Exploratory Factor Analysis and Canonical Correlation

Please download:
- SAQ.sav

3 Dec 2010
CPSY 501
Dr. Sean Ho
Trinity Western University

Outline for today

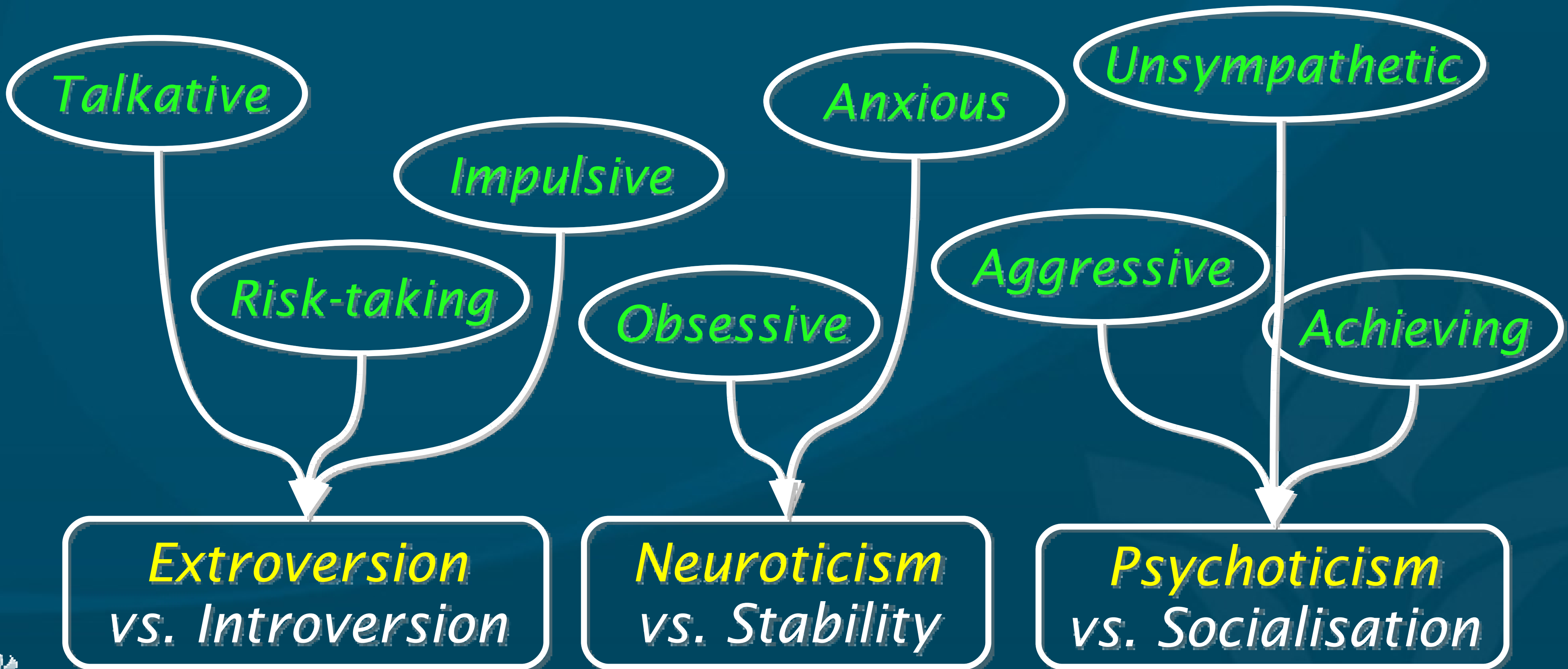
- Factor analysis
  - Latent variables
  - Correlation ("R") matrix
  - Factor extraction
  - Rotation
  - Assumptions
  - Example in SPSS
- Canonical correlation … and how it explains all!
- Structural Equation Modeling
  - Path analysis + factor analysis

Latent variables

- Often we have many, many variables (100s!) and don't know (from theory) which are important
  - e.g., Rosenberg Self-Esteem Scale (SES): rate 10 statements, each on 4-point Likert scale
- Latent variables: underlying quantity to measure
  - Often intangible, e.g., self-esteem
  - Latent vars are estimated via observed vars
- Variable selection: eliminate vars that don't help
  - → simplify model (parsimony)
- Orthogonalization: reduce multi-collinearity by combining highly correlated vars

