Fitting propensity analysis in R

Sonja D. Winter
&
Tive Khumalo

IMPS 2025

Motivation

Falk and Muthukrishna (2023) developed the *ockhamSEM* R package

- Made it easy to run fitting propensity analyses in R using familiar *lavaan* syntax
- Uses the *onion* method for uniformly sampling correlation matrices from the full data space

Several researchers have used this package to examine fitting propensity[1]

But! Analysis design was inconsistent across studies

- ➢ What part of the data space (all of it vs. positive manifold)?
- ➢ How many correlation matrices should be sampled from the data space?
- ➢ What fit indices should we focus on?
- ➢ Should fit indices based on non-converged analyses be considered?
- ➢ Is it fine to rely on default *lavaan* settings?

[1] Bader & Moshagen, 2022; Bonifay et al., 2025

Fitting propensity analysis in R

```r
library(ockhamSEM)
```

Set up temp. correlation matrix for number of variables in model

```r
p <- 3
temp_mat <- diag(p)
colnames(temp_mat) <- rownames(temp_mat) <- paste0("v", seq(1:p))
```

Specify model

```r
mod1 <- '
v3 ~ v1 + v2
v1 ~~ 0*v2
'
```

Fit model to temp. correlation matrix

```r
mod1.fit <- sem(mod1, sample.cov = temp_mat,
                start = "default",
                control = list(iter.max = 150),
                optim.force.converged = F,
                bounds = "none",
                std.lv = T,
                sample.nobs = 1000)
```

Run fitting propensity analysis

```r
res <- run.fitprop(mod1.fit,
                   fit.measure = c("logl", "cfi", "srmr", "rmsea"),
                   rmethod = "onion", reps = 1000000,
                   onlypos = FALSE,
                   seed = 3242)
```

Model estimation choices

Set up temp. correlation matrix for number of variables in model

Specify model

Fit model to temp. correlation matrix

Run fitting propensity analysis

```r
library(ockhamSEM)

p <- 3
temp_mat <- diag(p)
colnames(temp_mat) <- rownames(temp_mat) <- paste0("v", seq(1:p))

mod1 <- '
v3 ~ v1 + v2
v1 ~~ 0*v2
'

mod1.fit <- sem(mod1, sample.cov = temp_mat,
                start = "default",
                control = list(iter.max = 150),
                optim.force.converged = F,
                bounds = "none",
                std.lv = T,
                sample.nobs = 1000)

res <- run.fitprop(mod1.fit,
                   fit.measure = c("logl", "cfi", "srmr", "rmsea"),
                   rmethod = "onion", reps = 1000000,
                   onlypos = FALSE,
                   seed = 3242)
```

What starting values?

How many MLE iterations?

Force convergence?

Place constraints on variances?

What identification constraints?

What sample size?

Fitting propensity analysis choices

```
library(ockhamSEM)
```

Set up temp. correlation matrix for number of variables in model

```
p <- 3
temp_mat <- diag(p)
colnames(temp_mat) <- rownames(temp_mat) <- paste0("v", seq(1:p))
```

Specify model

```
mod1 <- '
v3 ~ v1 + v2
v1 ~~ 0*v2
'
```

Fit model to temp. correlation matrix

```
mod1.fit <- sem(mod1, sample.cov = temp_mat,
                start = "default",
                control = list(iter.max = 150),
                optim.force.converged = F,
                bounds = "none",
                std.lv = T,
                sample.nobs = 1000)
```

What starting values?

How many MLE iterations?

Force convergence?

Place constraints on variances?

What identification constraints?

What sample size?

Run fitting propensity analysis

```
res <- run.fitprop(mod1.fit,
                   fit.measure = c("logl", "cfi", "srmr", "rmsea"),
                   rmethod = "onion", reps = 1000000,
                   onlypos = FALSE,
                   seed = 3242)
```

What fit indices?

How many samples from data space?

Can correlations be any value or positive manifold?

Data and model characteristics

```r
library(ockhamSEM)

p <- 3
temp_mat <- diag(p)
colnames(temp_mat) <- rownames(temp_mat) <- paste0("v", seq(1:p))

mod1 <- '
v3 ~ v1 + v2
v1 ~~ 0*v2
'

mod1.fit <- sem(mod1, sample.cov = temp_mat,
                start = "default",
                control = list(iter.max = 150),
                optim.force.converged = F,
                bounds = "none",
                std.lv = T,
                sample.nobs = 1000)

res <- run.fitprop(mod1.fit,
                   fit.measure = c("logl", "cfi", "srmr", "rmsea"),
                   rmethod = "onion", reps = 1000000,
                   onlypos = FALSE,
                   seed = 3242)
```

Set up temp. correlation matrix for number of variables in model

Specify model

Fit model to temp. correlation matrix

Run fitting propensity analysis

How many *observed* and *latent* variables in the model?

What starting values?

How many MLE iterations?

Force convergence?

Place constraints on variances?

What identification constraints?

What sample size?

What fit indices?

How many samples from data space?

Can correlations be any value or positive manifold?

Research Design

Main Studies

- Pilot Tests
- Study 1: replicate and extend Preacher (2006) example with 2 simple path models
- Study 2: does the number of latent factors (1 to 6) matter?
- Study 3: does the number of observed variables (9 to 45) matter?

