Bayesian estimation of Confirmatory Factor Analysis for Count Data with extensions for cross-classification

Abstract

We develop a confirmatory factor analysis (CFA) model under the assumption that the data are Poisson. The estimation method is Bayesian, and we provide prior specifications for model parameters. Bayesian Poisson CFA of a single simulated dataset (n = 300) shows that the method performs acceptably in terms of parameter recovery, and the parameter estimates are similar to maximum likelihood estimates. We extend the Poisson CFA to cross-classified data. Analysis of a single simulated dataset (n = 300), with two crossed hierarchies shows acceptable parameter recovery for the Bayesian approach. These preliminary results suggest that the Bayesian approaches we developed may be adopted for CFA of single or multilevel Poisson item indicators.

Keywords: CFA, Bayesian estimation, Poisson, cross-classification, RStan

Word count: 1971

Bayesian estimation of Confirmatory Factor Analysis for Count Data with extensions for cross-classification

Confirmatory factor analysis (CFA) is a structural equation modeling (SEM) approach for measurement models. The major focus of a CFA is the relationship between item indicators and latent traits or factors as captured by factor loadings (Brown, 2006; Schumacker & Lomax, 2004). Additionally, the relationship between the factors is often of interest. These factor loadings and interfactor correlations may sufficiently characterize the relationship between the item indicators, such that CFA can be viewed as a data reduction technique (Bollen, 1989, p. 227).

In this paper, we present CFA methods assuming the data are Poisson distributed. Our estimation method is Bayesian primarily for the reason laid out by Kruschke (2013) − a Bayesian analysis yields more information about parameters of interest relative to an analogous frequentist analysis. And our Bayesian computational engine of choice is Stan (Carpenter et al., 2017), which we accessed using R, both software are free.[1]

Our contribution here is of value for a number of reasons. Outside of Mplus − which is not free software, we failed to find a CFA implementation for count data, although there are frequentist item response models for count data (e.g. Magnus & Thissen, 2017; Wang, 2010). Additionally, count data appear in educational settings, especially in the study of behaviours, see examples in Wang (2010). We also provide extensions to the Poisson CFA to account for multilevel data structures since such structures are commonplace in educational settings (Goldstein, 1995). Finally, the Poisson model we present is a foundation for more flexible modeling alternatives; hence, our work here is a first step in CFA analysis of count data.

In the remainder of this paper, we lay out the model specification for the Poisson CFA. We then present the Bayesian estimation approach with prior choice recommendations. Next,

[1] If our work is accepted, we will provide a link to a GitHub gist file containing all our R and Stan code.

we show the Bayesian method recovers the parameters of interest using data we simulate. We extend the Poisson CFA to account for cross-classified data and show acceptable parameter recovery using simulated data. Finally, we end with recommendations for future research.

Poisson CFA

With continuous indicator CFA, we assume (Bollen, 1989, Eq. 7.1):

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \quad \mathbb{E}(\boldsymbol{\xi}) = 0,\ \mathbb{E}(\boldsymbol{\delta}) = 0,\ \mathrm{Cov}(\boldsymbol{\xi}, \boldsymbol{\delta}) = 0 \tag{1}$$

where $\mathbf{x}$ is a p-dimensional vector of p observed variables (item indicators) in deviation form (mean-centered), $\boldsymbol{\Lambda}$ is a $p \times m$ factor loading matrix for p observed variables and m latent variables, $\boldsymbol{\xi}$ is an m-dimensional vector of m latent variables, and $\boldsymbol{\delta}$ is a p-dimensional vector of error terms, one for each observed variable. One can decompose the $p \times p$ model-implied covariance matrix of $\mathbf{x}$ (Bollen, 1989, Eq. 7.5):

$$\begin{aligned}
\mathbb{E}(\mathbf{x}\mathbf{x}') &= \mathbb{E}\Big[(\boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta})'\Big] = \mathbb{E}\Big[(\boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\xi}'\boldsymbol{\Lambda}' + \boldsymbol{\delta}')\Big] \\
&= \boldsymbol{\Lambda}\mathbb{E}(\boldsymbol{\xi}\boldsymbol{\xi}')\boldsymbol{\Lambda}' + \mathbb{E}(\boldsymbol{\delta}\boldsymbol{\delta}') \iff \mathrm{Cov}(\boldsymbol{\xi}, \boldsymbol{\delta}) = 0 \\
&= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}
\end{aligned} \tag{2}$$

