Generalised Bayesian SEM

K. Vamvourellis with K. Kalogeropoulos and I. Moustaki

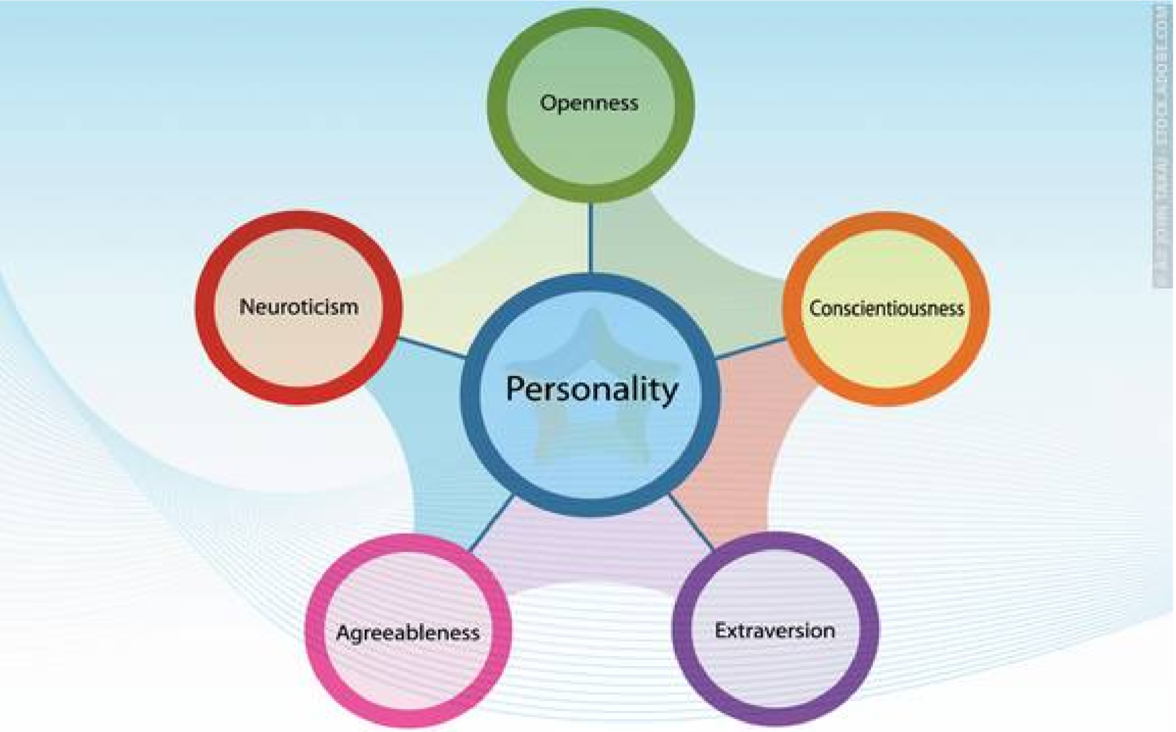
LSE

March 25, 2022

Outline

1. Introduction
2. Generalised Model
3. Model Assessment
4. Applications

Example: Big 5 personality factors in BHPS data



Model Definitions

Data consist of p variables that are observed on N individuals.

Latent factors $z_i = (z_{i1}, \ldots, z_{ik})$, $k < p$, distributed as $\mathrm{N}(0, \Phi)$.

Conditional on z_i, the $i-$th row of the data (y_i) is

$$y_i = \alpha + \Lambda z_i + \epsilon_i, \quad i = 1, \ldots, N,$$

where Λ is a $p \times k$ matrix of rank k, containing the factor loadings.
The error terms ϵ_i independent

$$\epsilon_i \sim N(0_p, \Psi), \quad \Psi = \mathrm{diag}(\psi_1^2, \ldots, \psi_p^2)$$

In this case of Normal distributions we also have a marginal distribution

$$y_i \sim N(\alpha, \Lambda \Phi \Lambda^T + \Psi).$$

Assessing a hypothesised structure

The hypothesised structured is usually assessed via the goodness of fit by contrasting the estimated covariance of y_i's implied by the model with the empirical one.

