# Modeling Composite-Based Populations Using Composite-Based Methods

2017 Modern Modeling Methods Conference

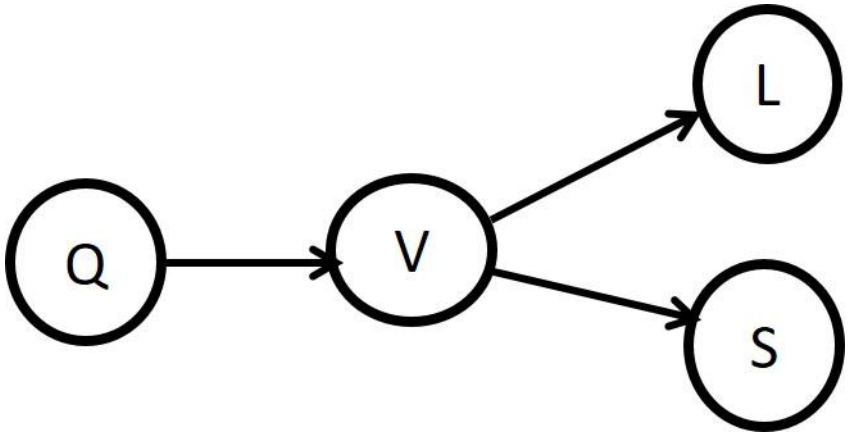
J.-M. Becker[a], E. Rigdon[b], A. Rai[b]

[a]University of Cologne, [b]Georgia State University

Contact: erigdon@gsu.edu

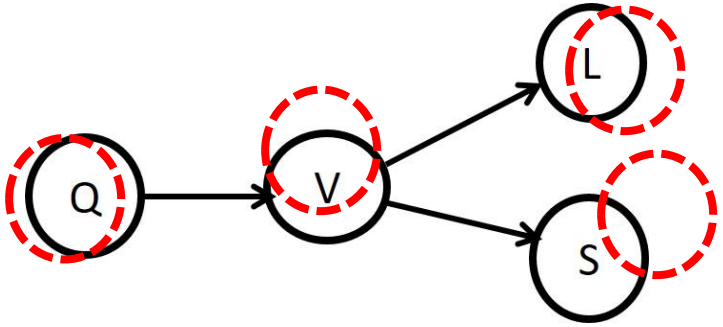
Disclosure: J.-M. Becker is a principal author of *SmartPLS*, a program for estimating PLS path models

# A conceptual model



# Representing unobservables

- Statistical methods work on data--must represent conceptual variables empirically
- Factor-based SEM uses common factors to represent conceptual variables
- Composite-based methods use weighted composites for the same purpose



# History

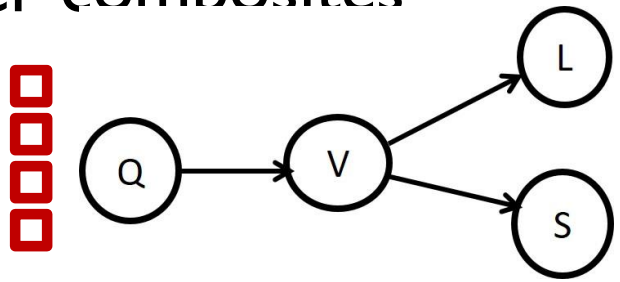
- Composite-based methods were first developed to approximate results from factor analysis, at lower cost
- Simulations show biased results when using composite methods to estimate factor model parameters*
- All such studies used data drawn from populations where a factor model, not a composite model, was correct

# Composite-Based Methods

- Partial least squares (PLS) path modeling
- Generalized structured component analysis (GSCA)
- Regularized generalized canonical correlation analysis (RGCCA)
- Principal components (PCA) + regression
- Unit weights
- Etc …

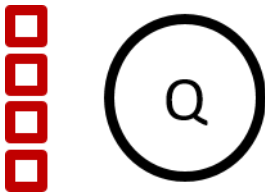
# PLS path modeling

- Each composite estimated alternately:
  - Weighted sum of its indicators
  - Weighted sum of other composites




- In the end, always the former
- End result is a super correlation matrix of indicators and composites, from which regression coefficients are estimated

# Mode A vs Mode B




- B: Regress composite on all components simultaneously
- A: Regress each component on composite separately (long mistaken for "something like" factor analysis)
- Mode B implies regression weights while Mode A implies correlation weights
- In PLS, must choose 1 for each composite

# GSCA

- Can estimate just weights ("formative") or weights and loadings ("reflective") for each composite
- Subsets of model parameters estimated in turns, using alternating least squares
- But with each step minimizing the same overall criterion
- Enables constraints on model parameters

# RGCCA

- Springs from the "multi-block data analysis" literature
- Enables Mode B, a variant of Mode A, and an in-between "Mode Ridge"

# Received view

- Composite based methods like PLS path modeling are inherently defective, yielding biased parameter estimates
- You might as well create composites with unit weights, or use regression following principal component analysis (PCA)