where $\boldsymbol{\Phi}$ is an $m \times m$ inter-factor covariance matrix, and $\boldsymbol{\Theta}$ is a $p \times p$ residual covariance matrix. Of concern is model identification or the existence of a unique solution for $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$. For identification purposes, we can standardize the latent variables, $\boldsymbol{\xi}$, to have unit variance $\Big(\mathrm{diag}(\boldsymbol{\Phi}) = \mathbf{1}_{\mathsf{m}}\Big)$ such that $\boldsymbol{\Phi}$ is the inter-factor correlation matrix. It is also typical to constrain several elements in $\boldsymbol{\Lambda}$ and most off-diagonal elements in $\boldsymbol{\Theta}$ to zero for model identification and to match the hypothesized factor structure (Bollen, 1989, p. 239).

Assuming that $\mathbf{x}$ is multivariate normal $\Big(\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta})\Big)$ provides a maximum likelihood (ML) approach for estimating the model parameters. At this juncture, we can develop the Poisson model. We replace $\mathbf{x}$ in the formulation above with $\boldsymbol{\eta}$, and assume $\boldsymbol{\eta}$ is a p-dimensional vector of p **latent** variables in deviation form. Hence:

$$\boldsymbol{\eta} = \boldsymbol{\Lambda}\boldsymbol{\xi}, \quad \mathbb{E}(\boldsymbol{\eta}\boldsymbol{\eta}') = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' \quad \text{and} \quad \boldsymbol{\eta} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}') \tag{3}$$

We represent the observed count indicators with $\mathbf{y}$, a p-dimensional vector of p count indicators. We represent $\mathbf{y}$ in long form as $\mathbf{y}_g$, a $(p \cdot n)$-dimensional column vector, where n is the number of respondents to the items. Similarly, $\boldsymbol{\eta}_g$ is the long-form version of $\boldsymbol{\eta}$, such that $\boldsymbol{\eta}_g$ is a $(p \cdot n)$-dimensional column vector. Then the Poisson CFA is:

$$\mathbf{y}_g \sim \text{Poisson}\Big(\exp\Big(\boldsymbol{\nu} + \boldsymbol{\eta}_g + \ln(\boldsymbol{\phi})\Big)\Big) \tag{4}$$

where $\boldsymbol{\nu}$ represents the intercept parameter on the log-scale, such that $\boldsymbol{\nu}$ contains p distinct elements. $\boldsymbol{\nu} + \boldsymbol{\eta}_g$ is the mean of the observed Poisson data on the log scale, and $\boldsymbol{\phi}$ is the offset variable which can be used to model the data as rates. If we set $\boldsymbol{\phi} = \mathbf{1}$, the offset has no effect on model estimation. We note that the assumption of multivariate normality on $\boldsymbol{\eta}$ in equation 3 is a simply "device" for model estimation, and a researcher may assume alternative multivariate distributions. Additionally, model identification considerations on $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ continue to apply.

Bayesian estimation

We follow a subjective Bayesian approach (Goldstein, 2006) in the manner precribed by Greenland (2006). Greenland (2006) recommended using the middle 95% interval of prior distributions to a-priori identify the plausible values for parameters; we adopt this recommendation. We assume that factor loadings span the real number line. However, we select one loading per factor that we constrain to be positive − the vector of these *marker* loadings is $\boldsymbol{\lambda}_{\mathsf{m}}$, and the vector of all other loadings in $\boldsymbol{\Lambda}$ that are not constrained to be zero or positive is $\boldsymbol{\lambda}_{\mathsf{nm}}$. Also, let $\mathbf{L}$ be the Cholesky factor of $\boldsymbol{\Phi}$, $\boldsymbol{\Phi} = \mathbf{L}\mathbf{L}'$. Then the Bayesian model is:

$$\begin{aligned}
\mathbf{y}_g &\sim \text{Poisson}\Big(\exp\big(\boldsymbol{\nu} + \boldsymbol{\eta}_g + \ln(\boldsymbol{\phi})\big)\Big) \quad \text{likelihood} \\
\boldsymbol{\eta} &\sim \mathcal{N}_p\big(\mathbf{0}, \boldsymbol{\Lambda}(\mathbf{L}\mathbf{L}')\boldsymbol{\Lambda}' + \delta\mathbf{1}\big), \quad \delta = 0.01 \text{ to ensure positive definiteness} \\
\boldsymbol{\lambda}_{\mathsf{m}} &\sim \mathcal{N}^+(0, \sigma_\lambda)\ , \quad \boldsymbol{\lambda}_{\mathsf{nm}} \sim \mathcal{N}(0, \sigma_\lambda)\ , \quad \sigma_\lambda \sim \mathcal{N}^+(0,1) \\
\mathbf{L} &\sim \text{LKJ}_{\mathsf{chol}}(\tau)\ , \quad \boldsymbol{\nu} \sim \mathcal{N}(0, \sigma_\nu)
\end{aligned} \tag{5}$$

Similarly to software defaults in Mplus (Muthén & Asparouhov, 2012) and blavaan (Merkle & Rosseel, 2018), we assume the loadings are normally distributed; the marker loadings are assumed half-normal. We assume a standard half-normal prior for σ_λ. Hence, there is an a-priori 95% chance that σ_λ will be under 2. This prior permits a considerable 95% plausible interval for non-zero loadings, approximately $(-4, 4)$. Precisely, $\Pr(0 < \sigma_\lambda < 1.96) = 95\%$, and $\Pr(-3.84 < \lambda < 3.84) = 95\%$. Other choices for the prior on σ_λ include inverse-gamma on σ^2_λ and half-Cauchy/t on σ_λ (Gelman, 2006).

For $\mathbf{L}$, we assume an LKJ-prior (Lewandowski, Kurowicka, & Joe, 2009) reparameterized for the Cholesky factor as recommended by the Stan manual (Stan

Development Team, 2018, sec. 63). For the shape parameter, τ, we recommend a value of 2 to reduce the probability of extremely positive or negative correlations between factors. And we recommend $\sigma_\nu = 2.5$, implying $\Pr(-5 < \boldsymbol{\nu} < 5) = 95\%$. This prior on $\boldsymbol{\nu}$ permits means over 100, $\exp(5) = 148$.

All prior recommendations here are defaults and we expect that in certain situations, such as when the data do not contain much information, researchers will need alternative or stronger priors.

Example using simulated data

Simulated dataset. We simulated a dataset according to equations 3 and 4 to test parameter recovery of the proposed Bayesian model. We set $\boldsymbol{\nu} = \ln([1, 1.25, 1.5, \ldots, 3])$, $n = 300$, $p = 9$, $\boldsymbol{\phi} = \mathbf{1}$ and:

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1.3 & 0 & 0 \\ 1.0 & 0 & 0 \\ 0.7 & 0.4 & 0 \\ 0 & 1.3 & 0 \\ 0 & 1.0 & 0 \\ 0.3 & 0.7 & 0 \\ 0 & 0 & 0.8 \\ 0 & 0 & 0.9 \\ 0 & 0 & 1.0 \end{pmatrix}, \quad \boldsymbol{\Phi} = \begin{pmatrix} 1.0 & 0.4 & 0.3 \\ 0.4 & 1.0 & -0.3 \\ 0.3 & -0.3 & 1.0 \end{pmatrix}$$

Given the parameters, there are cross-loadings in the data; factor 1 is positively related with factors 2 and 3, and factor 2 and 3 are negatively related. We added 10^{-9} to the diagonal of $\boldsymbol{\Lambda\Phi\Lambda'}$ to ensure the matrix was positive-definite. We present the simulated data in Figure 1. The items with larger factor loadings (e.g. 1 and 4) had larger ranges.

Bayesian analysis. We used Stan to fit the Bayesian model. Stan uses the No-U-Turn sampler (NUTS). NUTS, an algorithm for sampling continuous parameters, is more efficient than the Gibbs sampler (Hoffman & Gelman, 2014). For Bayesian parameter estimation, we drew 2000 posterior samples across 4 parallel chains, and retained the final half of the samples within each chain. This left us with 4000 samples for inference. As a crude check of sampling convergence, we computed the potential scale reduction factor ($\widehat{R}$) and effective sample size (n_{eff}) for the posterior samples of parameters. $\widehat{R}$ close to 1 and n_{eff} above a thousand are preferred (Carpenter et al., 2017). The marker iterms for the Bayesian analysis were items 1, 4 and 7, hence their loadings were constrained to be positive.