Total data space was represented by 1,000,000 sample correlation matrices in each study

Study Outcomes

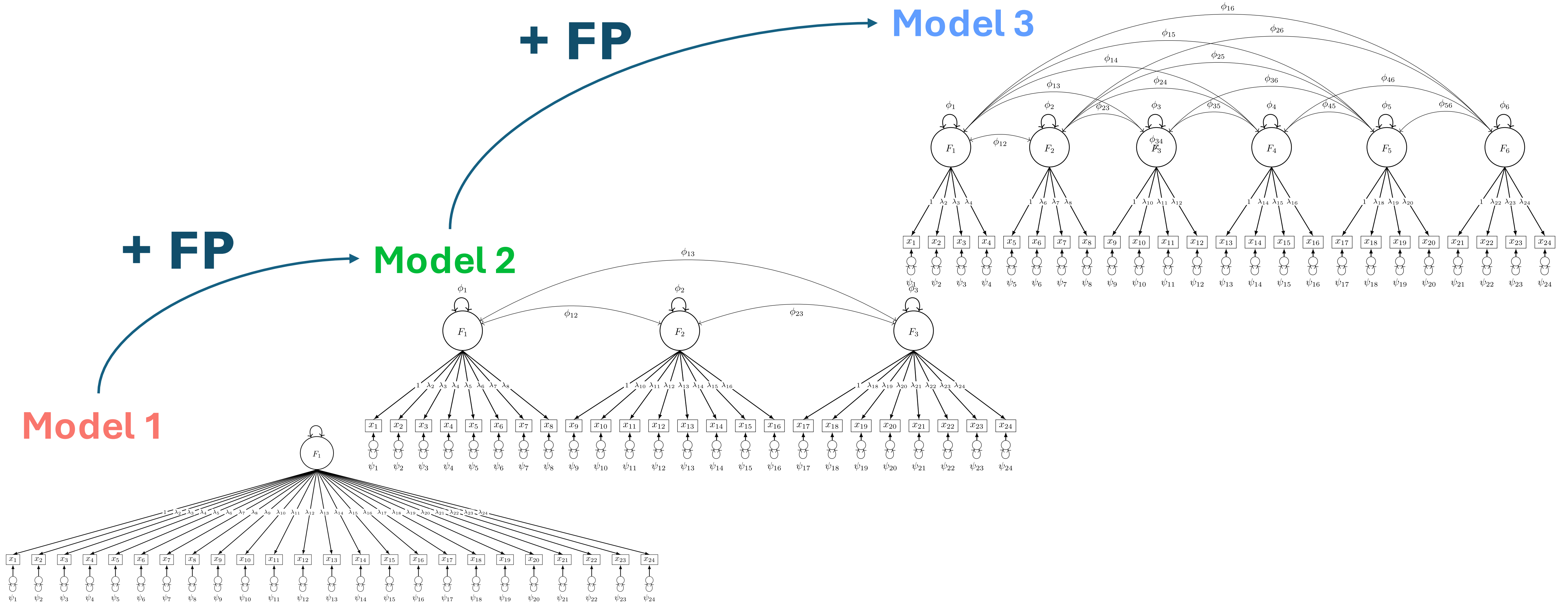
We do not (cannot) know the true fitting propensity that we are aiming for

- Thus, we do not focus on bias, accuracy, etc.

In these studies, focus lies on understanding which researcher decisions *matter* under what circumstances.

- Within each study, we can use knowledge about model complexity to know which model *should* have higher fitting propensity

Intuition about fitting propensity



Data and model characteristics

```r
library(ockhamSEM)

p <- 3
temp_mat <- diag(p)
colnames(temp_mat) <- rownames(temp_mat) <- paste0("v", seq(1:p))

mod1 <- '
v3 ~ v1 + v2
v1 ~~ 0*v2
'

mod1.fit <- sem(mod1, sample.cov = temp_mat,
                start = "default",
                control = list(iter.max = 150),
                optim.force.converged = F,
                bounds = "none",
                std.lv = T,
                sample.nobs = 1000)

res <- run.fitprop(mod1.fit,
                   fit.measure = c("logl", "cfi", "srmr", "rmsea"),
                   rmethod = "onion", reps = 1000000,
                   onlypos = FALSE,
                   seed = 3242)
```

How many *observed* and *latent* variables in the model?

What starting values?

How many MLE iterations?

Force convergence?

Place constraints on variances?

What identification constraints?

What sample size?

What fit indices?

How many samples from data space?

Can correlations be any value or positive manifold?

Data and model characteristics

```r
library(ockhamSEM)

p <- 3
temp_mat <- diag(p)
colnames(temp_mat) <- rownames(temp_mat) <- paste0("v", seq(1:p))

mod1 <- '
v3 ~ v1 + v2
v1 ~~ 0*v2
'

mod1.fit <- sem(mod1, sample.cov = temp_mat,
                start = "default",
                control = list(iter.max = 150),
                optim.force.converged = F,
                bounds = "none",
                std.lv = T,
                sample.nobs = 1000)

res <- run.fitprop(mod1.fit,
                   fit.measure = c("logl", "cfi", "srmr", "rmsea"),
                   rmethod = "onion", reps = 1000000,
                   onlypos = FALSE,
                   seed = 3242)
```

How many *observed* and *latent* variables in the model?

~~What starting values?~~

How many MLE iterations?

Force convergence?

Place constraints on variances?

~~What identification constraints?~~

~~What sample size?~~

What fit indices?

