
50	521	5	4	0
60	264	1	8	0
70	166	0	5	0
80	73	0	2	0
90	37	0	2	0
100	21	0	1	0

using wide bounds with lavaan 0.6-9

```
> fit.wide <- sem(model, data = Data, estimator = "ML", bounds = "wide")
> parTable(fit.wide)[,c("lhs", "op", "rhs", "free", "lower", "upper", "est")]
```

	lhs	op	rhs	free	lower	upper	est
1	Y	=~	y1	0	1.000	1.000	1.000
2	Y	=~	y2	1	-3.689	3.689	1.400
3	Y	=~	y3	2	-3.231	3.231	0.938
4	X	=~	x1	0	1.000	1.000	1.000
5	X	=~	x2	3	-4.907	4.907	1.787
6	X	=~	x3	4	-3.978	3.978	0.528
7	Y	~	X	5	-Inf	Inf	-0.055
8	y1	~~	y1	6	-0.097	2.329	1.584
9	y2	~~	y2	7	-0.102	2.445	0.915
10	y3	~~	y3	8	-0.078	1.875	1.059
11	x1	~~	x1	9	-0.064	1.526	0.636
12	x2	~~	x2	10	-0.118	2.834	-0.118
13	x3	~~	x3	11	-0.078	1.863	1.336
14	Y	~~	Y	12	0.005	2.802	0.570
15	X	~~	X	13	0.141	1.794	0.777

lavaan output with wide bounds

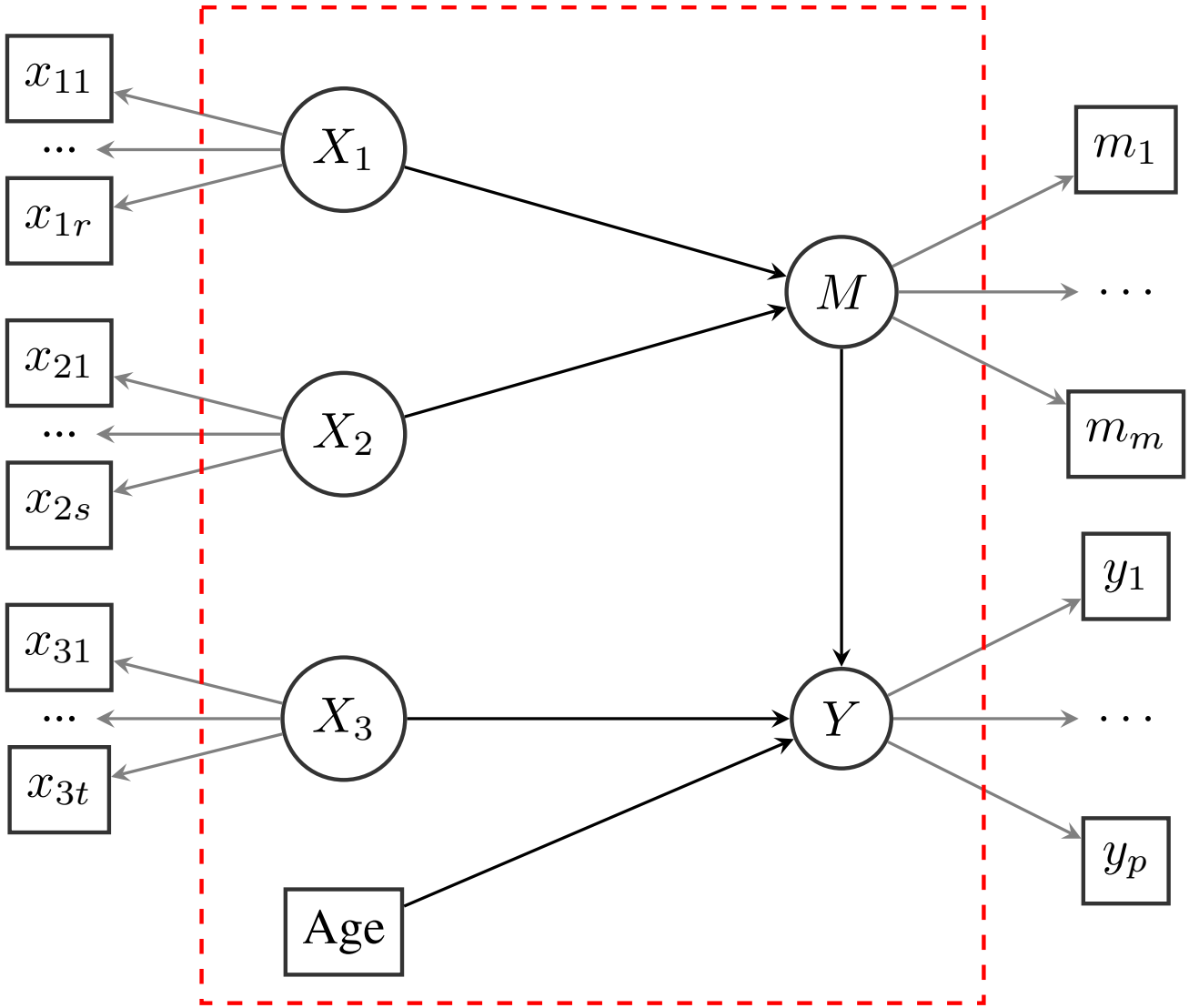
```
> parameterEstimates(fit.wide, header = FALSE, ci = FALSE, output = "text")[1:13,

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y =~
    y1                1.000
    y2                1.400    1.039    1.347    0.178
    y3                0.938    0.623    1.505    0.132
  X =~
    x1                1.000
    x2                1.787    0.932    1.918    0.055
    x3                0.528    0.305    1.733    0.083

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y ~
    X                -0.055    0.230   -0.239    0.811

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .y1                1.584    0.640    2.474    0.013
   .y2                0.915    0.832    1.100    0.271
   .y3                1.059    0.485    2.184    0.029
   .x1                0.636    0.424    1.498    0.134
   .x2        (lb)   -0.118    1.194   -0.099
   .x3                1.336    0.435    3.071    0.002
```

