Solving the 'many variables' problem in MICE with supervised principal component regression

In one sentence

Using **supervised principal component regression** as a univariate **imputation** model in **MICE** is a great way to solve the **many-variables** imputation problem.

Large data with missing values (-)

	x_1	x_2	x_3	x_4	...	w_{141}	w_{142}	w_{143}	w_{144}	...	$z_{(p-3)}$	$z_{(p-2)}$	$z_{(p-1)}$	z_p
Esther	3	-	4	6		7	6	2	2		5	4	9	8
Anton	-	-	3	1		8	3	7	10		8	10	3	7
Leonie	-	7	-	4		5	9	3	6		9	10	9	2
Joran	1	4	4	-		9	1	5	5		3	1	9	8
...														
Mihai	-	8	-	4		10	6	2	9		2	5	2	10

Expert imputation model specification

- Remove constants and collinear variables.
- Evaluate connection between variables in the data.
- Apply a correlation-threshold selection.
- Extra: use total scores for item scales.
- Extra: use single measurement in longitudinal data.

VS

Automatic imputation model specification

- MICE with Principal component regression (MI-PCR)
- MICE with Association-threshold supervised principal component regression (MI-SPCR)
- MICE with Principal covariates regression (MI-PCovR)
- MICE with Partial least square (MI-PLSR)

Percent relative bias

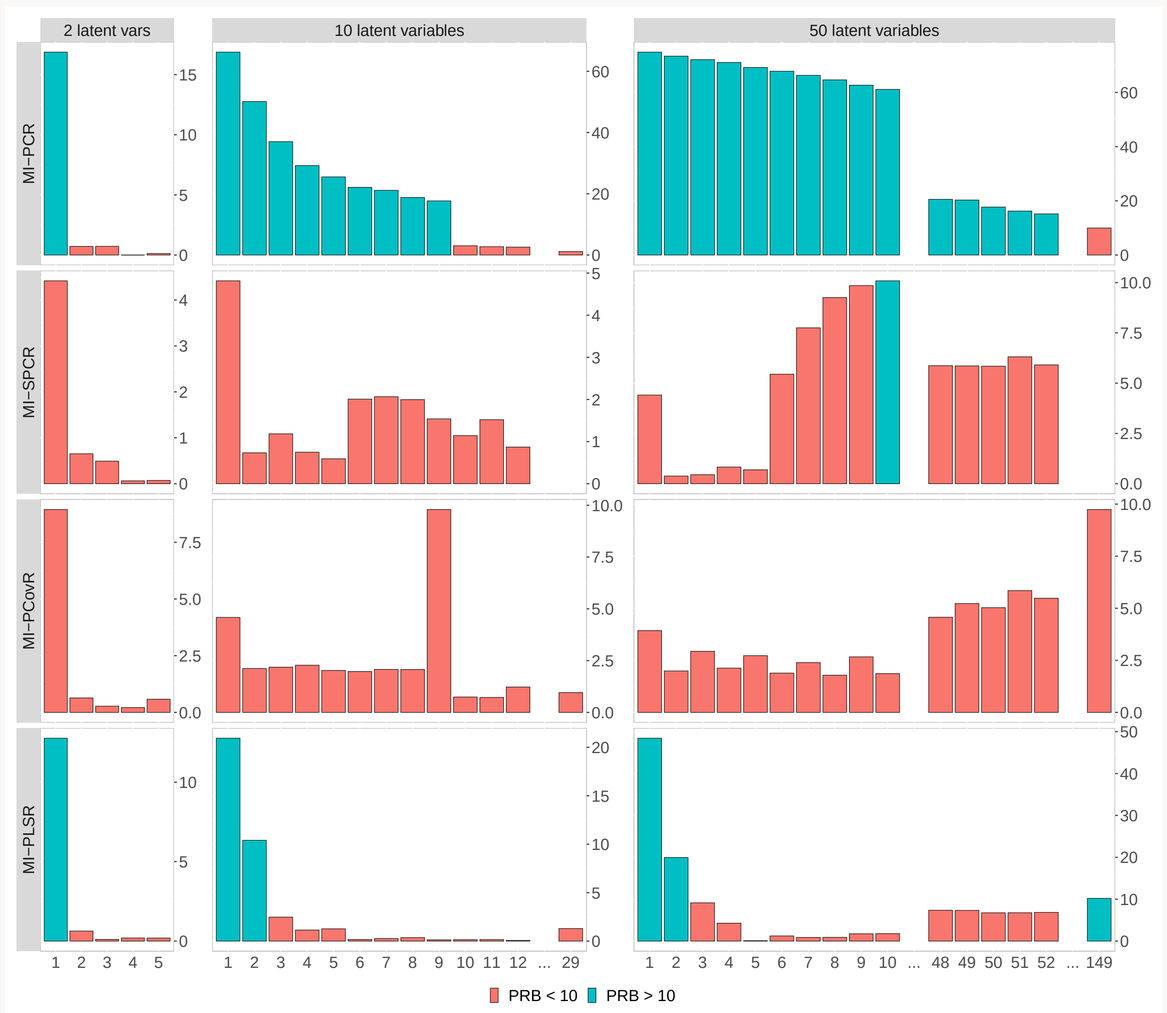

Figure: The percent relative bias (Y-axis) for the correlation coefficient between x_1 and x_2, obtained after imputing the missing values with the four PCR-based imputation methods (grid rows), is reported as a function of the number of components used (X-axis).