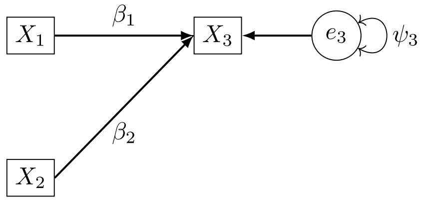
How many samples from data space?

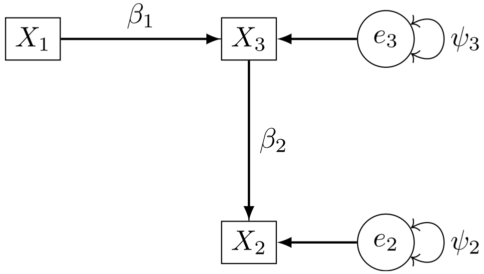
Can correlations be any value or positive manifold?

Duration of fitting propensity analysis

Analyses completed on HPC using 12 cpus per fitting propensity analysis (using one million correlation matrices)

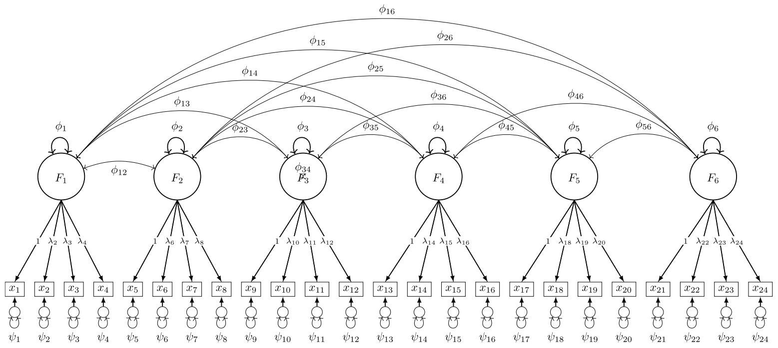
Study 1





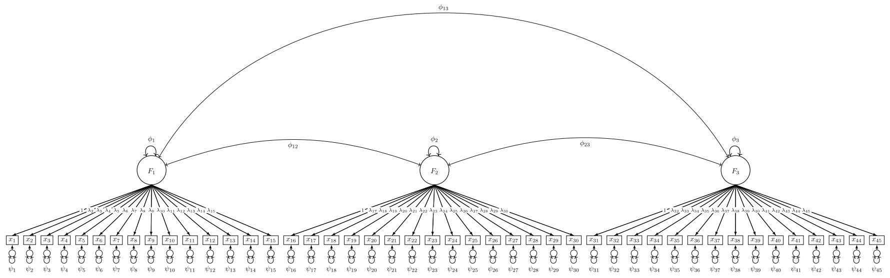
1.54-1.86 hrs

Study 2



105.790 - 643.59 hrs
(4.38 – 26.79 days!)

Study 3



116.42 - 1205.24 hrs
(4.83 – 50.21 days!)

Convergence rates

- Positive manifold data spaces: Convergence rates are high across the board
- All correlations data spaces: Convergence drops when there are…
  - more latent factors
  - more observed variables
  - no bounds on variances
  - fewer MLE iterations

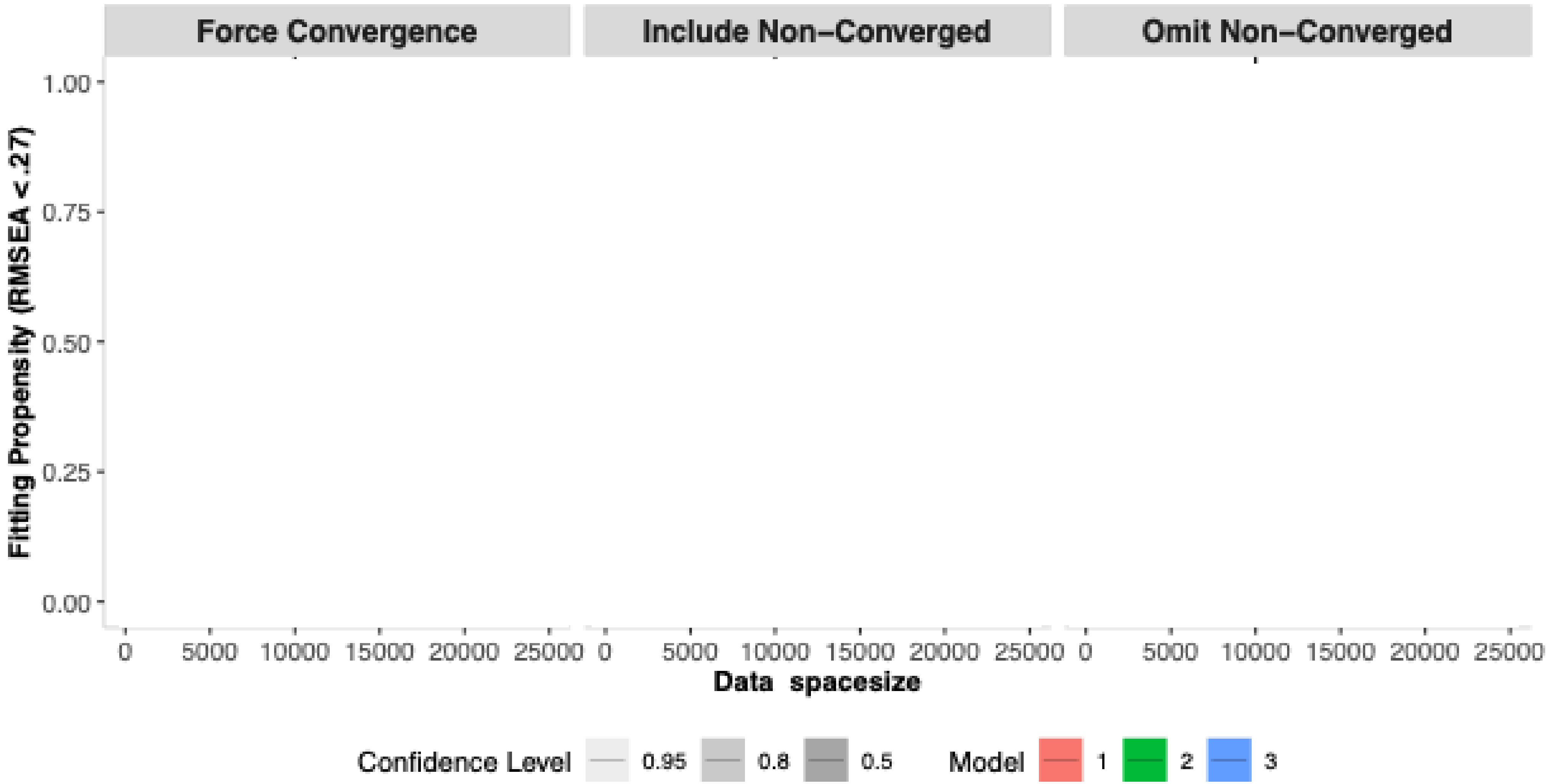
- In the most challenging condition, only **8270 of 1 million** correlation matrices converged (0.83%)
  - 24 indicator, 6-factor model with unit-variance constraints, no variance bounds, running for 150 MLE iterations using the data space based on all possible correlations