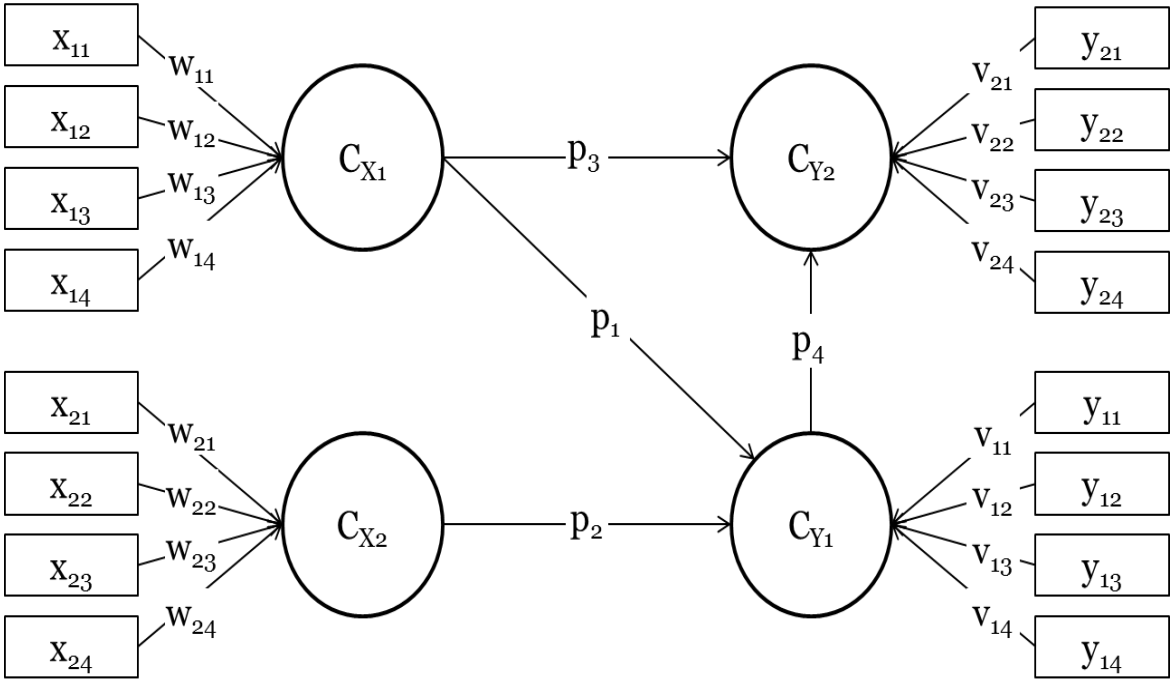
# Questions

- Given a correct model and composite population, are structural estimates from PLS / GSCA / RGCCA consistent?
- Are they "any better" than estimates from simpler techniques like unit weights?
- Does Mode A yield an advantage in out-of-sample R2 (and if so, at what cost)?

# Procedure

- Generate 10,000 observations from a composite-based population defined by parameter values
- Select sample of size n (no replacement) and estimate model
- Fix parameters at estimated values and predict dependent variables
- Repeat

# Simulation model



# Defining composite populations

$$C_X = WX, C_Y = VY$$

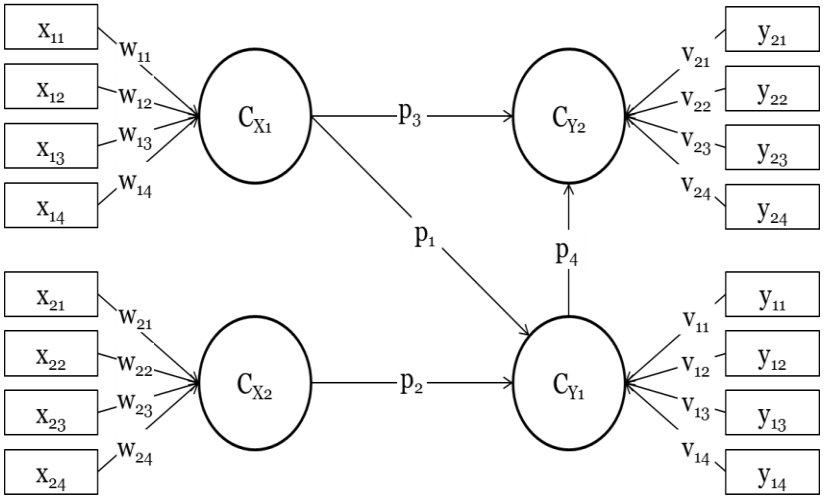
$$R_{XX} = R_{YY} = \begin{bmatrix} 1 & k & 0 & 0 \\ k & 1 & 0 & 0 \\ 0 & 0 & 1 & k \\ 0 & 0 & k & 1 \end{bmatrix}$$

# Defining composite populations

$$C_Y = C_X P + E$$

$$p_1 = 0.4$$

$$p_2 = \sqrt{R^2 - p_1^2}$$



# Defining composite populations

Defined cross-covariances (X,Y) using path model equations:

$$\Sigma_{XY} = \Sigma_{YY} V^{*T} P W^{*} \Sigma_{XX}$$

though there are other ways to do this, since you are working with composites

# Design dimensions

- Sample size: 40, 100, 500
- Indicator correlations: .00 to .95 by .05
- Population $R^2$ (.2, .3, .4, .5, .6, .7, .8)