Not clear what to do when the model does not exhibit 'good fit'.

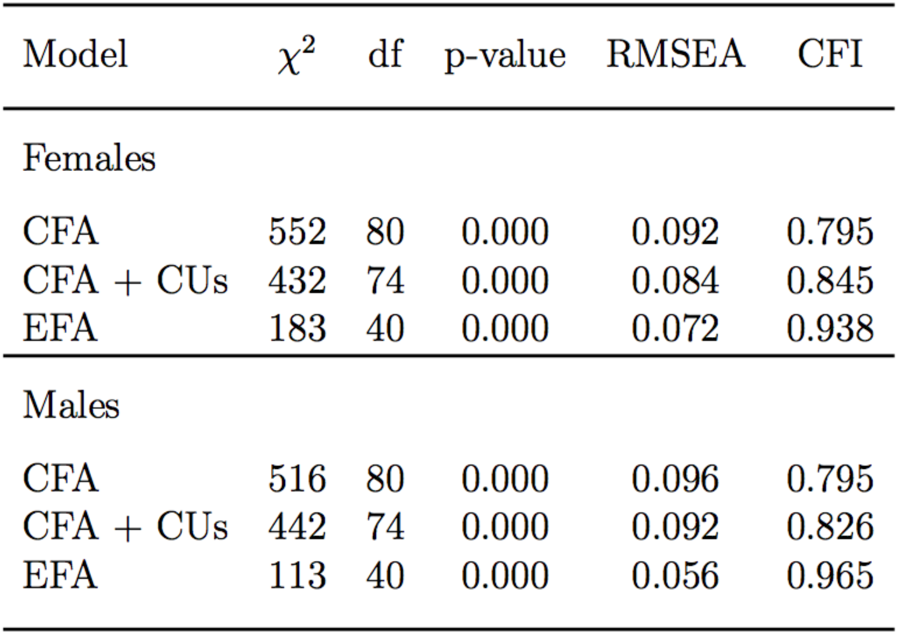
Is the structure inadequate?

Is the structure reflected too strictly in the model? e.g. zero cross-loadings, conditional independence.

Is there imperfect measurement for reasons unrelated to the structure? e.g. wording of questions or measurement error misspecification? (related to matrix Ψ)

Big 5 personality factors: goodness of fit

None of the goodness of fit statistics indicate a good fit. So what do we do next?



Model	χ^2	df	p-value	RMSEA	CFI
Females					
CFA	552	80	0.000	0.092	0.795
CFA + CUs	432	74	0.000	0.084	0.845
EFA	183	40	0.000	0.072	0.938
Males					
CFA	516	80	0.000	0.096	0.795
CFA + CUs	442	74	0.000	0.092	0.826
EFA	113	40	0.000	0.056	0.965

Modification Indices

A modification index measures the improvement in model fit that would result if a previously omitted parameter were to be freely estimated.

Issues:

- This can often lead to a model unsupported by the hypothesised substantive theory by capitalising on chance.
- "Greedy" procedure that is not guaranteed to convergence to an optimal model MA12, AMM15.

Bayesian SEM by [Muthén and Asparouhov, 2012]

Problem: If we free all cross-loadings we are led to EFA and non-identified for non-zero factor covariances.

Answer: [Muthén and Asparouhov, 2012] introduced *approximate zero framework*, treating parameters as approximate rather than exact zero via informative priors.

All the parameters are freed in a single stage, but by very 'little' and the model is now identified in the Bayesian sense.

Marginal model becomes $\mathbf{y}_i \sim N(\alpha, \Lambda\Phi\Lambda^T + \Omega + \Psi)$ where

- the cross-loadings in Λ have informative priors $N(0, 0.01)$
- the matrix $\Omega + \Psi$ has an informative inverse Wishart prior forcing towards being diagonal but allowing for some small covariances.

Goodness of fit of approximate zeros models

Goodness of fit can be assessed by the posterior predictive p-value (PPP) index (Gelman 1996).

Could get around rejecting valid models due to measurement error issues, but is it possible to get a good fit even for a bad model? In other words, could approximate zero models over-fit the data?