Intermezzo

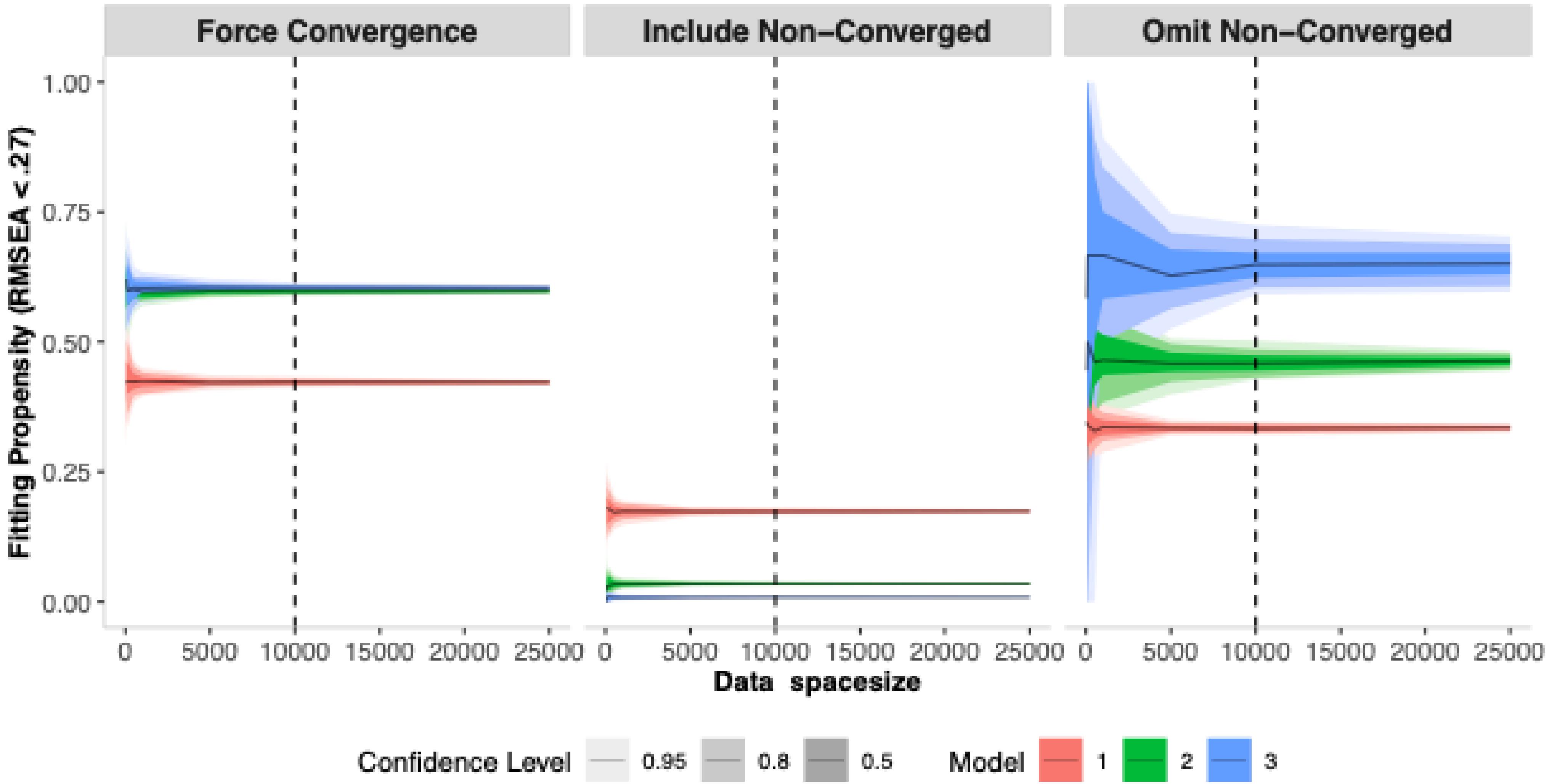
Based on these results, we had the following questions:

- Do we need one million correlation matrices to ensure a stable result?
- (How) do we incorporate model non-convergence?
- Do we need to include all correlations in the data space?