Latent vars: Eysenck's factors

- Eysenck Personality Questionnaire (EPQ–R, 1985):
  - 100 yes/no questions (observed)
  - 3 underlying (latent) dimensions of personality



Linear factor analysis

- Define factors as linear combos of observed vars:
  - $F_1 = b_1X_1 + b_2X_2 + \ldots + b_nX_n$
  - b_i are the loadings of F_1 on the observed X_i
  - If X_i does not contribute to F_1, then $b_i = 0$
- Exploratory Factor Analysis:
  - No prior theory: find the factors from the vars
  - Sample = pop: results may overfit to data
- Confirmatory Factor Analysis:
  - Already know the factors (from EFA, or theory)
  - Test their predictive power on larger dataset

Why use EFA?

- Parsimony: eliminate redundant vars and those with little variation (reduce 100's → 10's of vars)
  - c.f. multiple regression “backward selection”: drop non-significant IVs from model (for DV)
    - But in EFA there is no DV, just a collection of IVs
  - Combining similar/redundant variables may be useful for enhanced reliability
  - Simplify variables for subsequent analysis
- Discover latent variables, construct new theory:
  - Group together highly inter-correlated variables
  - Interpretability of factors is key!

Outline for today

- Factor analysis
  - Latent variables
  - Correlation ("R") matrix
  - Factor extraction
  - Rotation
  - Assumptions
  - Example in SPSS
- Canonical correlation … and how it explains all!
- Structural Equation Modeling
  - Path analysis + factor analysis

Factors in the R matrix

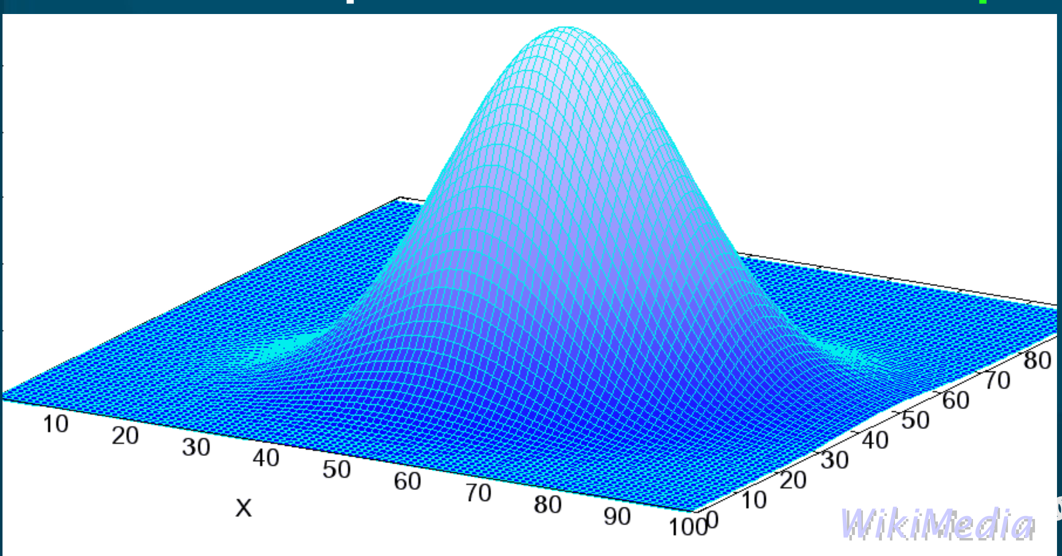
- The "R matrix" is the matrix of correlations between all pairs of (observed) variables
  - Symmetric, and all entries on diagonal are 1
- Groups of highly-correlated variables are good candidates for factors
  - Combine several variables into one factor
  - Internal consistency: Cronbach's α