Did NOT vary

- Unstandardized component weights .7, .6, .3, ~.25 (so composite variance = 1)
- Number of components: 4

# Simulations

1,000 replications x 7 x 20 x 3 = 420,000 for each method examined:

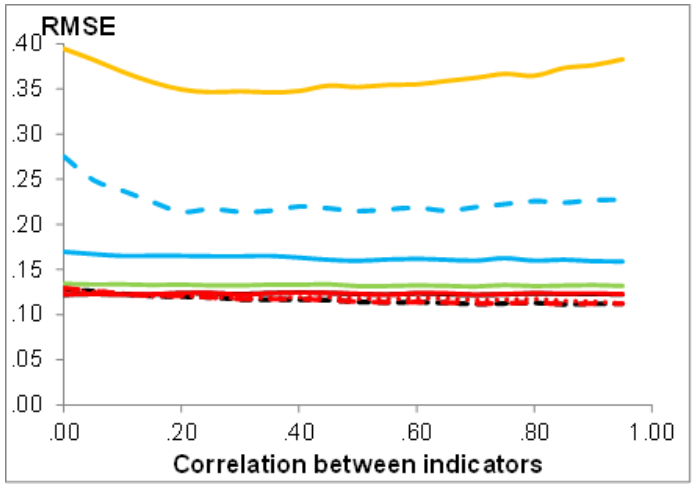
PLS Mode A, Mode B
GSCA "formative," "reflective"
RGCCA Mode B, new Mode A, Mode Ridge
Unit weights
PCA + regression

# Criteria

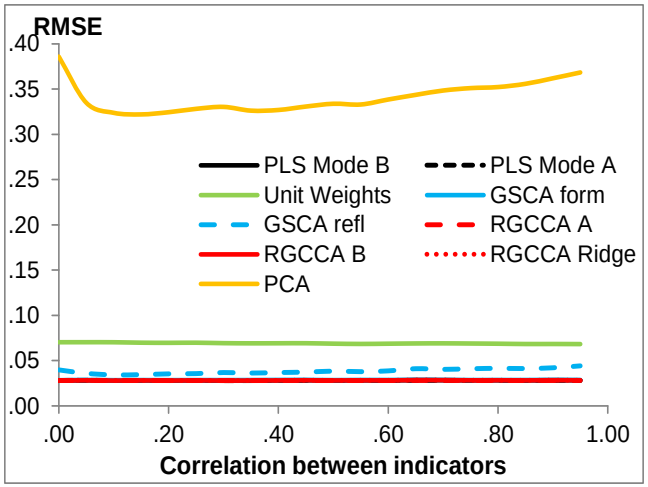
- RMSE
- Bias
- In-sample $R^2$
- Out-of-sample $R^2$