Muthen et al (2015) suggest another perspective, by gradually relaxing the prior until the fit becomes good and then exploring for the source of error. The final decision is left to the researchers depending on the case.

Either way more information is required to fully assess the models.

Our approach [Vamvourellis et al., 2021]

Contributions

1. MA12 definition is based on the marginal model $\mathbf{y}_i \sim N(\alpha, \Lambda\Phi\Lambda^T + \Omega + \Psi)$, but this is not available for e.g. logistic based IRT models. We propose a generalised approximate zero framework to handle a much wider model family.

2. We propose a model assessment framework that monitors collectively fit and predictive performance, implemented via cross-validation and scoring rules, to examine whether the approximate zero parameters are picking up noise rather than systematic patterns in the data.

Model

We suggest the following general model:

$$\mathbf{y}_i^* = \alpha + \Lambda \mathbf{z}_i + \mathbf{u}_i + \mathbf{e}_i, \tag{1}$$

$\mathbf{y}_i^*$ may be viewed as a latent variable to cover different data types, i.e. for continuous $\mathbf{y}_i = \mathbf{y}_i^*$, for binary $y_{ij} = \mathcal{I}(y_{ij}^* > 0)$ (ordinal are also possible).

$\mathbf{z}_i$ and $\mathbf{e}_i$ are defined as before. The u_{ij}s may be viewed as item-individual random effects.

For continuous $\mathbf{y}_i$, $\mathbf{z}_i \sim N(0, \Phi)$, $\mathbf{u}_i \sim N(0, \Omega)$ and $\mathbf{e}_i \sim N(0, \Psi)$, we get the MA12 model

$$\mathbf{y}_i \sim N(\alpha, \Lambda \Phi \Lambda^T + \Omega + \Psi).$$

But the framework covers any other choice.

Binary Data

For example, we can work with the following model for **binary data**

$$\begin{cases} y_{ij} = \mathcal{I}(y_{ij}^* > 0), \\ \mathbf{y}_i^* = \alpha + \Lambda \mathbf{z}_i + \mathbf{u}_i + \mathbf{e}_i, \\ \mathbf{e}_i \sim \prod_{j=1}^{p} \text{Logistic}(0, \pi^2/3) \;\text{ or }\; \prod_{j=1}^{p} N(0,1) \\ \mathbf{z}_i \sim N(0, \Phi) \\ \mathbf{u}_i \sim N(0, \Omega). \end{cases}$$

In the above models the $\mathbf{e}_i$s correspond to the logistic and probit specifications that are the most frequently used models, although other choices of distributions are also possible.

Models for ordinal data can be handled in a similar fashion.

Priors

parameters with informative priors

- Λ cross-loadings and $\Lambda_{ij} \sim N(0, 0.1^2)$ informative
- Ω non-diagonal error covariance matrix $\Omega \sim IW(I, p + 6)$

rest of parameters

- Λ major loadings $\Lambda_{ij} \sim N(0, 1)$ for the major i, j (non-informative)
- α mean parameter $\alpha_i \sim N(0, 10^2)$ for all i (non-informative)
- ψ^2 idiosyncratic variances $\psi_j^2 \sim \text{InvGamma}(c_0, (c_0 - 1)/(S_y^{-1})_{jj})$ for all j (only slightly informative to avoid Heywood cases)
- Φ latent factor covariance matrix $\Phi \sim IW(I, p + 4)$ under full covariance parametrisation, or $\Phi \sim LKJ(2)$ otherwise

Model Assessment Framework

We introduce a model assessment framework that collectively uses fit indices and cross-validation to detect overfit.

Complements PPP values, or other similar indices, with scoring rules to evaluate the prediction extracted from the model.

Models so far:

EZ exact zero model. Standard SEM model, defined by $\mathbf{y}_i^* = \alpha + \Lambda \mathbf{z}_i + \epsilon_i$ with the cross-loadings in Λ being fixed to zero.