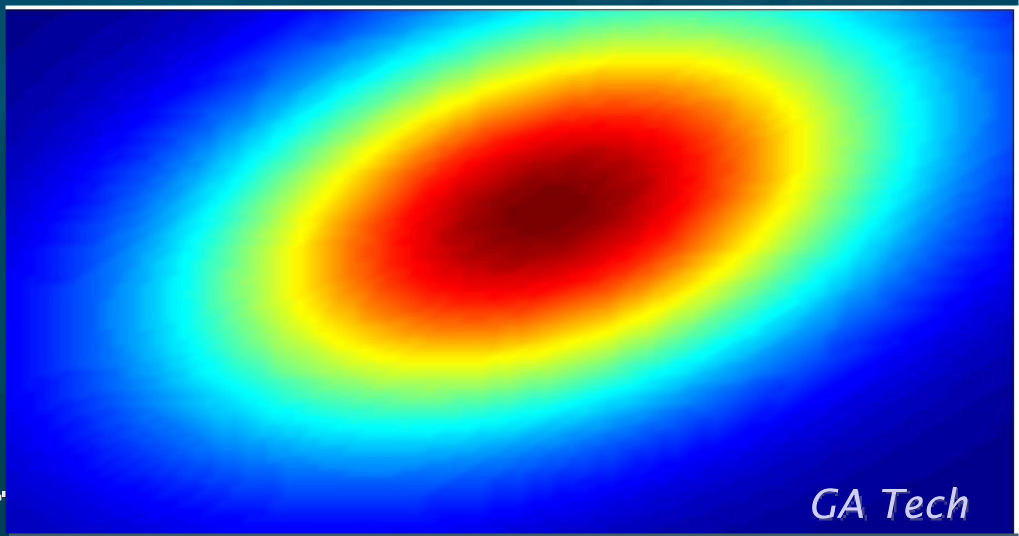
Potential factor

1	0.20	0.11	0.15
0.20	1	0.85	0.91
0.11	0.85	1	0.93
0.15	0.91	0.93	1

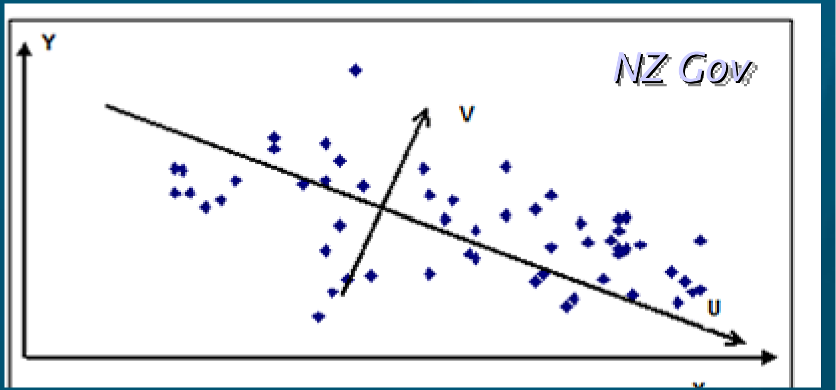
Multivariate normal distribution

- (Pearson) correlation assumes normality
  - Although EFA is fairly robust to non-normality
- For multiple vars: multivariate normal distribution
  - Quartiles, etc. form ellipsoids around mean
  - Area of ellipsoids $\leftrightarrow$ determinant of R matrix
- Multi-collinearity: some vars nearly perfect correl.
  - Ellipsoids become squashed; det(R) $\rightarrow$ 0





Factor extraction



- Each factor is an axis through the data
- PCA extracts one factor at a time:
  - PC1 is the longest axis of the ellipsoid: capture as much variance in vars as possible
  - PC2 is longest axis perpendicular to PC1: capture as much leftover variance as possible
  - Etc.: capture most of variance w/just a few PCs
- PAF is similar but only cares about the portion of variance which is shared with other variables
  - Uses R matrix but diagonals are %shared var
  - Useful for exploration, development of theory