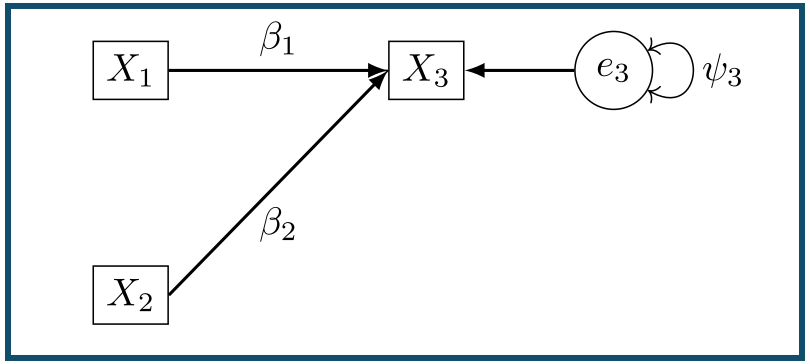
Do we need one million correlation matrices to ensure a stable result?



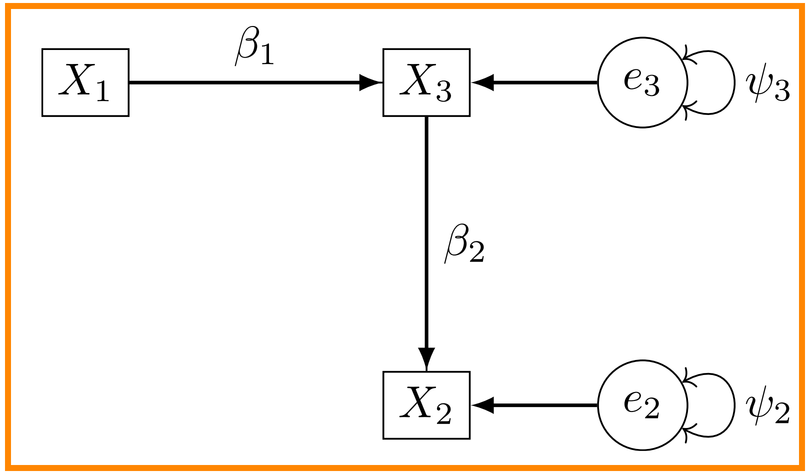
Do we need one million correlation matrices to ensure a stable result?



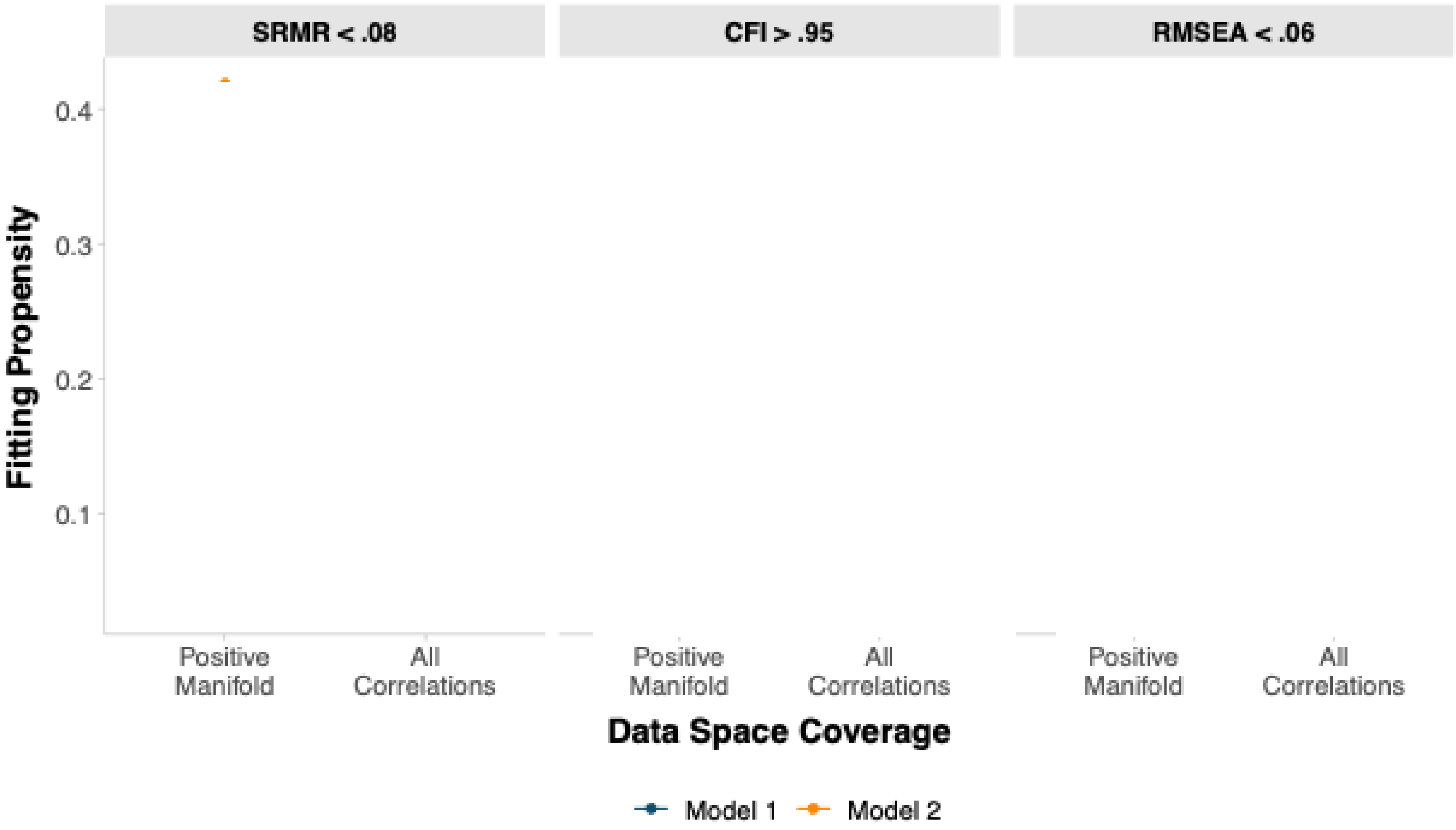
Do we need to include all correlations in the data space?