problematic setting 2: large models, small samples

- 'large' models with many (say > 100) parameters, but the sample size is relatively small (say $N = 150$)
- using (traditional) SEM using ML (system-wide estimation)
  - convergence issues, unstable estimation, extreme parameter values
  - very sensitive to (local) model misspecification
- proposed solution:

  Rosseel & Loh (submitted). Structural After Measurement (SAM) approach to SEM. Available from `https://osf.io/pekbm/`
- long–standing idea: decouple estimation of the measurement part and the structural part
- local SAM is a generalization of bias-corrected factor score regression
- global SAM (not discussed in this talk) can be used in settings where local SAM cannot be used

structural versus measurement



local SAM: rationale

- the measurement model:

$$\boldsymbol{y} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}$$

- to solve this for $\boldsymbol{\eta}$, we proceed as follows:

$$\begin{aligned} \boldsymbol{\nu} + \boldsymbol{\Lambda}\,\boldsymbol{\eta} + \boldsymbol{\epsilon} &= \boldsymbol{y} \\ \boldsymbol{\Lambda}\,\boldsymbol{\eta} &= \boldsymbol{y} - \boldsymbol{\nu} - \boldsymbol{\epsilon} \\ \mathbf{M}\boldsymbol{\Lambda}\,\boldsymbol{\eta} &= \mathbf{M}\left[\boldsymbol{y} - \boldsymbol{\nu} - \boldsymbol{\epsilon}\right] \\ \boldsymbol{\eta} &= \mathbf{M}\left[\boldsymbol{y} - \boldsymbol{\nu} - \boldsymbol{\epsilon}\right] \end{aligned}$$

  where $\mathbf{M}$ is $M \times P$ mapping matrix such that $\mathbf{M}\boldsymbol{\Lambda} = \mathbf{I}_M$

