lavaan: past, present and future

Yves Rosseel
Department of Data Analysis
Ghent University – Belgium

Gent – May 2022

software for SEM: commercial – closed-source

- the big four (and the main developer):
  - LISREL ('70s, Karl Jöreskog)
  - EQS ('80s, Peter Bentler)
  - AMOS ('90s, James Arbuckle)
  - Mplus (Bengt Muthén, 1998-now)
- SAS/Stat: proc CALIS, proc TCALIS
- Statistica (SEPATH), Systat (RAMONA), Stata 12
- Mx (Michael Neale, free, closed-source, '90s)
- what about SPSS?
  - SPSS bought AMOS and sells it as a separate product
  - SPSS is bought by IBM (quote from the AMOS website:)

    What it can do for your business

software for SEM: non-commercial – open-source

- outside the R ecosystem:
  - 'gllamm' in stata (Rabe-Hesketh, Skrondal & Pickles, since 2002)
  - 'semopy' in python (`https://pypi.org/project/semopy/`) (since 2018)
- R packages:
  - sem (John Fox, since 2001)
  - OpenMx (Steven Boker, Michael Neale, . . . since 2009)
  - lavaan (Yves Rosseel, since 2010)
  - lava (Klaus Holst, since 2012)
  - psychonetrics (Sacha Epskamp, since 2019)
- interfaces between R and commercial packages:
  - REQS (Patrick Mair, Eric Wu, since 2008)
  - MplusAutomation (Michael Hallquist, since 2010)

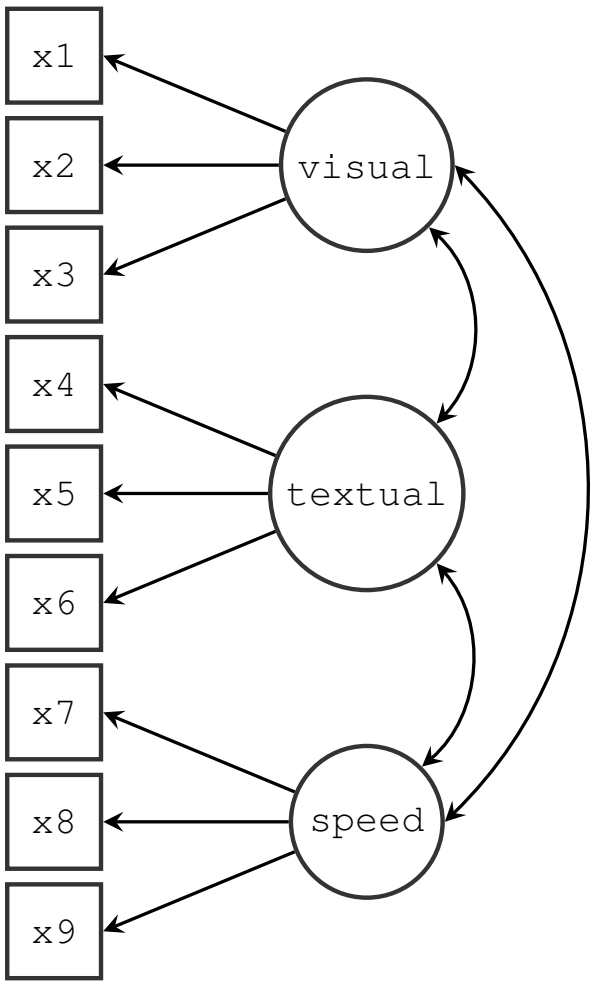
writing statistical software

- many (open-source) statistical software packages (written in R, Julia, Python, . . .) implement *new* statistical ideas
  - you can set your own standards
  - no comparison with other (existing) software
  - no community (yet)
- image writing a package for structural equation modeling (SEM)
  - there is a 'tradition', dating back more than 4 decades
  - there are already many (mostly commercial) software packages available
  - there is already a large community

the beginnings . . .