We performed the same analysis using ML estimation in Mplus. On a laptop with a 3.5 GHz i7 processor running on Ubuntu, Stan converged in 253 seconds while Mplus converged in 158 seconds. All n_{eff} values for the loadings, intercepts, and interfactor correlations were above 2000, and all $\widehat{R}$ were one to two decimal places. We compared the parameter estimates to the population values in Figure 2. The Bayesian point estimates and ML point estimates were near-identical, and for most parameters, these estimates were close to their population values. The advantage of the Bayesian approach is our ability to use the posterior samples to make probabilistic statements about the estimated parameters given the data. For example, an analysis of the posterior samples of the loadings revealed that there was a four in five chance that the loading of item 1 on factor 1 (λ_{11}) was greater than all other loadings, $\#\Big(\lambda_{11} = \max(\lambda_{\mathsf{ij}})\Big)/4000 = .83$, for $\mathsf{i} = [1, 2, \ldots, 9]$, $\mathsf{j} = [1, 2, 3]$, where $\#(\cdot)$ is the count function.

Extension to multilevel data structures

To demonstrate the flexibility of a Bayesian approach, we extend the Poisson CFA to cross-classified data structures:

$$\mathbf{y}_g = \text{Poisson}\Big(\exp\Big(\boldsymbol{\nu} + \boldsymbol{\eta}_g + \mathbf{Z}\boldsymbol{\gamma} + \ln(\boldsymbol{\phi})\Big)\Big)\ ; \quad \mathbf{Z} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{1})\ , \quad \boldsymbol{\gamma} \sim \mathcal{N}^+(0, 1) \tag{6}$$

where $\mathbf{Z}$ is a $(p \cdot n) \times k$ matrix of k uncorrelated random effects, and each random effect is standard normal. $\boldsymbol{\gamma}$ is a k-dimensional vector of loadings, one for each random effect in $\mathbf{Z}$. In this model, we claim that each random effect/latent variable in $\mathbf{Z}$ represents a grouping structure. The effect of each grouping structure i.e. the loading (e.g. γ_1 for grouping structure 1) is the same across all items − this is a restriction that can be relaxed. Finally, we assume a standard half-normal prior for the loadings. The specification for the cross-classified Poisson CFA in equation 6 leverages the connection between generalized linear mixed models and structural equation modeling, and subsumes nested data structures (e.g. Rabe-Hesketh, Skrondal, & Pickles, 2004; Rabe-Hesketh, Skrondal, & Zheng, 2007).

To make the above concrete, we provide an example. Assume five count indicators measure a single factor. The respondents are 300 students who attend 50 different schools and live in 30 different neighbourhoods. And both schools and neighbourhoods influence the students' responses on the count indicators. Figure 3 captures this structure, and we simulated data according to this structure. The data were balanced with 6 students per school and 10 students per neighbourhood.

We analyzed the simulated data using both the original Poisson CFA and the cross-classified Poisson CFA (CCPCFA). We used item one as our marker variable. We retained the same estimation settings from the earlier example. The results are available in Figure 4. Stan converged in 154 and 251 seconds for the Poisson CFA and CCPCFA respectively. For the Poisson CFA, all n_{eff} values for the loadings, and intercepts were above 3000, and all $\widehat{R}$ were one to two decimal places. However, for the regular Poisson CFA, n_{eff} values were somewhat lower but still aceptable, we report them in Figure 4. $\widehat{R}$ for estimated parameters were at most one to two decimal places.

The standard Poisson CFA consistently overestimated the student-level loadings. On the other hand, the CCPCFA model demonstrated adequate recovery of the student-level loadings. The estimate for the school loading was closer to the parameter than the estimate of the neighbourhood loading. We found that we could increase the n_{eff} for the school loading (985) by reducing the scale of the prior on the loadings, e.g. $\boldsymbol{\gamma} \sim \mathcal{N}(0, 0.5)$; we do not report those results here.