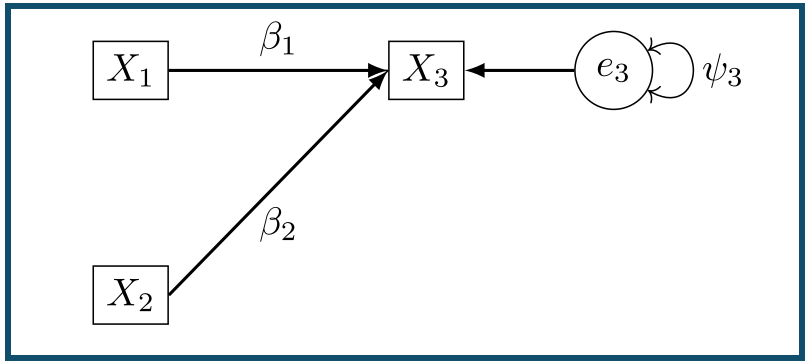
Assumes $Cov(X_1, X_2) = 0$



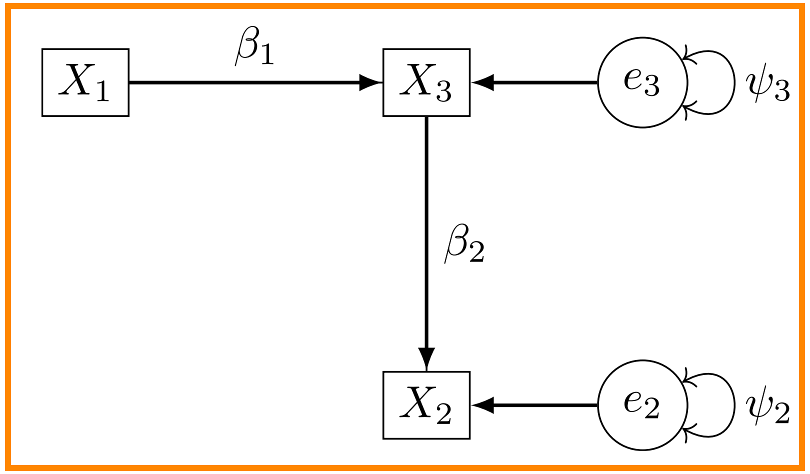
Assumes $Cov(X_1, X_2) = \beta_1 \times \beta_2$