- the context:
  - in my statistical consultancy years (2000–2008), I often used LISREL, EQS or Mplus, depending on the experience of the client
  - mostly just confirmatory factor analyses (CFA)
  - often very repetitive (same model, multiple datasets)
  - it would be great if we could do everything in R, but (around 2008–2009) the only option was the **sem** package, which was too limited for my purposes
- the initial plan:
  - create a small (private) R package to do only 1 thing: CFA
  - do one thing, do it well (cfr. the Unix philosophy)
  - would be great for teaching too
  - first package (March 2009, never published): **cfa2000**

cfa2000 example: Holzinger & Swineford (1939) 3-factor CFA

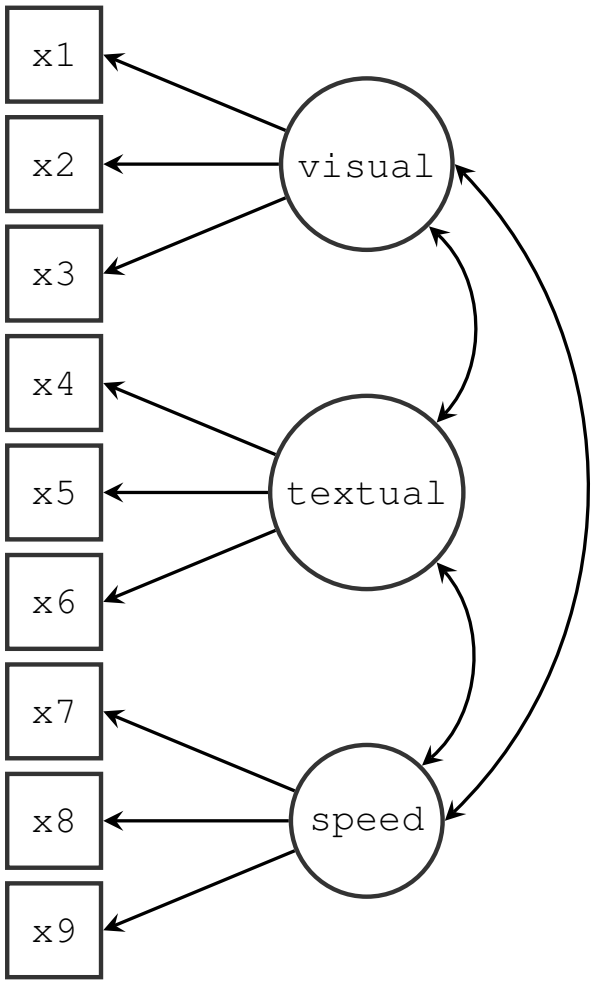



```
library(cfa2000)

# specify 3-factor CFA model
measurement.model <-
    list( visual  = c("x1","x2","x3"),
          textual = c("x4","x5","x6"),
          speed   = c("x7","x8","x9") )

# fit the model
fit <- cfa(measurement.model = measurement.model,
           data = HolzingerSwineford1939)
summary(fit)
```

cfa2000 partial output

```
Model converged normally after 35 iterations (0.146s)

  Chi-square test full model          85.306
  Degrees of freedom                     24
  P-value                            0.0000

Factor loadings:     Estimate      S.E.   z value  Pr(>|z|)
  visual
     x1                 1.000
     x2                 0.554     0.100     5.554     0.000
     x3                 0.729     0.109     6.685     0.000
  textual
     x4                 1.000
     x5                 1.113     0.065    17.014     0.000
     x6                 0.926     0.055    16.703     0.000
  speed
     x7                 1.000
     x8                 1.180     0.165     7.152     0.000
     x9                 1.082     0.151     7.155     0.000

Factor var/cov:      Estimate      S.E.   z value  Pr(>|z|)
  visual
     visual             0.812     0.146     5.564     0.000
     textual            0.410     0.074     5.552     0.000
     speed              0.263     0.056     4.660     0.000
...
```

