R in Psychometrics and Psychometrics in R

Jan de Leeuw

R in Psychometrics and Psychometrics in R

Jan de Leeuw, UCLA Statistics

In psychometrics, and in the closely related fields
of quantititative methods for the
social and educational sciences, R is not yet used
very often. Traditional mainframe packages such
as SAS and SPSS are still dominant at the user-level,
Stata has made inroads at the teaching level,
and Matlab is quite prominent at the research level.

In this paper we define the most visible techniques
in the psychometrics area, we give an overview of
what is available in R, and we discuss what is missing.
We then outline a strategy and a project to fill in
the gaps. The outcome will hopefully be a more
prominent position of R in the social and behavioral
sciences, and as a result less of a gap between these
disciplines and mainstream statistics.

1. What is Psychometrics ? How is it related to other Foometrics ?

2. How much R is there in Psychometrics ? Can there be more ? Should there be more ?

3. How much Psychometrics is there in R ? Will there be more ? What is missing.

A recent overview of what Psychometricians themselves think about Psychometrics is in Statistica Neerlandica, 60, 2006, 135-144.

If *Foo* is a science then *Foo* often has both an area *Foometrics* and an area *Mathematical Foo*.

Mathematical Foo applies mathematical modeling to the *Foo* subject area, while *Foometrics* develops and studies data analysis techniques for empirical data collected in *Foo*.

What we call statistics is the union of the various *Foometrics* over all *Foo*. Not the intersection, but the union.

Each of the social and behavioural sciences has a form of *Foometrics*, although they may not all use a name in this family.

Clearly Economics, Psychology, Biology, Archeology, Anthropology, and Environmental Science have their own *Foometrics*.

And then there are various recent upstarts such as *Cliometrics*, *Informetrics*, *Bibliometrics*, *Behaviormetrics*, *Ecolometrics*, *Cybermetrics*, and *Scientometrics*.

Sociology would like to have *Sociometrics*, but the name was already in use for something quite different. *Historiometrics* and *Archeometrics* are there, but struggling.

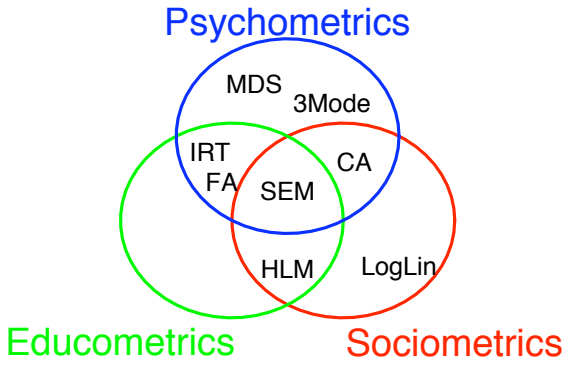
Education does not really have *Educometrics*, but we'll use it anyway.

Social sciences in which data are less prominent usually have books and conferences with titles such as *Statistics in Foo* -- they will have their very own *Foometrics* in the future.

In this presentation we'll look at *Psychometrics* and *Educometrics*, with a dash of *Sociometrics* and *Econometrics*.

Psychometrics and *Educometrics* have been around for a long time, at least since Galton, and their development has been very closely linked and often the two have been indistinguishable.

So we do not distort reality too much if we just simply call the body of techniques we discuss *Psychometrics*.



R in Psychometrics

Traditionally psychologists doing data analysis use SPSS, some use SAS.

Psychometricians developing data analysis techniques use Matlab, sociometricians and econometricians (at least in the US) tend to use Stata.

The situation in France or England may be quite different.

This has mainly historical reasons -- it has to do with where these packages originated.

But it also has to do with the rather large distance between areas such as psychometrics and (academic) statistics, which again has historical reasons, most of them silly. Typically, there is not much interaction, despite institutions like ETS and Bell Labs.

And thus the R revolution has largely passed psychometrics by.