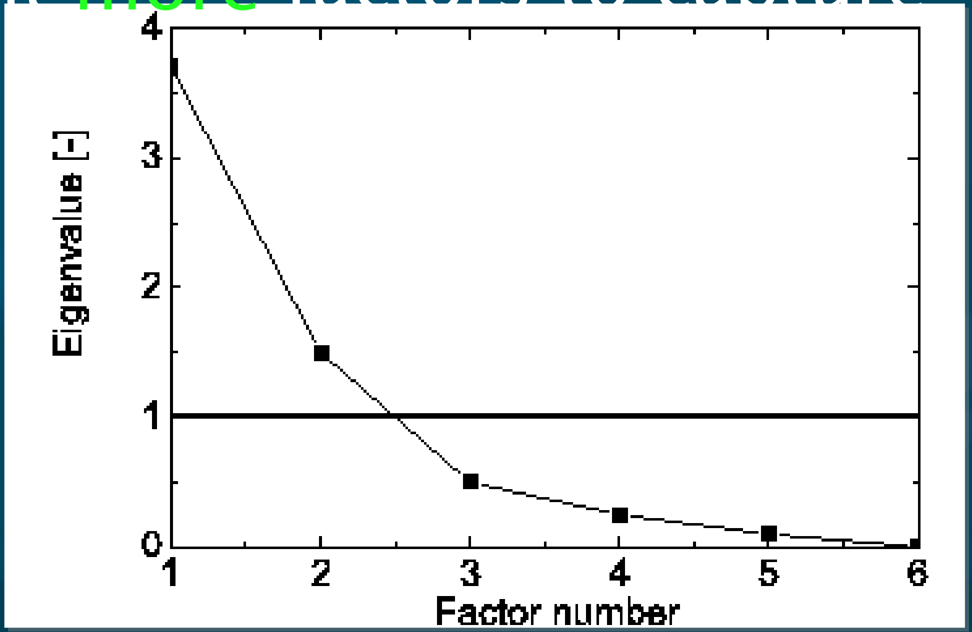
Communality and eigenvalues

	F_1	F_2
X_1	b_{11}	b_{21}
X_2	b_{12}	b_{22}
…	…	…
X_n	b_{1n}	b_{2n}

- Linear combinations of variables:
  - $F_1 = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1n}X_n$
  - $F_2 = b_{21}X_1 + b_{22}X_2 + \ldots + b_{2n}X_n$
- Factor matrix lists the loadings
- Communality of a variable is
  % of its variance explained by the selected factors
  - Sum of squared loadings across the factors
- Eigenvalue of a factor is
  % of total variance explained by that factor
  - Sum of squared loadings across all variables
  - PCs extracted in descending order by eigenval

How many factors to use?

- Guided by theory: # underlying dimensions
  - Can you understand/interpret each factor?
- Kaiser's criterion: eigvals < 1 → unstable factors
- Not adding appreciably to total explained variance
- If a (theoretically) important variable has low communality, may want more factors to capture its variance
- Scree plot: eigenvalues of successive factors



Rotation

- Factors as extracted using PCA or PAF are often difficult to interpret
  - Each variable may show up in several factors
- Rotation can shuffle the extracted factors to ease interpretation
- Several methods of rotation. The most common:
  - Varimax (orthogonal): minimise #factors each var shows up in ("Are you in or out?")
  - Oblimin (oblique): increases eigenvalues by allowing factors to be a bit correlated (use δ)
    - → in psychology, latent vars are rarely uncorrelated!

Outline for today

- Factor analysis
  - Latent variables
  - Correlation (“R”) matrix
  - Factor extraction
  - Rotation
  - Assumptions
  - Example in SPSS
- Canonical correlation … and how it explains all!
- Structural Equation Modeling
  - Path analysis + factor analysis

Assumptions for EFA

- Sufficient sample size
  - 300 or more; rule of thumb is ~20 per var
  - Kaiser-Meyer-Olkin (r^2 / partial-r^2) close to 1
  - Diagonals of anti-image correlation matrix >0.5 (drop any vars <0.5) (partial correl)
- Linearity of relationships amongst vars (scatter)
  - Normality is good, too (but not critical)
- Good correlations in R matrix (many >0.5)
  - Bartlett's test of sphericity (tests for goodness): if R matrix differs significantly from identity, i.e., off-diagonal entries not all zero

EFA in SPSS: SAQ.sav

- Dataset: SAQ.sav (artificial data from textbook)
  - 23 questions on "SPSS anxiety" (5pt-Likert)
- EFA: Analyze → Dimension Reduction → Factor
  - Variables: select all 23 questions
- Descriptives: Correlation Matrix:
  - Coeffs, Det., Anti-image, KMO+Bartlett
- Extraction: Method → Principal Components
  - Display: Unrotated, Scree; Extract: Eigvals >1
- Rotation: Varimax; Display: Rotated, Loading plot
- Options: Sort, Suppress <0.4