cfa2000, August 2009, using formula-like expressions



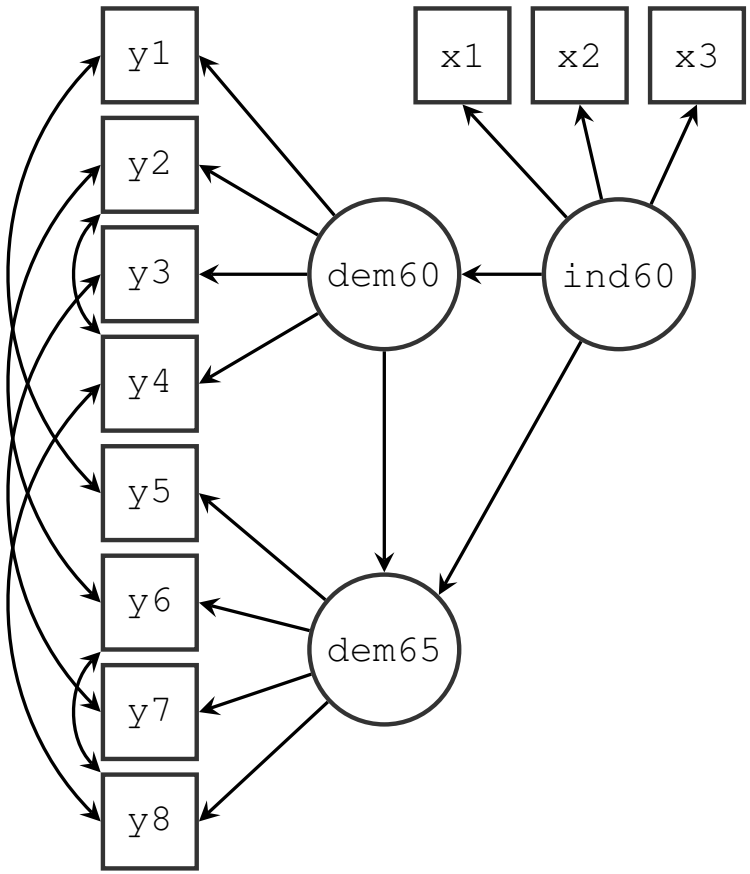

```
visual  =~ x1 + x2 + x3
textual =~ x4 + x5 + x6
speed   =~ x7 + x8 + x9

fit <- cfa(measurement.model = list(visual,
                                    textual,
                                    speed),
           data = HolzingerSwineford1939)

summary(fit)


fit.measures(fit, c("cfi", "rmsea", "srmr"))
```

- nice and easy
- but what about exogenous covariates?
- we need a SEM package after all

second package: ‘semplus’, September 2009

```
Package: semplus
Type: Package
Title: Structural Equation Modeling
Version: 0.9-9
Date: 2009-09-16
Author: Yves Rosseel <yves.rosseel@ugent.be>
Maintainer: Yves Rosseel <yves.rosseel@ugent.be>
Description: Structural Equation Modeling with a formula-based interface
Depends: methods, MASS
License: GPL version 2 or later
LazyLoad: yes
LazyData: yes
Packaged: 2009-10-13 08:18:48 UTC; yves
```

semplus: political democracy example




```
# measurement part
mm <- list(ind60 =~ x1 + x2 + x3,
           dem60 =~ y1 + y2 + y3 + y4,
           dem65 =~ y5 + y6 + y7 + y8)

# correlated errors
ce <- list(y1 ~~ y5,
           y2 ~~ y4 + y6,
           y3 ~~ y7,
           y4 ~~ y8,
           y6 ~~ y8)

# structural part
eqs <- list(dem60 ~ ind60,
            dem65 ~ ind60 + dem60)

fit <- sem(measurement.model = mm,
           eqs = eqs,
           ce = ce,
           data = BollenDemocracy)
```