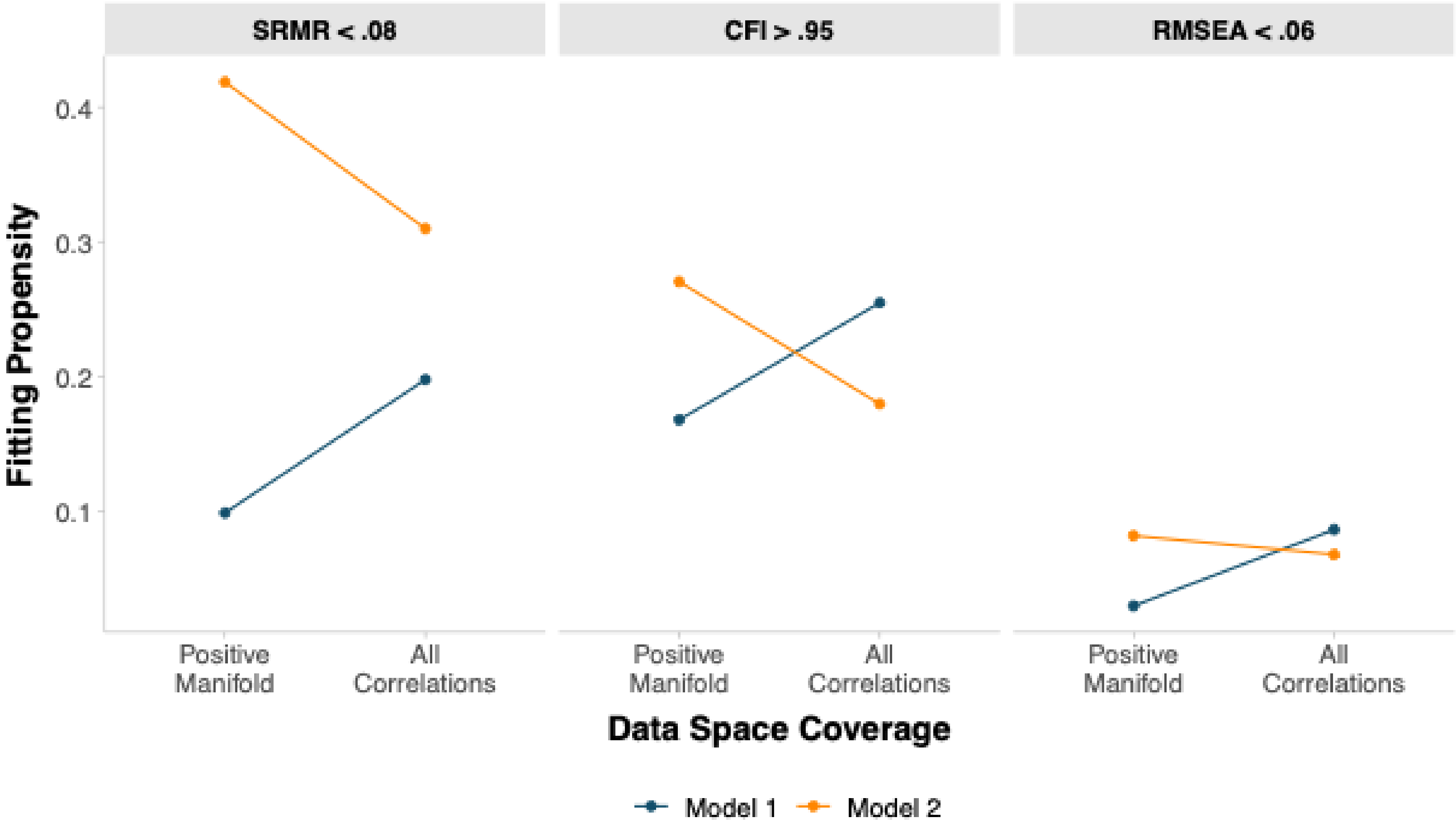
Do we need to include all correlations in the data space?