Interpreting output: SAQ.sav

- Assumptions: correlation matrix
  - Ensure each var has several $|r| > 0.3$
    - Should really look at significance of correl., too
  - Determinant not close to 0? (multi-collinearity)
- Diagonal entries of anti-image: >0.5?
- KMO close to 1?
- Bartlett's test? (sig is good)

Interpreting output: factors

- How many factors were selected?
  - Total Variance Explained
  - Scree plot
- Component Matrix: shows unrotated loadings
- Rotated Component Matrix:
  - Each var should show up in only 1 or 2 factors
  - We asked for cut-off at 0.4
  - Factors indicate groupings of vars and relative importance of each var within a grouping
- Try to interpret the factors? Fear of computers, fear of stats, fear of math, peer evaluation

EFA summary

- A way of reducing the number of variables and finding underlying structure
  - Factors are linear combinations of the variables
- Extraction: one factor (axis) at a time
  - Largest eigenvalues first
  - PCA: total variance, good for variable reduction
  - PAF: shared variance, good for theory explor.
- Rotation: to ease interpretation of the factors
  - Varimax: orthogonal, each var in 1–2 factors
  - Oblimin: oblique, may reflect theory better

Factor Analysis: References

- Kristopher J. Preacher & Robert C. MacCallum (2003).
  Repairing Tom Swift's Electric Factor Analysis Machine.
  Understanding Statistics, 2(1), 13–43.
- Anna B. Costello & Jason W. Osborne (2005).
  Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis.
  Practical Assessment Research & Evaluation, 10(7), 1–9.

Outline for today

- Factor analysis
  - Latent variables
  - Correlation (“R”) matrix
  - Factor extraction
  - Rotation
  - Assumptions
  - Example in SPSS
- Canonical correlation … and how it explains all!
- Structural Equation Modeling
  - Path analysis + factor analysis

Canonical correlation

- Hotelling, 1936: an idea to help understand many analyses: multiple regression, MANOVA, EFA, etc.
- Two groups of variables X_i and Y_j (not necc. IV/DV)
- Let U, V be linear combinations of the X_i and Y_j:
  - $U = b_1X_1 + b_2X_2 + \ldots + b_mX_m$
  - $V = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n$
- The first canonical pair is the (U_1,V_1) which have max correlation; this corr. is 1st canonical corr.

$$X_1, X_2, \ldots, X_m \longrightarrow U_1 \leftrightarrow V_1 \longleftarrow Y_1, Y_2, \ldots, Y_n$$

Canonical corr. and regression

- Second canonical pair is the (U_2,V_2) which are orthogonal to (U_1,V_1) and maximize correlation
  - 3rd canonical pair, 3rd canonical correlation, etc.
  - # canonical pairs = min(#X, #Y) = min(m, n)
- Multiple regression is a special case:
  - Only one Y → only one canonical pair (U,V)
  - The canonical correlation is R (→ R^2!)
  - Coefficients of U are the b's (slopes)!

Canonical corr. and MANOVA

- MANOVA: multiple DVs → multiple canon. corrs.
  - Each V_k is a linear combo of the DVs:
    → they are factors of the DVs!
  - Different from EFA, in that the factors are chosen to maximize correlation of Xs with Ys
  - Optimal factor of the IVs, plus
    optimal factor of the DVs → to maximize correl.
- Canonical factor loadings: correlation between a variable (X_i or Y_j) and one of its factors (U_k or V_k)
  - These are equivalent to factor loadings in EFA