Psychometric software is often distributed by incorporating it as modules in the standard packages (SPSS, SAS, Stata), using either native matrix routines if available or linking in compiled code. This guarantees good distribution, some money, but certainly not efficient computation.

Examples are CATEGORIES for CA in SPSS, PROC CALIS for SEM and PROC GLM for MLA in SAS, and gllamm for SEM and MLA in Stata.

In addition, psychometricians tend to write stand-alone packages for specific families of techniques. This is often compiled code combined with a suitable GUI.

The prototypical example are SEM packages like LISREL, EQS, M-PLUS, AMOS, or MLA packages such as HLM or ML-WIN -- but there many similar stand-alone packages for IRT and CA and LLA as well. In fact the number of CA packages in marketing, for example, is staggering.

Writing stand-alone compiled packages often means that the psychometrician is a small company, trying to make money. It also means a certain form of competition, which does not really belong in academia. And it means proprietary software, which costs money.

More seriously, perhaps, is that this approach means black-box software, in which the machinery is almost completely hidden. This means the user often will not even try to understand what is going on.

The techniques implemented in the black-box packages are often complicated (many parameters, complicated optimizations, doubtful standard errors).

This is necessarily true: simpler techniques are already implemented in SAS or SPSS and usually the institution has a site license for those.

Thus we have *Deus Ex Machina* software: it transforms large datasets into rather mysterious pictures or tables that are nevertheless acceptable, and often even encouraged, by peers and journals.

Promoting the teaching and the use of R in psychometrics has some major advantages.

1. The distance to academic statistics becomes smaller.

2. Software is more transparent -- driven by interpreted code. Reproducible results are more likely.

3. One can teach *with* R. One can teach SAS, but one cannot teach *with* SAS (or LISREL).

4. Software should be free.

Psychometrics in R

We give a quick inventory of the psychometric software now available or soon to be available in R.

I shall concentrate on CRAN, of course, while mentioning some additional easily available packages on other servers.

We shall see there is quite an abundance, although in most cases all forms of organization is lacking and duplications abound.

The psychoR project.

I have been writing and planning a substantial number of psychometric techniques in R. Eventually they will grow up to be packages.

They are not intended to replace existing packages: let a thousand flowers bloom. They are written following the familiar programming philosophy that you can write FORTRAN in any language. You can find them at http://www.cuddyvalley.org/psychoR

JSS (www.jstatsoft.org) is planning a number of special issues, with appropriate guest editors, and names such as

-- *R in Psychometrics*
-- *R in Econometrics*
-- *R in Sociometrics*

and whatever else anyone suggests along these lines. Of course there is an inherent risk in actually making constructive suggestions -- you may wind up to be a guest editor.

1. Simple and Multiple Correspondence Analysis.

There is CA and MCA both in *MASS*, in *ade4,* in *FactoMineR*, and in *homals*. Many variations (Canonical CA, Fuzzy CA, Detrended CA, Multiway CA, Discriminant CA, Co-CA) in *ade4*, *PTAk, cocorresp, vegan, made4.* At least three more CA packages (Greenacre, Beh, De Leeuw) with various options are currently being prepared.

An Embarrassment of riches.

The *homals* (soon *gifi*) package does what SPSS Categories does, and more. It has many forms of multivariate analysis with optimal scaling, organized as extensions of MCA. But it is rather poorly documented.

$$\min_{X'X=I}\min_{Y_j\in\mathcal{Y}_j}\sum_{j=1}^m \mathbf{tr}\,(X-G_jY_j)'(X-G_jY_j)$$

CA and MCA are extended in the *psychoR* project with distance association models (*distassoc*, *scalassoc*, *singlepeaked, logithom*), which also generalize many common IRT models.

2. Item Response Theory

ltm fits the simple Rasch model, the graded logistic model for polytomous data, and the linear multidimensional logistic model.

mprobit fits the multivariate binary probit model.

Logistic IRT is related to Gaussian ordination, implemented in various forms in *VGAM*.