Discussion

In this study, we have presented a Bayesian CFA for count indicators that are Poisson, with extensions for multilevel data. We are not aware of other instances in the literature of Bayesian estimation of count CFA for multilevel data. Second, the approach presented here will scale to other types of non-normal data. For example, one can use a negative binomial likelihood in place of a Poisson likelihood to model overdispersed counts, while retaining the same structure in equations 5 and 6. When the data are truncated, e.g. “how many days in the last 30 days did …?”, then a truncated Poisson or negative binomial distribution may be adequate. All of these options are possible in Stan.[2]

We simplified the CFA by omitting residual covariances between count indicators. Selectively estimating residual covariances poses a challenge for Bayesian analysis because one cannot simply estimate relevant covariances independently, or we run the risk of estimating non-positive definite covariance matrices. We see two promising approaches. One is the approach adopted by Muthén and Asparouhov (2012) for Gaussian data where one estimates the complete residual covariance matrix, using strong priors to ensure the model is identified. Another is the parameter extension approach recommended by Palomo, Dunson,

[2] The code snippet we intend to provide on GitHub will include a negative-binomial CFA example in Stan.

and Bollen (2007), and implemented by Merkle and Rosseel (2018) for Gaussian CFA.

Finally, we did not explore model fit assessment. We see two promising approaches. One is to develop a Poisson CFA where the data on the log-scale are multivariate normal with an intercept per indicator, and the full covariance matrix of all indicators − hence a baseline model. The practitioner can then use Bayesian model comparison (Vehtari & Ojanen, 2012) to select a working model from one or more hypothesized factor structures and the baseline model. Another approach is the posterior predictive checking approach used by Muthén and Asparouhov (2012) to assess a model without needing to estimate a baseline model. We intend to explore residual covariances and model fit assessment in future studies.

References

Bollen, K. A. (1989). *Structural equations with latent variables* (p. 514). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9781118619179

Brown, T. A. (2006). Introduction. In *Confirmatory factor analysis for applied research* (pp. 1–11). New York: The Guilford Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi:10.18637/jss.v076.i01

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. doi:10.1214/06-BA117A

Goldstein, H. (1995). *Multilevel statistical models*. London: Arnold.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*(3), 403–420. doi:10.1214/06-BA116

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology*, *35*(3), 765–775. doi:10.1093/ije/dyi312

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623. Retrieved from https://dl.acm.org/citation.cfm?id=2627435.2638586

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi:10.1037/a0029146

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. doi:10.1016/J.JMVA.2009.04.008

Magnus, B. E., & Thissen, D. (2017). Item response modeling of multivariate count data with zero inflation, maximum inflation, and heaping. *Journal of Educational and Behavioral Statistics*, *42*(5), 531–558. doi:10.3102/1076998617694878

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. doi:10.18637/jss.v085.i04

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. doi:10.1037/a0026802

Palomo, J., Dunson, D. B., & Bollen, K. A. (2007). Bayesian structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 163–188). Elsevier/North-Holland.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. doi:10.1007/BF02295939

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 209–227). Elsevier/North-Holland.

Schumacker, R. E., & Lomax, R. G. (2004). Factor analysis. In *A beginner's guide to structural equation modeling* (pp. 85–105). Routledge.

Stan Development Team. (2018). Stan modeling language users guide and reference manual.

Retrieved from http://mc-stan.org

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. doi:10.1214/14-ss105

Wang, L. (2010). IRT–ZIP modeling for multivariate zero-inflated count data. *Journal of Educational and Behavioral Statistics*, *35*(6), 671–692. doi:10.3102/1076998610375838