# Results

# RMSE $P_1$, $P_2$, $P_3$, $P_4$

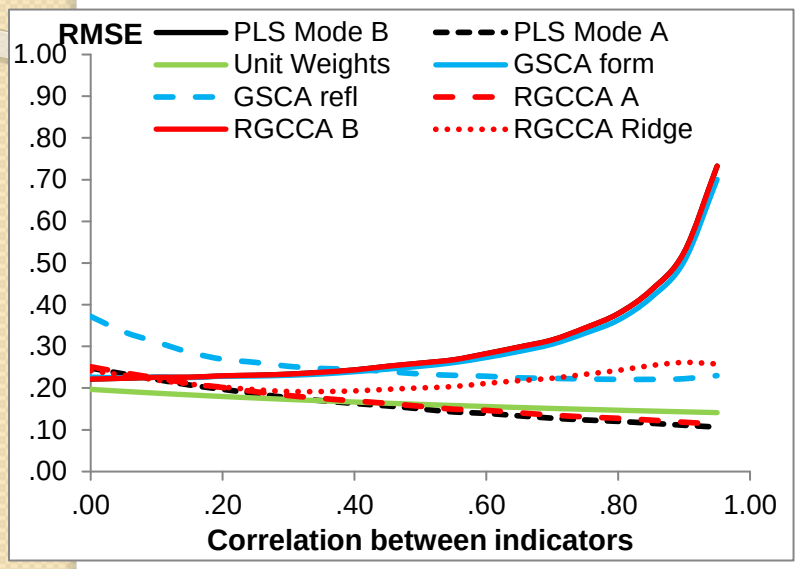



n = 40

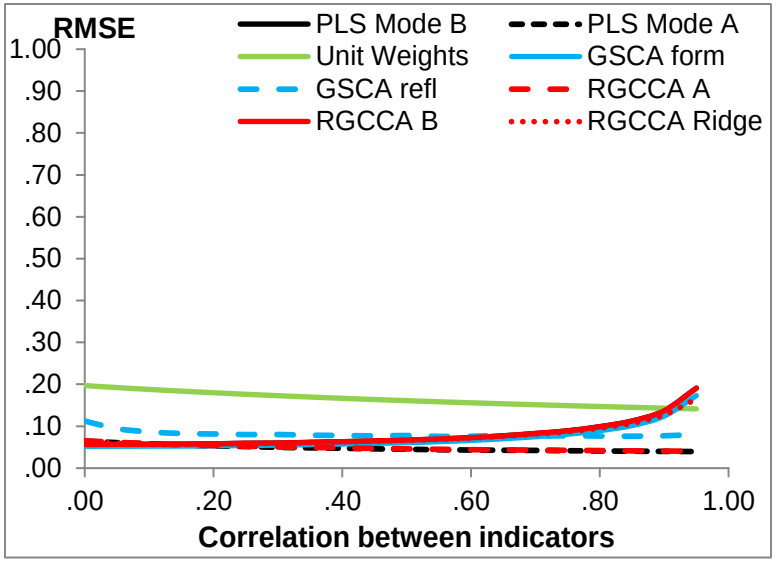



n = 500

# RMSE estimated weights

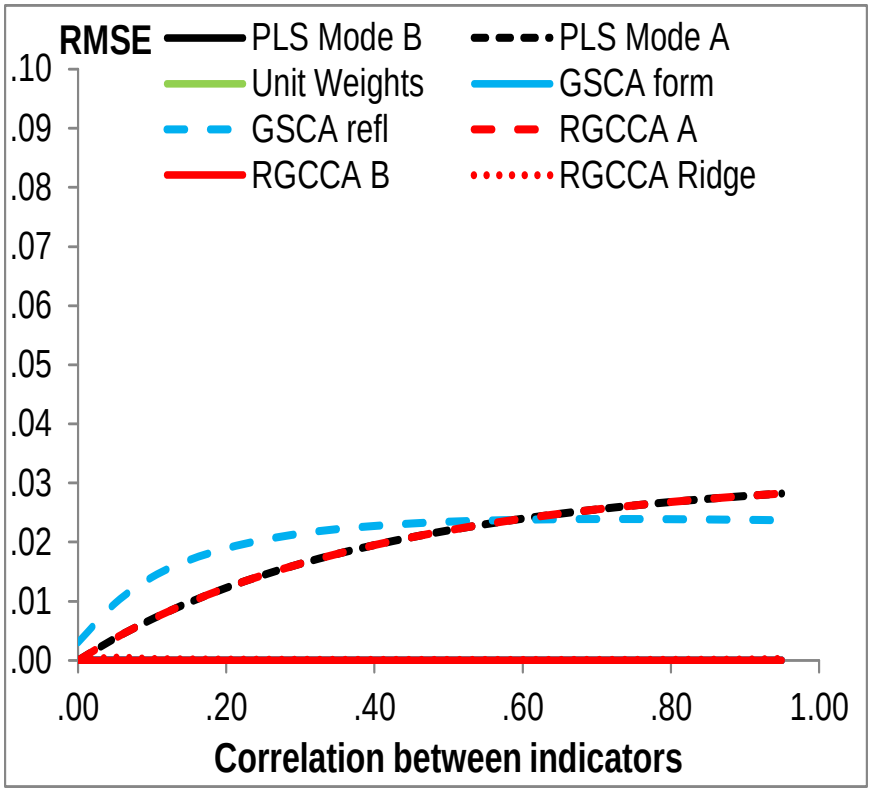



n = 40

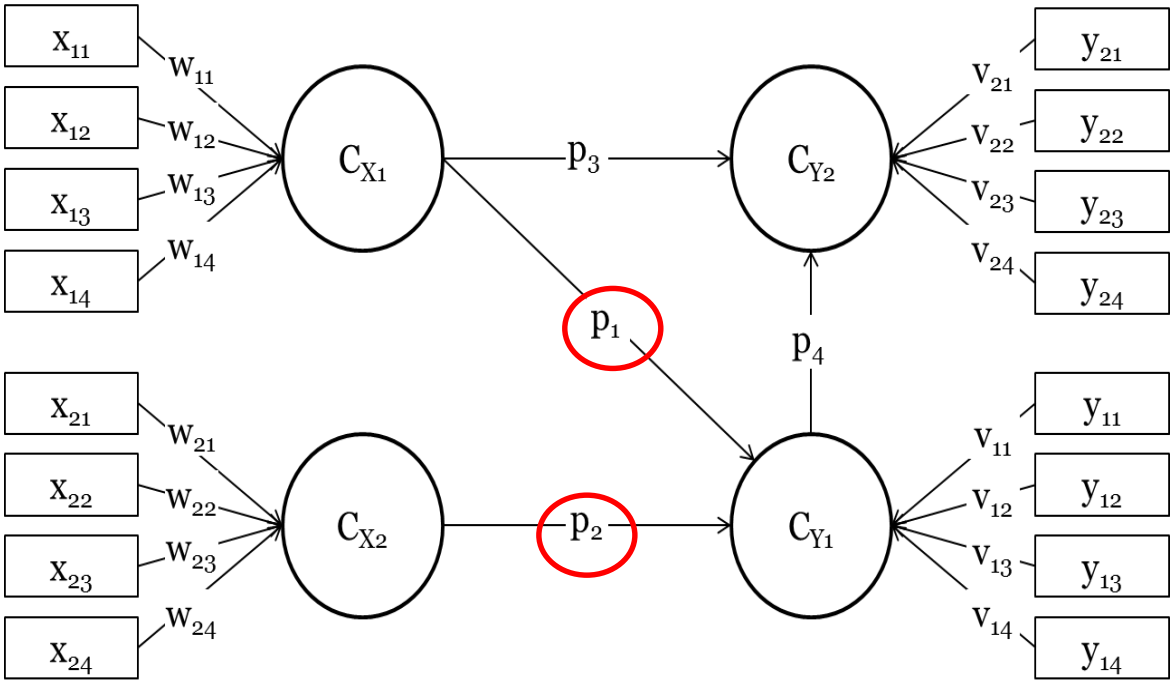



n = 500

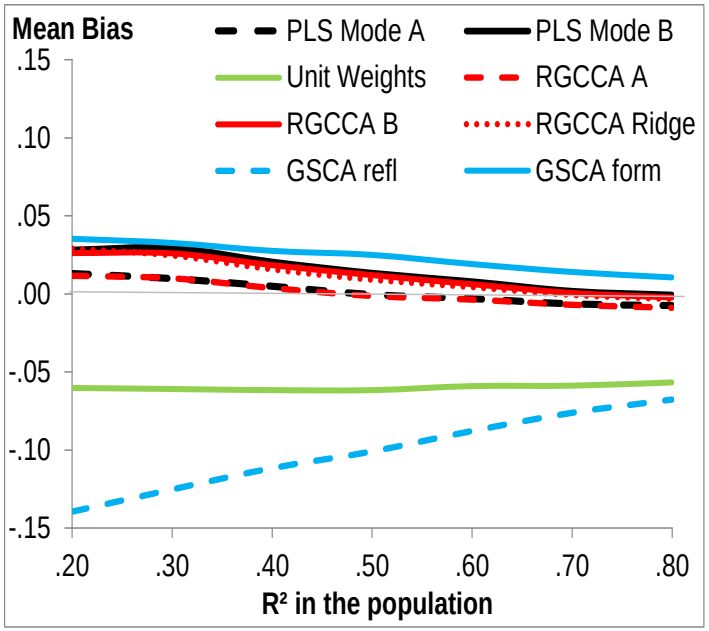
# RMSE for weights when n = 10,000



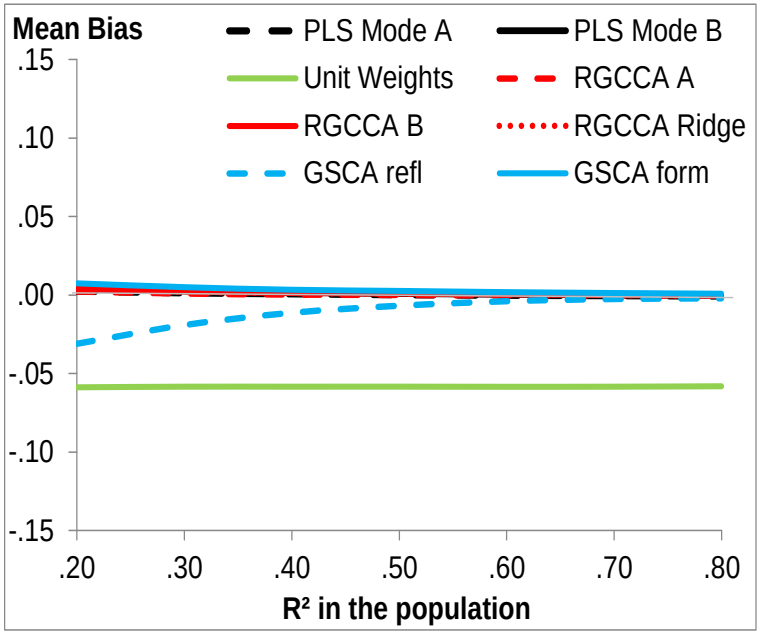
# Reminder: simulation model



# Path estimate bias: $p_1$

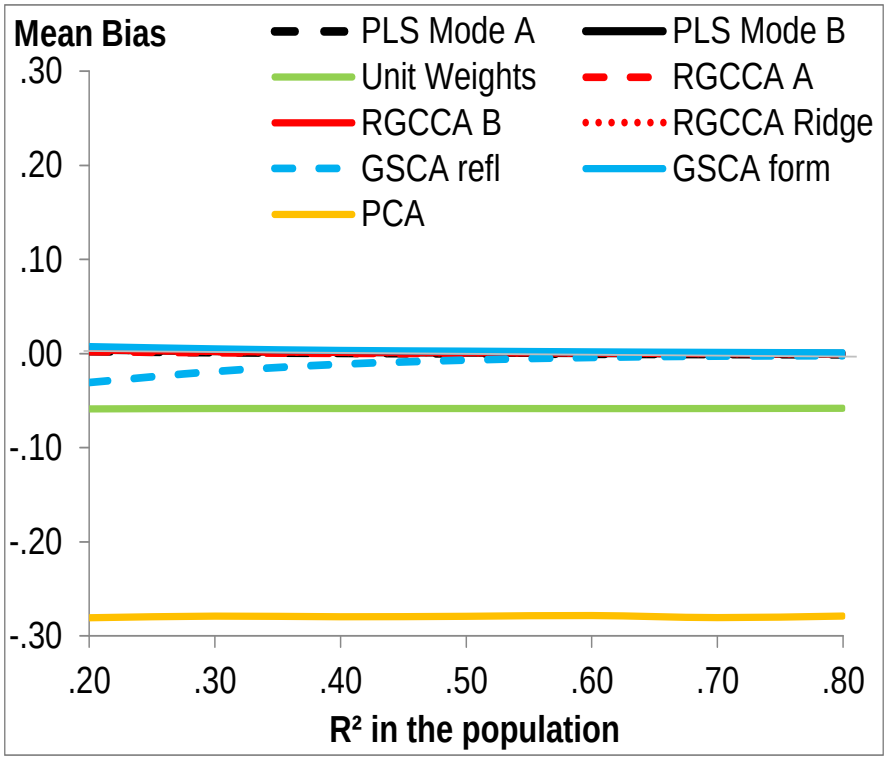