Outline for today

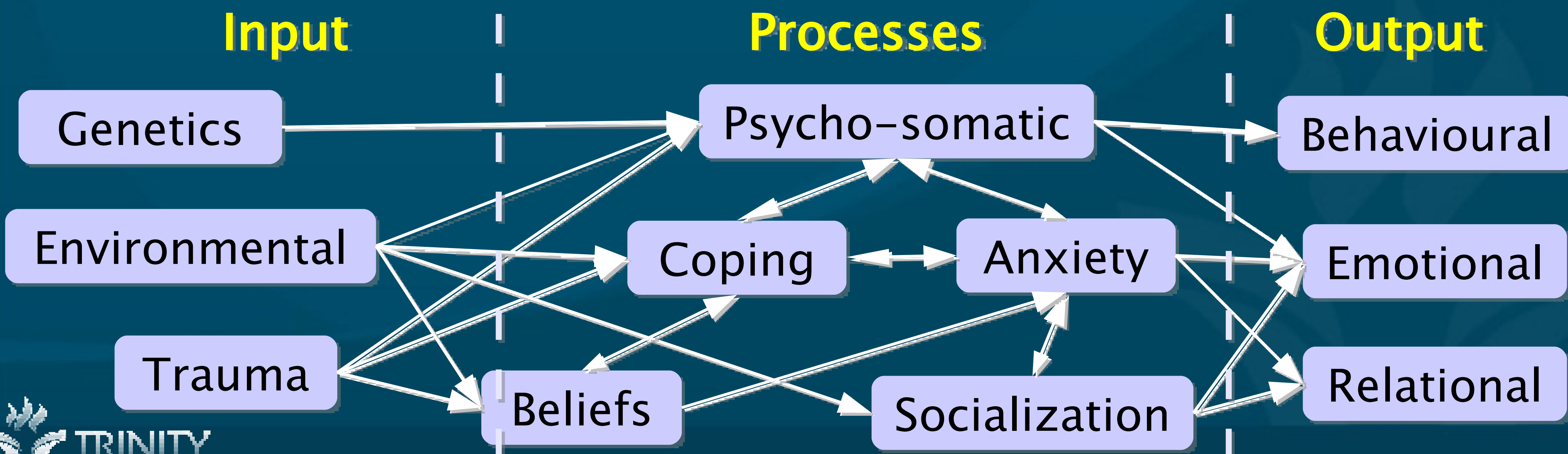
- Factor analysis
  - Latent variables
  - Correlation (“R”) matrix
  - Factor extraction
  - Rotation
  - Assumptions
  - Example in SPSS
- Canonical correlation … and how it explains all!
- Structural Equation Modeling
  - Path analysis + factor analysis

Structural Equation Modelling

- A way to do regression on latent variables
- SEM = Path analysis + Factor analysis
  - Structural model + Measurement model
- Structural model: causal links between variables
  - Vars may serve as both IV and DV (simultaneous equation analysis: econometrics)
- Measurement model:
  - Latent variables underlying the observed vars
- SEM may be exploratory: don't know structure, don't know latent vars (EFA)
- Or confirmatory: theory tells structure, latent vars

SEM: Path analysis

- Path diagrams indicate causation amongst vars
  - E.g., mediation diagram
- Often group vars as: Inputs → Processes → Output
- May be built from theory (confirmatory analysis)
  - Or derived from correlations in the data (exploratory analysis) (might not generalize)



SEM: Simultaneous equations

- Originally developed for econometrics theory
- Exogenous vars serve only as IVs (inputs)
- Endogenous vars are DV or both (process, output)
  - $Y_1 = a_{11}Y_1 + a_{12}Y_2 + b_{10} + b_{11}X_1 + b_{12}X_2 + \ldots$
  - $Y_2 = a_{21}Y_1 + a_{22}Y_2 + b_{20} + b_{21}X_1 + b_{22}X_2 + \ldots$
- In matrix form: $Y = \alpha + BY + \Gamma X$ (plus resids)
  - Y: vector of endogenous variables
  - X: vector of exogenous variables
  - α: vector of intercepts
  - B: matrix of endogenous inter-relationships
  - Γ: matrix relating exogenous → endogenous

SEM: putting it together

- SEM = Path analysis + Factor analysis
- The vars X, Y in the simultaneous equations need not be observed vars: can be latent vars (factors)
- Measurement model maps observed vars onto latent vars via factor loadings (either EFA or CFA)
- Model estimation step:
  - Build structural model (path diagram) w/latents
  - Build measurement model from observed vars
- Model testing step (goodness of fit): χ^2, AIC, etc.
- Repeat to refine model

SEM: steps for modelling