AZ approximate zero (AZ) model. First introduced in MA12 and generalised in VKM2021. Defined as $\mathbf{y}_i^* = \alpha + \Lambda \mathbf{z}_i + \mathbf{e}_i + \mathbf{u}_i$. The cross-loadings in Λ are no longer being fixed to zero. It is a model to be used only in the Bayesian sense, as the informative priors on the Ω and on the cross-loadings in Λ are essential to ensure identification.

Posterior Predictive Values

Most popular method to assess model fit in the Bayesian SEM framework. It is defined for discrepancy function denoted by $D(\mathbf{Y}, \theta)$ that quantifies how far the fitted model is from the data.

For **continuous data**, $D(\mathbf{Y}, \theta)$ is often set to the LRT statistic that compares the covariance of the estimated model against that of the saturated (empirical covariance).

At each (or some) of the MCMC samples θ_m, $m = 1, \ldots, M$, do the following:

1. Compute $D(\mathbf{Y}, \theta_m)$.
2. Draw $\tilde{\mathbf{Y}}$ having the same size as $\mathbf{Y}$, from the likelihood function $f(\mathbf{Y}|\theta_m)$ of the implied model and using the current value θ_m.
3. Calculate $D(\tilde{\mathbf{Y}}, \theta_m)$ and $d_m = \mathcal{I}\big[D(\mathbf{Y}, \theta_m) < D(\tilde{\mathbf{Y}}, \theta_m)\big]$

Return PPP$= \frac{1}{M} \sum_{m=1}^{M} d_m$.

Posterior Predictive Values

For **binary and ordinal** data, use the parametrisation by response patterns. denoted by $\{\mathbf{y}_r\}_{r=1}^{R}$, with corresponding observed frequencies denoted by O_r where $r = 1, \ldots, R$.

The probability of a response pattern, based on the logistic model defined earlier, is the following integral that can be computed via Monte Carlo

$$\pi_r(\theta) = \int \prod_{j=1}^{p} \text{Bernoulli}\left\{[\mathbf{y}_r]_j | \sigma([\eta]_j)\right\} f(\mathbf{z}) f(\mathbf{u}) d\mathbf{z} d\mathbf{u},$$

The discrepancy function can again be set to the LRT, or else G^2, that compares the model in question against a multinomial model with a separate parameter for each response pattern.

Scoring Rules

Scoring Rules are indices that assess the quality of predictive distributions. Typically small values indicate good performance.

The predictions for unseen data that come in the form of a distribution $h(\mathbf{Y}^{te}|\mathbf{Y}^{tr})$, that is contrasted against the actual test data $\mathbf{Y}^{te}$.
z- The standard Bayesian choice is the posterior predictive distribution

$$f(\mathbf{Y}^{te}|\mathbf{Y}^{tr}) = \int f(\mathbf{Y}^{te}|\theta)\pi(\theta|\mathbf{Y}^{tr})d\theta.$$

A standard choice of a scoring rule is the log score that takes the form of $LS(\mathbf{Y}^{te}) = -\log h(\mathbf{Y}^{te})$, although $h(\cdot)$ is not always available.

Scoring Rules for SEM

In our case, the log score is only available for categorical data under the response patterns formulation. One can show that

$$LS(\mathbf{O}^{te}, \pi^{tr}) = -\log f(\mathbf{O}^{te}|\pi^{tr}),$$

where $f(\cdot)$ is the pmf of the multinomial distribution.

For **continuous data** we use the variogram rule which does not require a closed form posterior predictive distribution

$$VS(\mathbf{y_i}, \tilde{\mathbf{y}}) = \sum_{j=1}^{p}\sum_{k=1}^{p} w_{j,k}\left(|y_{ij} - y_{ik}|^P - \frac{1}{m}\sum_{m=1}^{M}|\tilde{y}_{mj} - \tilde{y}_{mk}|^P\right)^2$$

Cross-validation

We propose using scoring rules and cross validation to measure the predictive performance of our model.

1. Split the data randomly into K parts.
2. For each of the K parts repeat the following steps:
   - Designate the selected group as the test data set and use the other $K-1$ groups together as the training data set.
   - Fit the model in the training data set and draw samples from its posterior and its posterior predictive distribution to predict the data in the test set.
   - Evaluate the predictions via the chosen scoring rule against the test data.
3. Aggregate the values of the scoring rules across all K groups by summing or averaging.