Jan 2010 – semplus using list() to specify the model

```
model <- list(

    # latent variable definitions
        ind60 =~ x1 + x2 + x3,
        dem60 =~ y1 + y2 + y3 + y4,
        dem65 =~ y5 + y6 + y7 + y8,

    # regressions
        dem60 ~ ind60,
        dem65 ~ ind60 + dem60,

    # residual (co)variances
        y1 ~~ y5,
        y2 ~~ y4 + y6,
        y3 ~~ y7,
        y4 ~~ y8,
        y6 ~~ y8
)
```

March 2010 – semplus using a string literal

```
model <- '
  # latent variable definitions
     ind60 =~ x1 + x2 + x3
     dem60 =~ y1 + y2 + y3 + y4
     dem65 =~ y5 + y6 + y7 + y8

  # regressions
    dem60 ~ ind60
    dem65 ~ ind60 + dem60

  # residual correlations
    y1 ~~ y5
    y2 ~~ y4 + y6
    y3 ~~ y7
    y4 ~~ y8
    y6 ~~ y8
'

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

from semplus to lavaan

- the package was named ‘semplus’ because it could do ‘more’ than the sem package
- and it contained the word ‘mplus’
- I contacted the Mplus team (24-02-2010), with some technical questions
- and received an email back (03-03-2010) saying:

> *We own the Mplus trademark. Using the name “semplus” can be construed as a trademark infringement and might also imply our endorsement.*

- our legal department was eager to fight them
- eventually, I changed the name to ‘lavaan’ (latent variable analysis)
- (I never got an answer from Mplus for my technical questions)

lift off

- lavaan 0.3-1 (about 6470 lines) was released on CRAN on 11 May 2010
- presented at useR 2010 (NIST, Gaithersburg, Maryland, USA)
  - no interest at all...
- invited speaker at Psychoco 2011 in Tübingen:
  - much interest among the participants
  - Achim Zeileis proposed to create a 'special volume' for the journal of statistical software on 'Psychometric Computing in R': `https://www.jstatsoft.org/issue/view/v048`
  - this is where the 'lavaan paper' was published (in 2012)
- from 2013 onwards: many workshops and presentations

why do we need lavaan?

- original propaganda, from 2010–2013 . . .
  1. **lavaan** is for statisticians working in the field of SEM (and beyond)
     - it seems unfortunate that new developments in this field are hindered by the lack of open source software that researchers can use to implement their newest ideas
  2. **lavaan** is for teachers
     - teaching these techniques to students was often complicated by the forced choice for one of the commercial packages
  3. **lavaan** is for applied researchers
     - keep the syntax simple, provide all the features they need
- . . . still true today

the next years

- more and more features were added
- HUGE step: 0.5 added support for categorical data (binary/ordinal)