n = 40



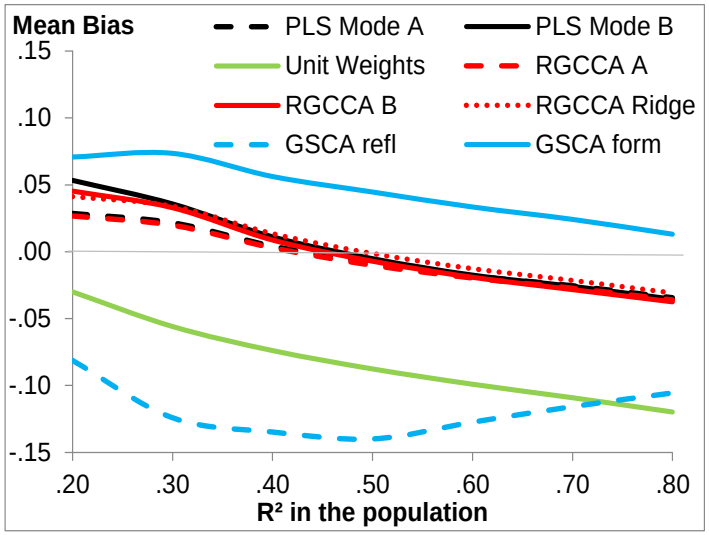

n = 500

# Bias in $p_1$ including PCA

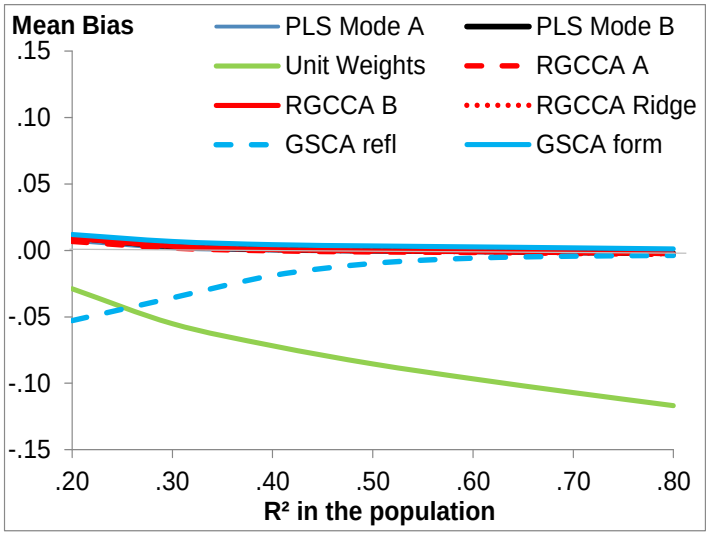



n = 500

# Path estimate bias: $p_2$

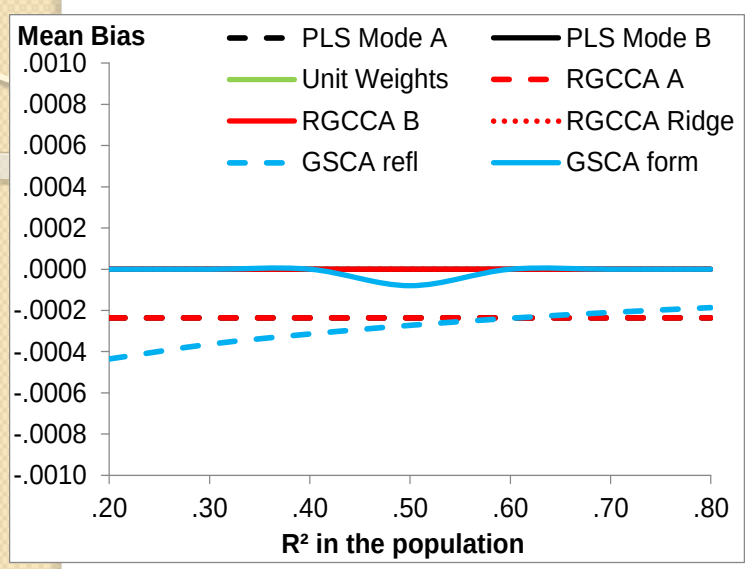


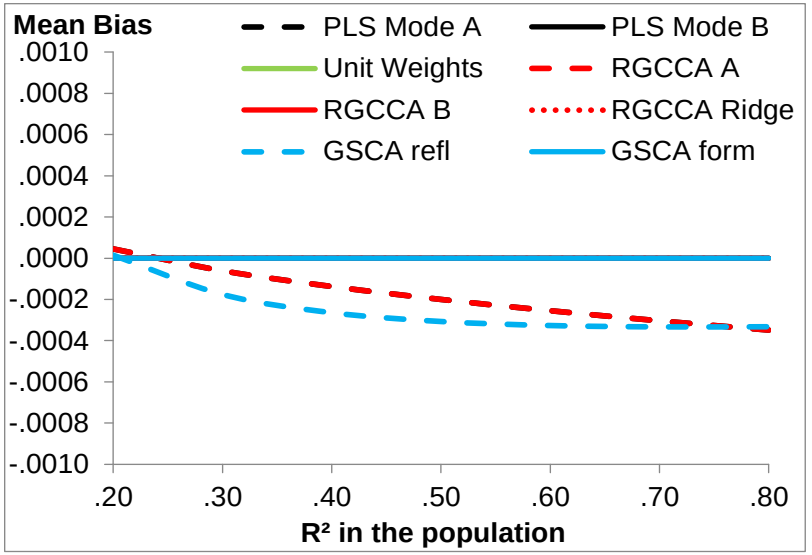

n = 40




n = 500

# Path estimate bias at n = 10,000

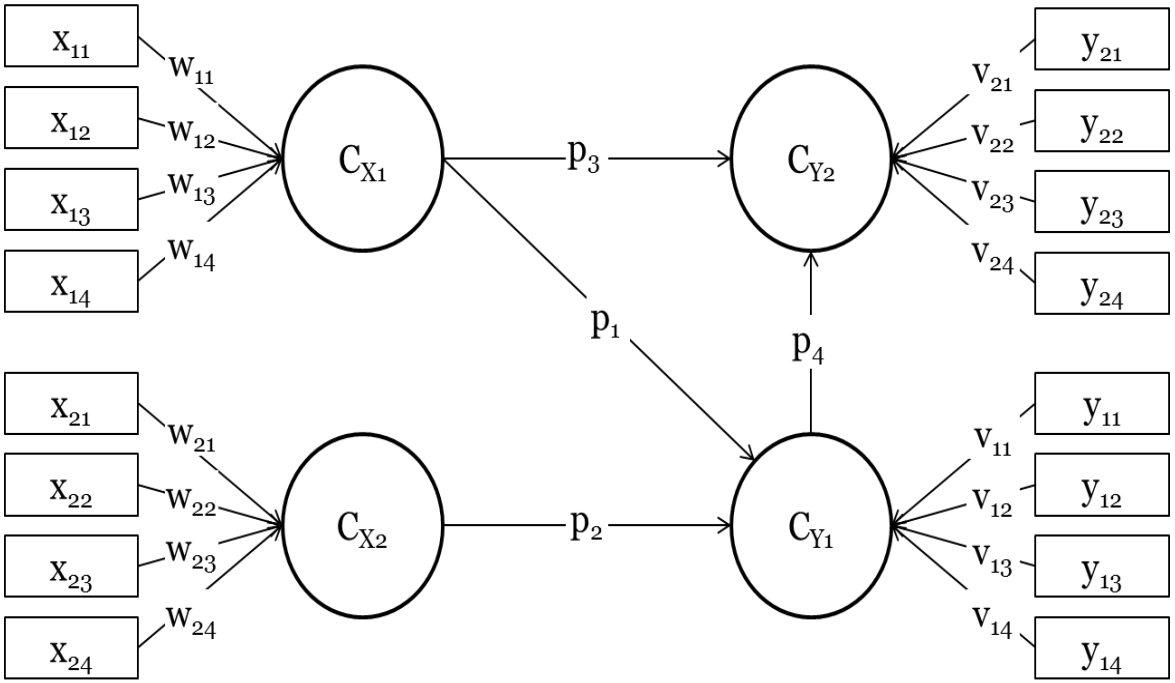


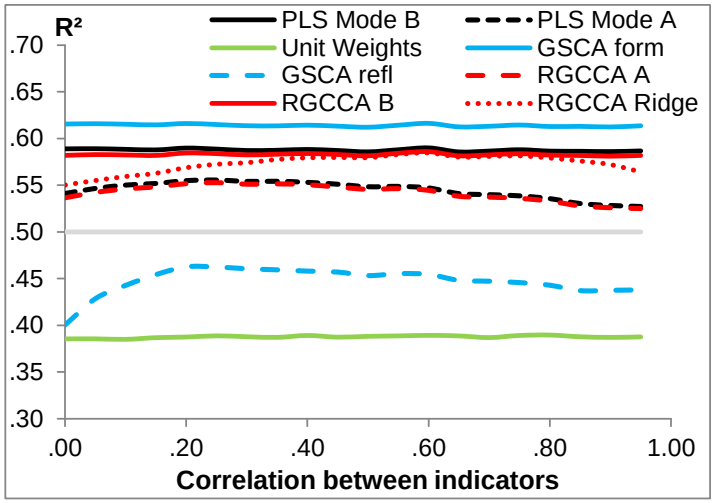

$P_1$

$P_2$

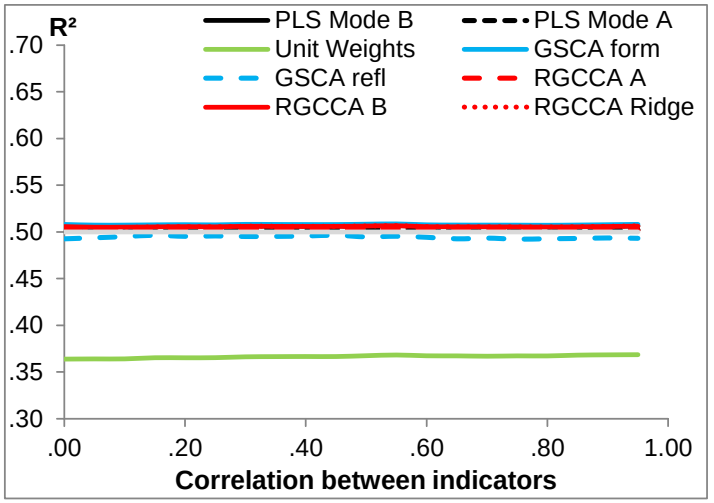