This way all data points appear in both training and test sets, and the result is more stable. The number of folds K can be chosen to satisfy the trade off between train time and test sample size.

Model Assessment Framework

The existing model assessment procedure can be summarised as follows:

Step 1 Try EZ model. If it achieves good fit, conclude support from the data. Otherwise go to step 2.

Step 2 Try the AZ model. If there is no good fit, then the hypothesised structure is not supported by the data. If it achieves good fit, go to step 3.

Step 3 EZ does not fit, but AZ does. We need to check if the AZ model overfits the data.

Our approach requires benchmark model for predictive perfomance performance to guard against overfitting. The saturated model is not recommended because it is over-parametrised and as such can lead to poor predictions.

Model Assessment Framework (cont'd)

We propose using the **corresponding EFA model** as the benchmark.

This model has generally fewer parameters than the saturated, and should achieve high predictive performance as it is allowed to search for systematic patterns in the data without any restrictions, other than having k factors.

We propose two types of EFA:

EFA The standard exploratory factor analysis model (EFA). It is the standard EFA model, defined by $\mathbf{y}_i = \alpha + \Lambda z_i + \epsilon_i$ where low informative priors are assigned to all the components of Λ and $\Phi = I$.

EFA-C The EFA model with item-individual random effects. Defined as $\mathbf{y}_i = \alpha + \Lambda z_i + u_i + e_i$ allows for a small amount of item dependencies conditional on the extracted independent factors.

Model Assessment Framework (cont'd)

We can now define the *Step 3* used earlier explicitly.

Step 3:

(a) Compute the scoring rules of EZ and AZ. If EZ is better than AZ, then AZ is likely overfitting and there is little support towards the hypothesised model.

(b) Else, compare the AZ against the EFA models. In cases of comparable or improved performance there is supporting evidence towards the hypothesised model.

Simulations Continuous Data

Generated data from a 2-factor 6-item model according to the following scenarios:

- Scenario 1: Data generated from the EZ model.
- Scenario 2: Data generated from the AZ model with small error correlations, introduced by item-individual random effects, and without cross loadings.
- Scenario 3: Data generated from the AZ model with two non-negligible cross loadings (shown below) and without correlated item-individual random effects.

Scenario 1		Scenario 2		Scenario 3	
z_1	z_2	z_1	z_2	z_1	z_2
1	0	1	0	1	0
.8	0	.8	0	.8	0
.8	0	.8	0	.8	.6
0	1	0	1	.6	1
0	.8	0	.8	0	.8
0	.8	0	.8	0	.8

Simulations Continuous Data

	Scenario 1		Scenario 2		Scenario 3	
Model	PPP	VS	PPP	VS	PPP	VS
EZ	0.66	**0**	0.00	6.93	0.00	17.79
AZ	0.51	4.28	0.31	**0**	0.53	1.58
EFA	0.62	2.06	0.00	0.23	0.59	**0**
EFA-C	0.53	1.05	0.38	0.03	0.56	1.45

Scenario 1: If EZ is correct, it will dominate.

Scenario 2: AZ goes all the way to *Step 3b* where it is very competitive against EFA models. Strong support despite measurement error issues.

Scenario 3: As Scenario 2 but AZ underpeforms the EFA models in *Step 3b*. Little support towards the model that misses a substantial cross-loading.

Simulations Binary Data

Performed the same simulation using binary data under the logit model.

	Scenario 1		Scenario 2		Scenario 3	
Model	PPP	LS	PPP	LS	PPP	LS
EZ	0.52	**0**	0.02	4.19	0.00	7.31
AZ	0.50	0.68	0.12	**0**	0.52	1.90
EFA	0.59	1.45	0.13	0.09	0.45	**0**
EFA-C	0.54	3.27	0.17	0.24	0.50	2.96

Notice how the guidelines can be used in a similar fashion to assess whether the hypothesised structure is valid or not.

Application Continuous Data

Data collected by **British Household Panel Survey in 2005-06** focusing on female subjects between the ages of 50 and 55; the sample size consists of 589 individuals.