  - getting the 'asymptotic covariance matrix of the sample statistics' right was a challenge
- more attention for:
  - optimization, scaling, stopping criteria, ...
  - numerical stability, numerical methods
  - what to do if a covariance matrix is not positive-definite?
  - speed
  - ...
- biggest challenge in the early years:
  - the lavaan output was not (always) identical to the output of other (commercial) packages

my program gives (slightly) different results!

- example: Satorra-Bentler scaled test statistic for a 3-factor CFA model using the 'classic' Holzinger and Swineford 1939 data (N=301)

program	SB test statistic
lavaan 0.5-22	80.872
Mplus 7.11	81.908
EQS 6.1	81.141
LISREL 8.72	77.396

- experts (often) could not explain these differences
- users of lavaan complained and believed that lavaan's results could not be trusted

the 'mimic' argument

- all fitting functions in lavaan have a `mimic` argument:
  - `mimic="EQS"` to mimic EQS computations
  - `mimic="Mplus"` to mimic Mplus computations
  - `mimic="LISREL"` to mimic LISREL computations (in dev)
  - this was originally intended to convince users that lavaan could produce 'identical' results as the (commercial) competition
  - it became a design goal on its own, but I gave up eventually
- example:

program	SB test stat	lavaan + mimic	SB test stat
lavaan 0.5-22	80.872	`mimic="lavaan"`	80.872
Mplus 7.11	81.908	`mimic="Mplus"`	81.908
EQS 6.1	81.141	`mimic="EQS"`	81.141
LISREL 8.72	77.396	`mimic="LISREL"`	77.396

studying the black box (closed-source) software

- I spent a ridiculous amount of time trying:
  1. to understand (and document) why we observe many subtle (and less subtle) numerical differences in the output of current modern SEM programs
  2. to reproduce results computed by older versions of SEM programs (reproducibility)
  3. to study and compare these (computational and numerical) differences in order to better understand their characteristics
- this is not unlike software archeology
- I learned a lot, and I am still processing the 'data'
- I discovered the lost art of coding numerical software in an efficient, stable, and elegant way

lavaan today

- current version 0.6-11 (about 70,000 lines)
- the official website:

  https://lavaan.org

- the lavaan paper:

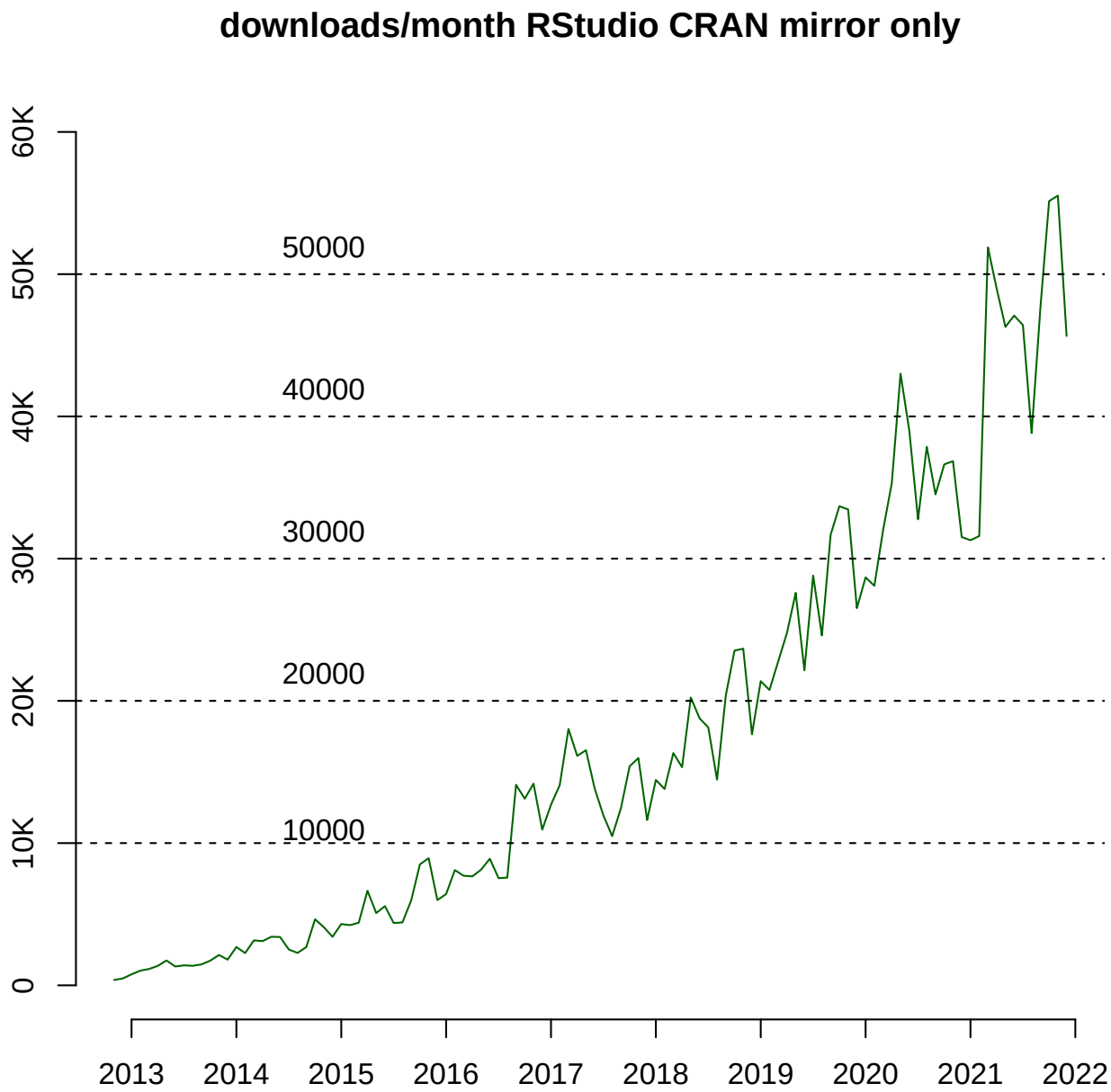
  Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.

- lavaan source code:

  https://github.com/yrosseel/lavaan

- lavaan discussion group (mailing list)

  https://groups.google.com/d/forum/lavaan


downloads/month RStudio CRAN mirror only
0
10K
20K
30K
40K
50K
60K
10000
20000
30000
40000
50000
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022

lavaan today (2)

- lavaan is a mainstream package for SEM
