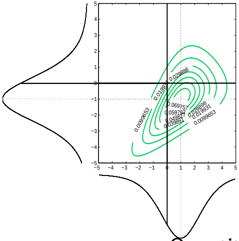
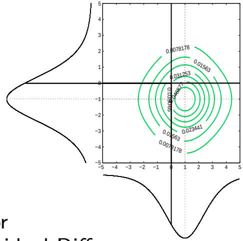
Modeling residual dependencies in latent variable models with copulas.



J. Braeken



Research group for
Quantitative Psychology and Individual Differences
K.U. Leuven

2008 / Leuven

Direct Measurement!

Example: Weight

- To be isolated physically
- Operational method: balance
- Fixed scale convention: kilogram

“Don’t step on it . . . it makes you cry.”

Direct Measurement?

Example: Optimism

- Not to be isolated physically
- No commonly agreed operational method
- What scale?

The four basic personality types

Indirect Measurement

Problem: you can not directly touch what you want to measure!

Develop Operational Method

- Gather other manifest observable variables that can provide indirect evidence
- Multiple observables
  $\Rightarrow$ Reliability
- Spread out over domain you want to say something about
  $\Rightarrow$ Validity

From these indicators,
inferences about the construct of interest can be made.

Latent Variable Model

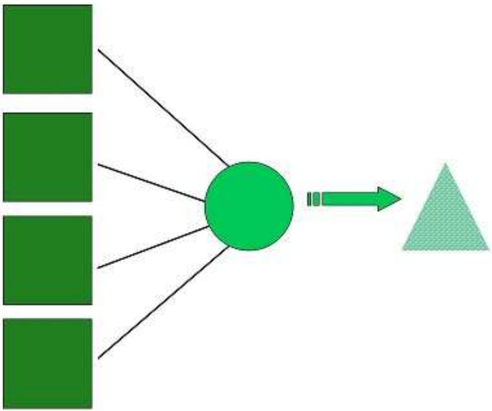
Indirect measurement: All "Art and contents Expertise" ?

Statistical model

AIM = help objectify and quantify the inferences described above

- Deriving a latent variable underlying the set of observable variables
- Acting like a summary of the gathered information
- Reflecting the construct of interest

Latent Variable Model



Working Example

Spelling proficiency

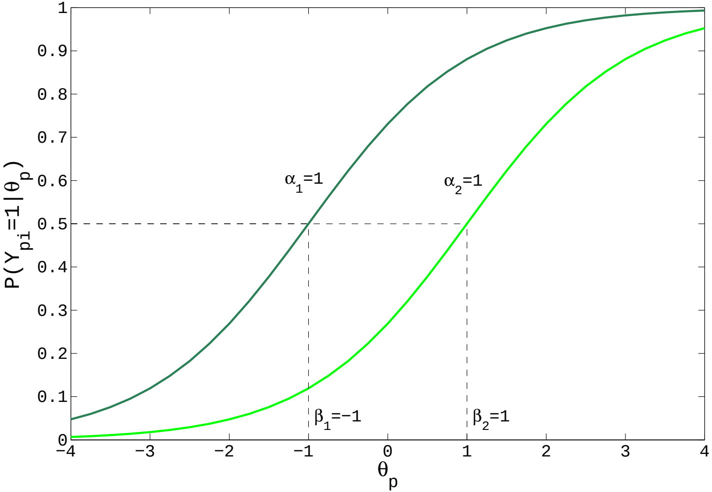
- Not something physically ⇒ indirect measurement
- Operational method: set of words to construct a ‘spell test’
- Scale: number of words correct?

To think about

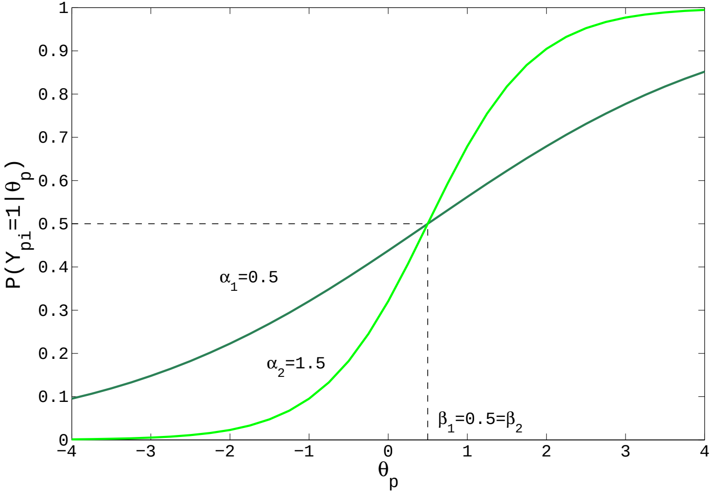
Not all words

- equally difficult to spell
- equally differentiating between good and bad spellers

Model: word Difficulty β_i



Model: word Differentiation α_i



Model for one word

Probability function

$$\Pr(Y_{pi} = y_{pi} | \theta_p) = \frac{\exp(y_{pi}\alpha_i(\theta_p - \beta_i))}{1 + \exp(\alpha_i(\theta_p - \beta_i))}$$

Distribution function

$$F_{Y_{pi}|\theta_p}(y_{pi}) = \begin{cases} 0 & \text{for } y_{pi} < 0 \\ \Pr(Y_{pi} = 0 | \theta_p) & \text{for } y_{pi} = 0 \\ 1 & \text{for } y_{pi} = 1. \end{cases}$$

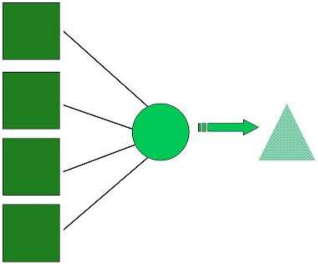
Model for the whole set of words

Logic

- θ_p increases $\Rightarrow Pr(Y_{pi} = 1)$ increases
- $Y_{pi} = 1 \Rightarrow Pr(Y_{pj} = 1)$ increases

Thus, you can learn something about Y_{pi} from Y_{pj} = Dependence.