# Reminder: simulation model



# In-sample $R^2$ $C_1$

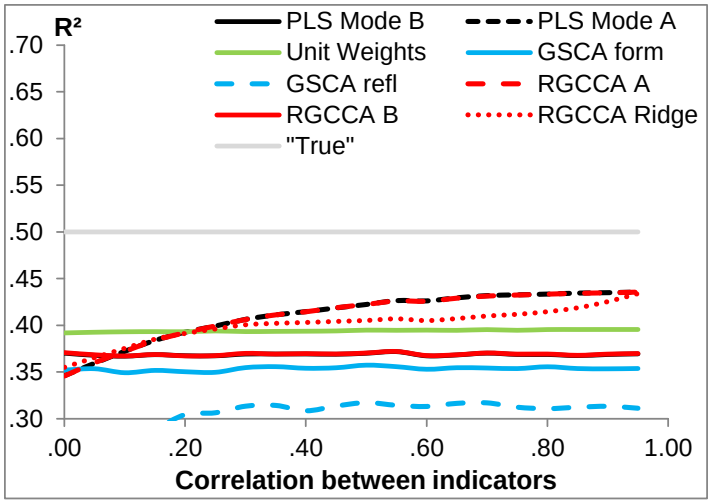



n = 40

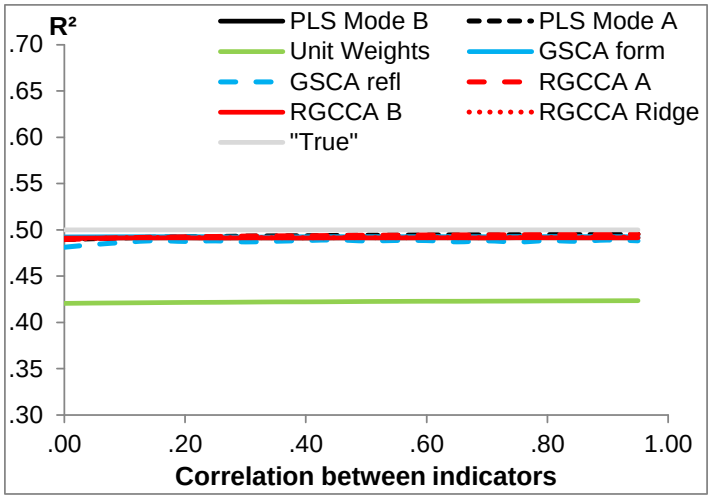



n = 500

# Out-of-sample $R^2$ $C_1$




n = 40

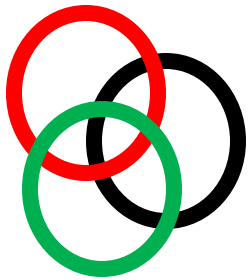



n = 500

# Conclusions / takeaways I

- PCA + regression is a poor choice
- Simple unit weights may outperform weighted methods when n is small
- Mode B (PLS / RGCCA) yields consistent estimates within the context of correctly specified composite models
- Mode A results may be preferable at moderate n, with high item covariance, and if the goal is max out-of-sample $R^2$

# Questions?



# Selected references


- Dana & Dawes (2004). The superiority of simple alternatives to regression for social science predictions. Journal of Educational and Behavioral Statistics, 29, 317-331.
- Dijkstra & Henseler, J. (2015). Consistent partial least squares path modeling. MIS Quarterly, 39, 297-316.
- Haig & Evers (2016). Realist inquiry in social science. London: Sage.
- Hwang & Takane (2015). Generalized structured component analysis: A component-based approach to structural equation modeling. Boca Raton, FL: CRC Press / Chapman & Hall.
- Rigdon (2012). Rethinking partial least squares path modeling: In praise of simple methods. Long Range Planning, 45(5/6), 341-358.
- Rozeboom (1979). Sensitivity of a linear composite of predictor items to differential item weighting. *Psychometrika,* 44, 289-296.
- Tenenhaus & Tenenhaus (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76, 257-284.