- we assume $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and write $\mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Theta}$; it follows that

$$\begin{aligned} \mathrm{E}(\boldsymbol{\eta}) &= \mathbf{M}\left[\mathrm{E}(\boldsymbol{y}) - \boldsymbol{\nu}\right] \\ \mathrm{Var}(\boldsymbol{\eta}) &= \mathbf{M}\left[\mathrm{Var}(\boldsymbol{y}) - \boldsymbol{\Theta}\right]\mathbf{M}^T \end{aligned}$$

local SAM: two–stage estimation

- first stage: estimation of the measurement part of the model (only); this results in estimates for $\boldsymbol{\nu}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$, collected in $\boldsymbol{\theta}_1$; three options:
  - single CFA, multiple CFAs, as many CFAs as latent variables
- three possible solutions for the mapping matrix $\mathbf{M}$:

$$\mathbf{M} = (\boldsymbol{\Lambda}^T \boldsymbol{\Theta}^{-1} \boldsymbol{\Lambda})^{-1} \boldsymbol{\Lambda}^T \boldsymbol{\Theta}^{-1} \qquad (ML)$$

$$\mathbf{M} = (\boldsymbol{\Lambda}^T \mathbf{S}^{-1} \boldsymbol{\Lambda})^{-1} \boldsymbol{\Lambda}^T \mathbf{S}^{-1} \qquad (GLS)$$

$$\mathbf{M} = (\boldsymbol{\Lambda}^T \boldsymbol{\Lambda})^{-1} \boldsymbol{\Lambda}^T = \boldsymbol{\Lambda}^+ \qquad (ULS)$$

- we estimate $\mathrm{E}(\boldsymbol{\eta})$ and $\mathrm{Var}(\boldsymbol{\eta})$ as follows:

$$\widehat{\mathrm{E}(\boldsymbol{\eta})} = \hat{\mathbf{M}} \left[\bar{\boldsymbol{y}} - \hat{\boldsymbol{\nu}}\right]$$

$$\widehat{\mathrm{Var}(\boldsymbol{\eta})} = \hat{\mathbf{M}} \left[\mathbf{S} - \hat{\boldsymbol{\Theta}}\right] \hat{\mathbf{M}}^T$$

- second stage: $\widehat{\mathrm{E}(\boldsymbol{\eta})}$ and $\widehat{\mathrm{Var}(\boldsymbol{\eta})}$ are used to estimate $\boldsymbol{\theta}_2$, the parameters related to the structural part of the model

local sam in lavaan

- lavaan version 0.6-9 or higher: function `sam()`

```
> set.seed(1234)
> Data <- simulateData(pop.model, sample.nobs = 150L)
> fit.lsam <- sam(model, data = Data)

> parameterEstimates(fit.lsam, ci = FALSE, output = "text")

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y ~
    X                 0.133    0.102    1.309    0.191

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .Y                 0.625    0.229    2.732    0.006
    X                 1.008    0.290    3.476    0.001
```

- see `?sam` for more information

local SAM: comments

- special cases of local SAM:
  - sum scores (biased)
  - sum scores + reliability (= single-indicator models)
  - factor scores (biased)
  - factor scores + Croon's correction
  - (linear) measurement error models (Fuller, 1987)
  - two-stage method-of-moments estimator (Wall & Amemiya, 2000)
- extends to the multigroup and multilevel setting
- for each submodel, a different estimation technique can be used (iterative, non-iterative)
- twostep-corrected standard errors, local and global fit measures are available

last slide

- small samples pose a challenge for SEM estimation and inference
- setting 1: small models, very small sample sizes (nonconvergence!)
  - solution: bounded estimation
- setting 2: large models, small sample sizes (misspecifications!)
  - solution: structural-after-measurement approach (SAM)
  - for each submodel, we can use a different (consistent) estimator
  - this includes bounded estimation, and/or non-iterative estimators
  - many previous solutions turn out to be a special case
- future work:
  - extension to $\mathsf{Var}(\boldsymbol{\eta} \otimes \boldsymbol{\eta})$ (Elissa Burghgraeve)
  - small-sample corrections (Jasper Bogaert)
  - improve quality of small-sample estimators (Sara Dhaene)
  - noniterative estimators

Thank you!

(questions?)

`http://lavaan.org`