That "something" is due to their common ground θ_p.



Model for the whole set of words

Local Stochastic Independence

$$F_{\mathbf{Y}_p|\theta_p}(\mathbf{y}_p) = F_{Y_{p1}|\theta_p}(y_{p1}) \times \ldots \times F_{Y_{pI}|\theta_p}(y_{pI}) = \prod_{i=1}^{I} F_{Y_{pi}|\theta_p}(y_{pi})$$

Once you know θ_p, you can't learn something extra about Y_{pi} from Y_{pj}.

Implications: latent variable θ_p completely accounts for

- person ($p = 1, \ldots, P$) heterogeneity = individual differences
- variable ($i = 1, \ldots, I$) homogeneity = dependence

Each variable contains unique additive information on θ_p.

Common Reality: Residual dependence

Consider following test structure

A subset

- moose
- elephant
- leopard

B subset

- offside
- penalty
- referee

Current model: All dependence between a person's responses Y_{pi} on the words $i = 1, \ldots, 6$ are due to θ_p

$$F_{\mathbf{Y}_p|\theta_p}(\mathbf{y}_p) = \prod_{i=1}^{I} F_{Y_{pi}|\theta_p}(y_{pi})$$

Common Reality: Residual dependence

Problem

homogeneity within subsets of Y_{pi}

- not only due to θ_p
- BUT also due to the common subset themes!

So in fact, even if you know θ_p,
you can learn something about Y_{pi} from some Y_{pj}!!

$\Rightarrow$ basic model fails to correctly account for the dependence structure among the observables

$\Rightarrow$ biased results and inferences

Logic of our Approach

Problem

$$
\begin{aligned}
F_{\mathbf{Y}_p|\theta_p}(\mathbf{y}_p) &= \prod_{i=1}^{I} F_{Y_{pi}|\theta_p}(y_{pi}) \\
&= \prod_{s=1}^{S} \left(\prod_{i=1}^{I} F_{Y_{pi}|\theta_p}(y_{pi}); i \in J_s \right) \\
&=> \prod_{s=1}^{S} C\left(F_{Y_{pi}|\theta_p}(y_{pi}); i \in J_s \right)
\end{aligned}
$$

Logic of our Approach

Solution

Look for more general type of functions C that

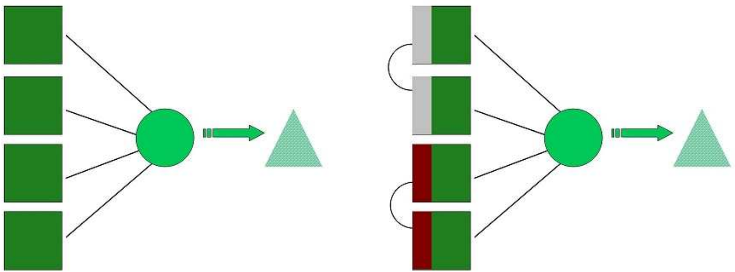
- allow for extra dependence within word cluster
- don't change the model for an individual word $F_{Y_{pi}|\theta_p}(y_{pi})$
- have simple structure and clarity

Such function = Copula Function

$\Rightarrow$ *Results in better description and prediction of the gathered data and hence of the construct of interest.*

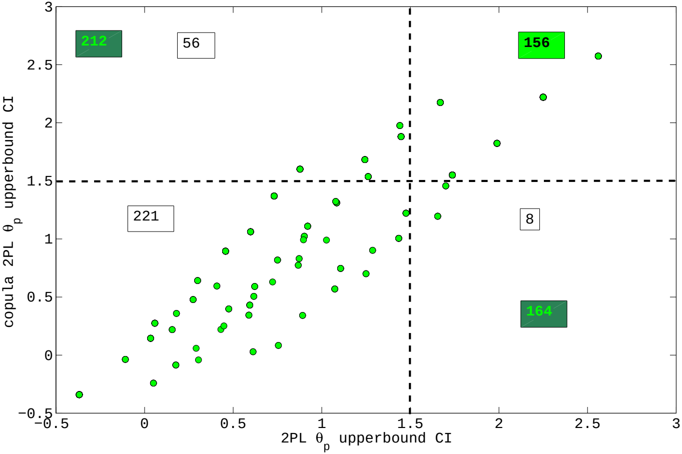
Ignoring vs Modelling

Information about θ_p is artifically large. [more information $\Rightarrow$ smaller estimation error bounds]



Ignoring vs Modelling

Entrance exam: θ_p+upper error boundary ?¿=¡? level



That's all folks !