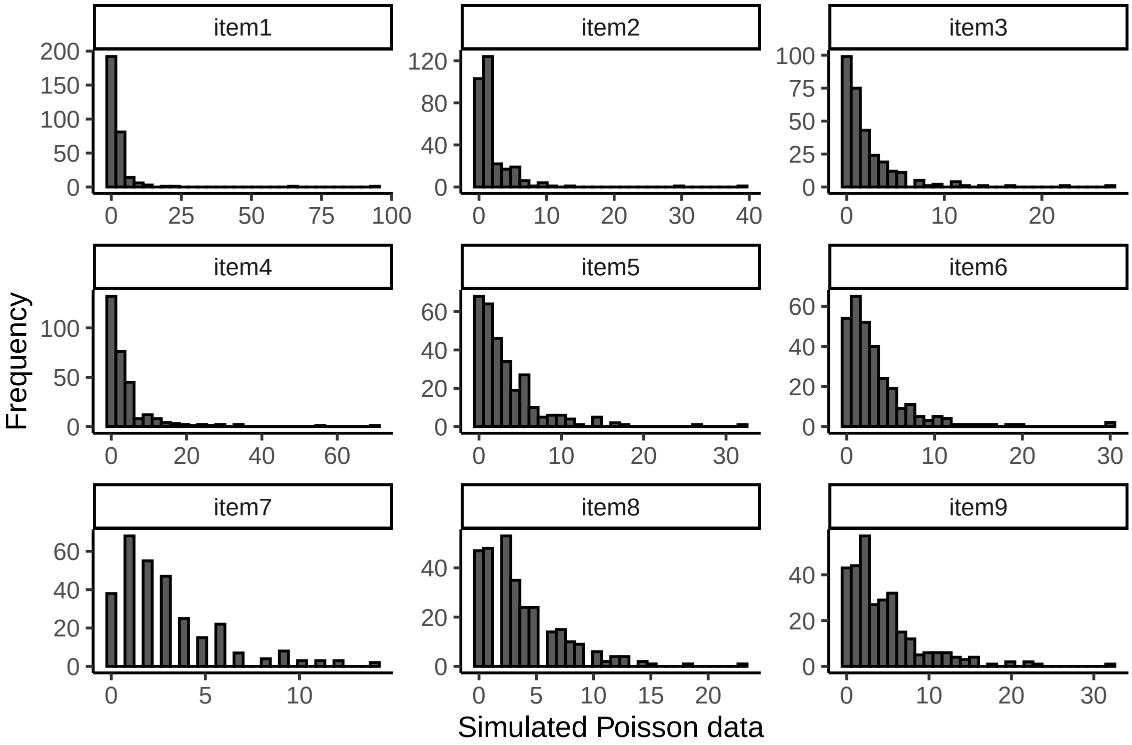

Figure 1. Histograms of simulated data. The axes are different for each item.