The ‘Big 5 Personality Test’, 15-item questionnaire on topics of social behaviour and emotional state.

Participants answer each item on a scale from $1-7$, 1 being ‘strongly disagree’ and 7 being ‘strongly agree’. Items are treated here as continuous.

The test is designed to measure five major, potentially correlated, personality traits.

Each trait corresponds to a factor, and each factor is hypothesised to explain exactly 3 out of 15 items.

‘Big 5’ Results

Model	PPP	VS
EZ	0.0	56.43
AZ	0.23	0
EFA	0.00	94.35
EFA-C	0.38	78.47

Very similar to simulation Scenario 2, yet much more pronounced. Our analysis confirms the poor fit of the EZ and the EFA with five factors. Both AZ and EFA-C models have reasonably good PPP values. This implies that error correlations contribute to the lack of fit to a large extent.

The variogram score of the AZ model clearly dominates all the other models. Strong support towards the ‘Big 5’ scale, attributing the fit issues of the EZ model to error correlations that could have been caused by the wording and other issues often present in survey data like the BHPS.

Application Binary Data

Data on 566 patients available through the National Institute on Drug Abuse.

The **Fagerstrom Test for Nicotine Dependence** (FTND) was designed to provide a measure of nicotine dependence related to cigarette smoking.

It contains six items that evaluate the quantity of cigarette consumption, the compulsion to use, and dependence.

No conclusive hypothesised theory is reached yet. Previous models included a single factor, a correlated two factor, and a two factor model with one cross loading. It is conjectured that first item loads on both factors. These models were also considered in our analysis and are denoted as 1F, 2F-EZ, and 2F EZ-b respectively.

FTND Survey Data

The original scale consists of 4 binary and 2 ordinal items for self-declared smokers:

1. FNFIRST: How soon after you wake up do you smoke your first cigarette?
2. FNGIVEUP: Which cigarette would you hate most to give up?
3. FNFREQ: Do you smoke more frequently during the first hours after waking than during the rest of the day?
4. FNNODAY: How many cigarettes/day do you smoke?
5. FNFORBDN: Do you find it difficult to refrain from smoking in places where it is forbidden (e.g., in church, at the library, in cinema, etc.)?
6. FNSICK: Do you smoke if you are so ill that you are in bed most of the day?

For the purposes of our analysis, item FNFIRST was dichotomised as '1'=[3] and '0'=[0,1,2] and item FNNODAY as '1'=[2,3] and '0'=[0,1].

FTND Results

Model	PPP	LS
1F	0.01	15.98
1F-C	0.32	6.63
2F-EZ	0.04	10.45
2F-AZ	0.40	6.23
2F-EZ-b	0.41	0.00
2F-AZ-b	0.44	2.01
2F-EFA	0.44	2.66
2F-EFA-C	0.58	2.38

The best model is the 2F-EZ-b correcting the misspecifications of 2F-EZ with a single additional parameter. The fact that the log score of 2F-EZ-b is smaller than that of the EFA models provides support towards the scale with two correlated factors where the item ‘FNFIRST’ loads on both of them.

Discussion-Extensions

- Assessing goodness of fit is an important question more broadly in Bayesian semi/non-parametrics, not just for SEM.
- Useful in multi-group factor analysis, where some parameters are hypothesised to be fixed across groups.
- Our simulations are meant to be a proof of concept, more elaborate experiments are needed to clarify terms such as ‘comparable’.
- In this work we complemented goodness of fit measures with out of sample performance. An alternative method is based on *model evidence* or *Bayes Factors* (more on this on the thesis).
- Here we looked at models for continuous or binary data. It is possible to extend this framework to accommodate data of mixed type (more on this on the thesis).
- All code is available at `github.com/bayesways/bayes-sem/`.

References I

Muthén, B. and Asparouhov, T. (2012).
Bayesian Structural Equation Modeling: A more flexible representation of substantive theory.
Psychological Methods, 17:313–335.

Vamvourellis, K., Kalogeropoulos, K., and Moustaki, I. (2021).
Generalised bayesian structural equation modelling.
arXiv preprint arXiv:2104.01603.