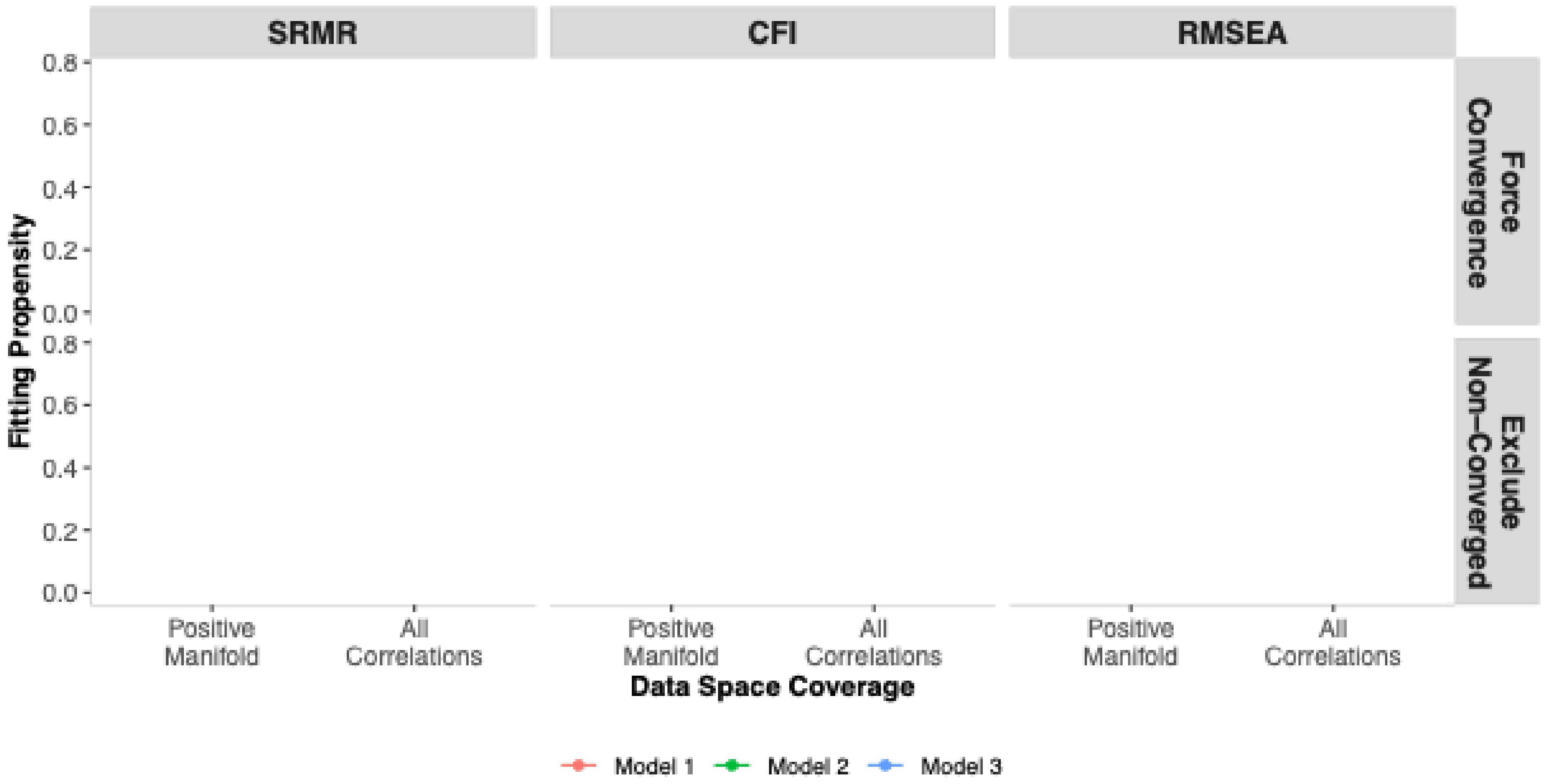
Assumes $Cov(X_1, X_2) = 0$



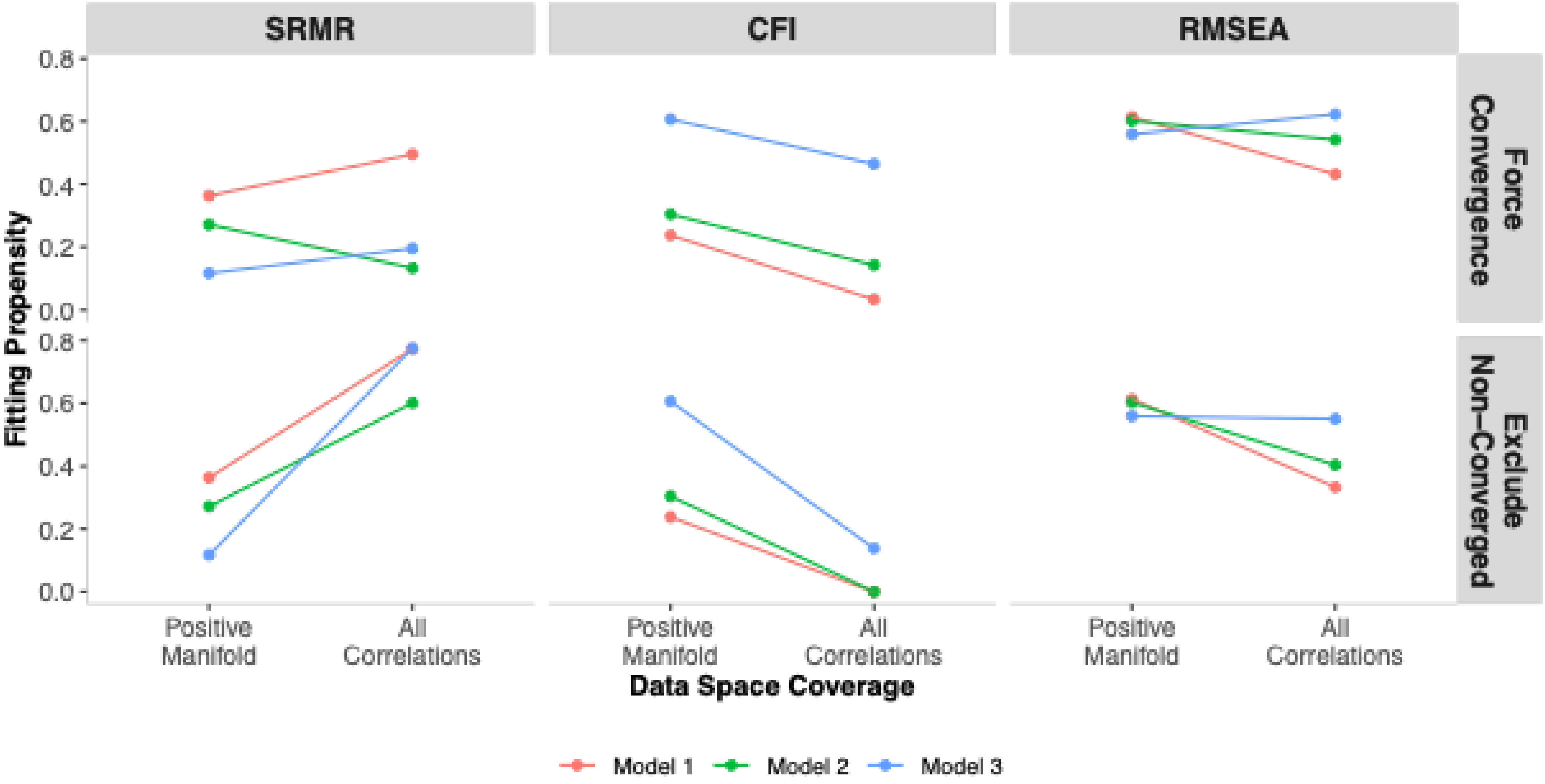
Assumes $Cov(X_1, X_2) = \beta_1 \times \beta_2$



What about for CFAs?



What about for CFAs?



Recommendations

Fitting propensity analysis choices

- If correlations can theoretically be negative, look at full data space
- 10,000 correlation matrices appear sufficient
- Different fit indices can reveal different aspects of a model's fitting propensity

Model estimation options

- Starting values, identification constraints, and sample size do not matter
- Increasing MLE iterations improves convergence up to a point

Next steps

- Using variance bounds improves convergence, but tends to lower FP, does that matter?
- How best to handle non-convergence?
- How to explain discrepancies between fit indices?

Thank you!

Sonja Winter

sdwinter@missouri.edu