Confidence interval coverage

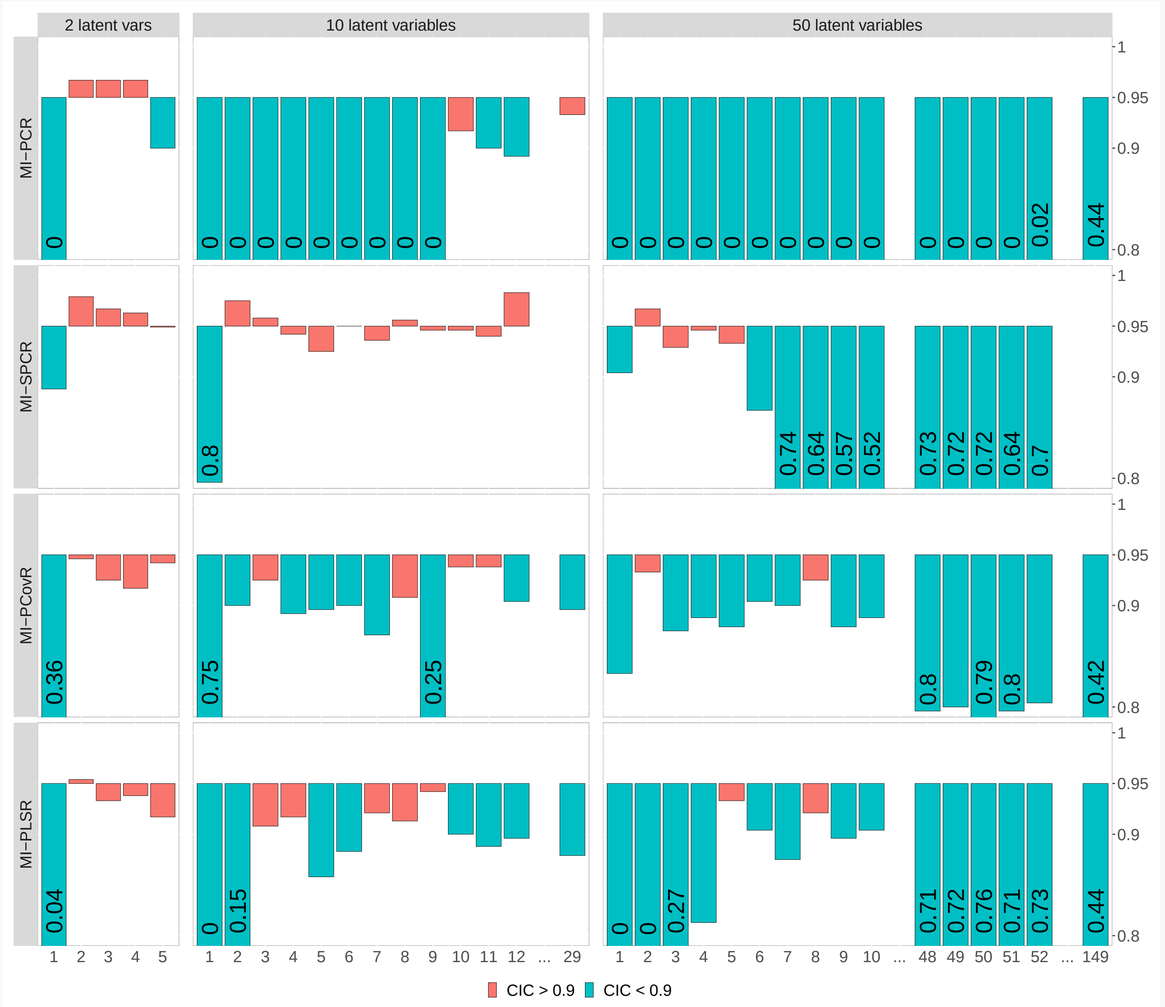

Figure: The confidence interval coverage for the correlation coefficient between x_1 and x_2, obtained after imputing the missing values with the four PCR-based imputation methods (grid rows), is reported as a function of the number of components used (X-axis).

Project summary and code

Play with the Shiny app

More research like this

Edoardo Costantini

e.costantini@tilburgunviersity.com

Tilburg University, Netherlands