  - in the last years, 3 books about 'lavaan' have been published
- lavaan is widely used for both teaching and research
- the lavaan 'ecosystem' contains about 90 packages that depend on, or extend lavaan:

bain Bayesrel BifactorIndicesCalculator blavaan bmem bnpa bruceR coefficientalpha conmet CoTiMA covsim cSEM detectR discnorm dmacs eatRep EFAtools EffectLiteR EGAnet eqs2lavaan equaltestMI ezCutoffs faoutlier fSRM gimme gorica IIVpredictor influence.SEM IPV jmv JWileymisc kfa lavaan.shiny lavaan.survey lavaanPlot lcsm lsl lslx lvnet matrixpls MBESS medmod MedSurvey merDeriv metaSEM MIIVsem misty MonteCarloSEM multid multilevelTools nlsem nonnest2 pathmodelfit pompom processR profileR PROsetta pscore psychonetrics psycModel pwr2ppl qgraph RAMpath regmed regsem Replication restriktor RMediation rosetta RSA rsem semdrw SEMgraph seminr semnova semPlot semptools SEMsens semTable semTools semtree sesem ShortForm simsem simstandard thurstonianIRT tidySEM umx unusualprofile vampyr WebPower

growing pains

- the more users, the larger my reponsibility
- I spend more time 'testing' than coding
- each update is somewhat of a nightmare
  - you shall not break a package that depends on your package!
  - you shall not alter the way the output looks!
  - you shall not change the numbers in the output! (same model, same data)
  - you shall not make any mistakes! (lavaan users have come to expect that everything works perfectly, all the time)
- software design: if only I could start over, with the knowledge I have today
- users almost want more features
- coding becomes increasingly more complex

why do we keep doing this?

- why stop:
  - no funding (in Belgium)
  - no support at the faculty/university level: writing software is not in my job description
  - a few (lavaan) users are not very friendly
- why continue?
  - it is (for me) a way to learn about SEM, numerical techniques, statistics, mathematics, ...
  - it feels more useful than writing yet another paper
  - you meet interesting people
  - open-source (statistical) software is too important
- but at some point (lavaan 1.0?), others will have to take over

wishlist

- more features
- technical documentation
- a nice user manual (or a book)
- a C++ backend (for speed)
- more connections (with other frameworks)
- more modularity
- a beautiful path diagrammer
- ...
- more help from (skilled) programmers
- someone (or some organization) to take over

thanks!

- you and all other lavaan users
- teachers using lavaan
- researchers using lavaan to advance the field of SEM
- code contributors, package developers
- discussion group maintainers
- workshop organizers
- seminar organizers
- ...
- please keep on using open-source software!