More Rasch model fitting packages are on their way.

In *psychoR* we have

$$\sum_{i=1}^n\sum_{j=1}^m\sum_{\ell=1}^{k_j} y_{ij\ell}\log\frac{\beta_{j\ell}\exp(\eta(x_i,y_{j\ell}))}{\sum_{\nu=1}^{k_j}\beta_{j\nu}\exp(\eta(x_i,y_{j\nu}))}$$

$$\sum_{i=1}^n\sum_{j=1}^m\sum_{\ell=1}^{k_j} y_{ij\ell}\log\Phi(\tau_{j\ell}-\eta(x_i,y_{j\ell}))-$$

$$-\Phi(\tau_{j\ell-1}-\eta(x_i,y_{j\ell-1}))$$

$$\eta(x_i,y_j)=\begin{cases}x_i'y_j,\\ -\|x_i-y_j\|,\\ -\|x_i-y_j\|^2.\end{cases}$$

This covers most IRT models, and then some. There are also versions for marginal maximum likelihood estimation, and for cross tables with frequencies in the form

$$\sum_{i=1}^n\sum_{j=1}^m y_{ij}\log\lambda_{ij}-\lambda_{ij},$$

$$\lambda_{ij}=\alpha_i\beta_j\exp(\eta(x_i,y_j))$$

This generalizes CA, the RC model, Quasi-Symmetry, and so on.

3. Factor Analysis (see also under SEM)

factanal in *stats* can do exploratory maximum likelihood factor analysis.

MCMCpack has some options for sampling from the posterior for ordinal and mixed factor models. These are related to IRT.

homals can do various forms of mixed data principal component analysis, which the French sometimes call FA. See also *FactoMineR*.

4. Three-mode Analysis

PTAk has various forms of k-mode component analysis or singular value decomposition, popular in both psychometrics, chemometrics, and fMRI analysis.

Although there is a three-mode slot in the *psychoR* project, currently *PTAk* seems to cover most of the useful analysis.

5. Structural Equations Models

sem fits SEM's using the RAM specification. This is quite general, and allows one to specify arbitrary path models with observed and latent variables.

In order to compete with the stand-alone programs *sem* may need various constraints, confirmatory analysis, asymptotically distribution free methods, ordinal variables, and hierarchical structures.

psychoR has a slot for least squares SEM. Find a patterned matrix A of coefficients and a matrix of transformed (quantified and standardized) variables B such that

$$\min_{A\in\mathcal{A}}\min_{B\in\mathcal{K}\cap\mathcal{S}}\sum_{i=1}^{n}\mathbf{tr}\,A'B'BA$$

Some of the blocks in B can also be "latent variables", which basically means they are completely missing and are only defined by the orthogonality constraints.

6. Multidimensional Scaling

There is non-metric MDS in *MASS*, *labdsv*, *ecodist*, *vegan* and *xgobi/ggobi*. These are all Kruskal-type least squares loss function using step-size gradient optimization methods.

There is classic (Torgerson) metric MDS in stats, and Principal Coordinate Analysis (Gower) in *ecodist*, *ade4*, *labdsv*, and *vegan*.

psychoR has metric and non-metric least squares multidimensional scaling, including unfolding individual difference models, using the *SMACOF* majorization algorithm.

$$\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ijk}(\delta_{ijk} - d_{ij}(X_k))^2$$

It also has least squares squared-distance multidimensional scaling, using either the *ALSCAL* or the *ELEGANT* algorithm.

$$\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ijk}(\delta^2_{ijk} - d^2_{ij}(X_k))^2$$

We do not discuss HLM and LogLin because they are mostly outside Psychometrics.

In any case, it seems that quite a few procedures (in many cases packages) are available, and more are coming on line regularly.

It seems that providing more options and better plots will pay off in the long run, but GUI's and spreadsheet data editors (for instance, a diagram editor for SEM) also seem to be a necessary condition for acceptance.