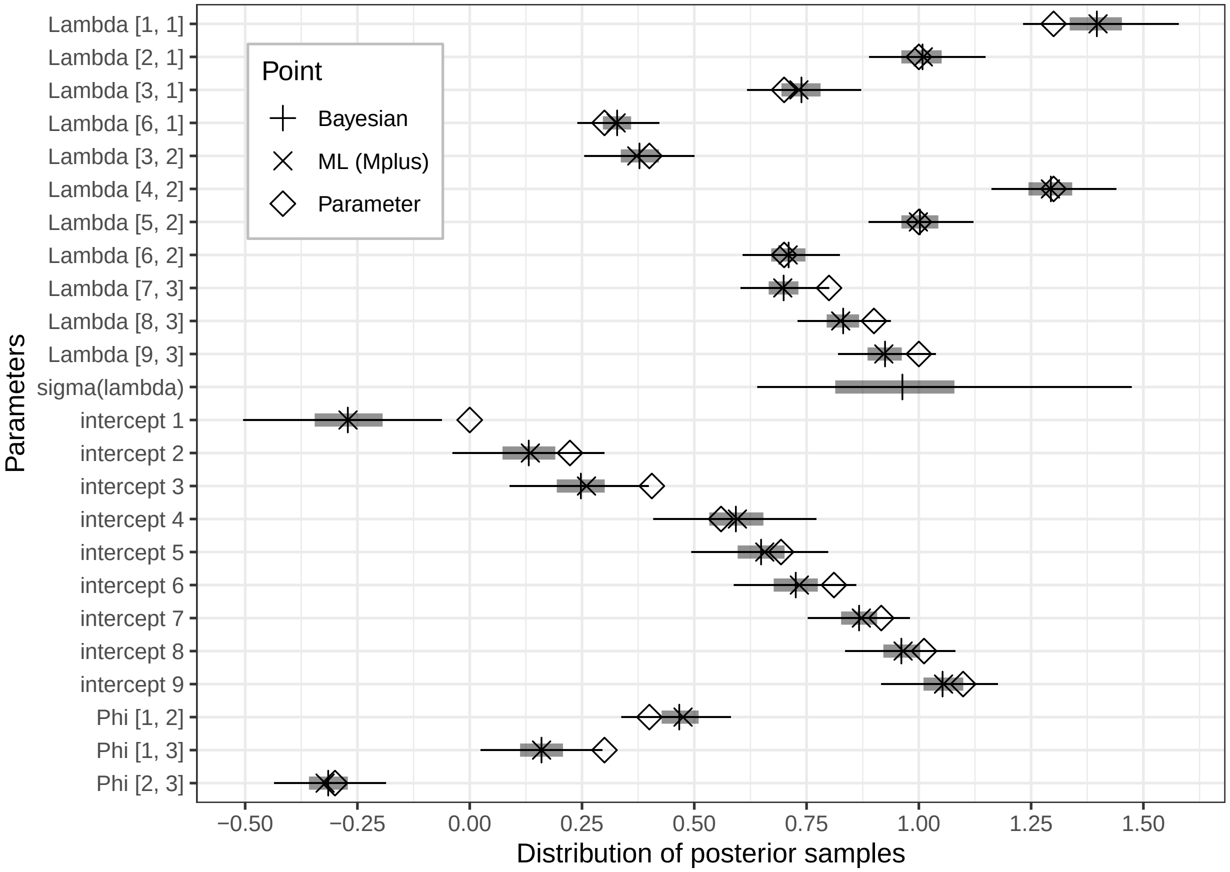

Figure 2. Result of Bayesian Poisson CFA applied to simulated data. The thick lines are the interquartile range, and the thin lines are the middle 95% of posterior samples. There is no population value for sigma(lambda) as it is a hyperparameter for Bayesian model estimation. We only provide the Mplus point estimate.

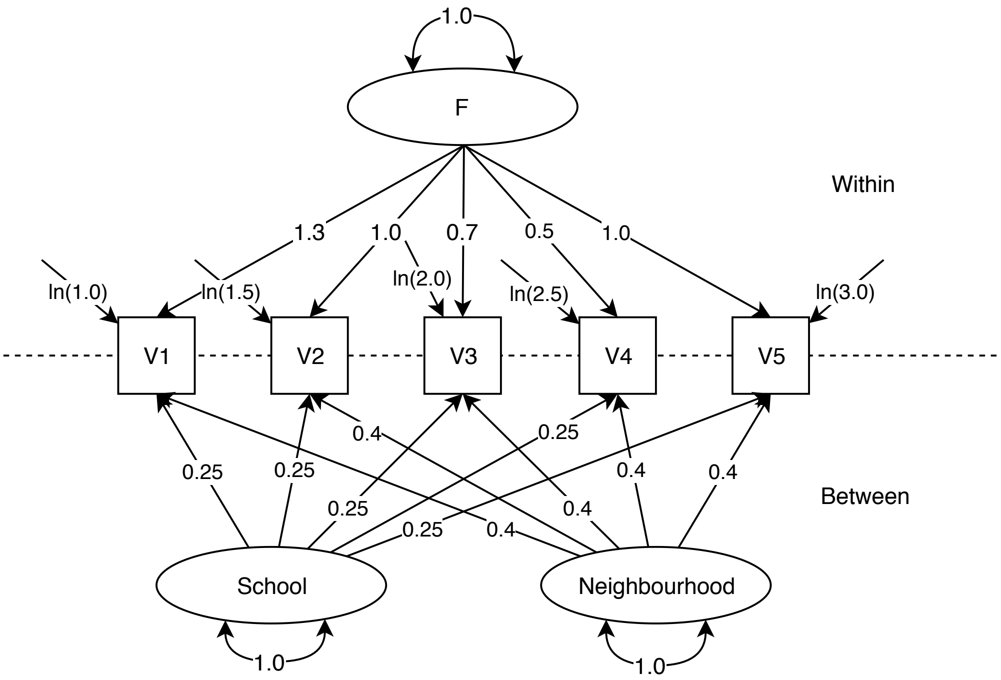

Figure 3. Cross-classified data structure for unidimensional construct, showing population parameters. The intercepts are at the person level, and we assumed the intercepts at the school and neighbourhood levels are zero.

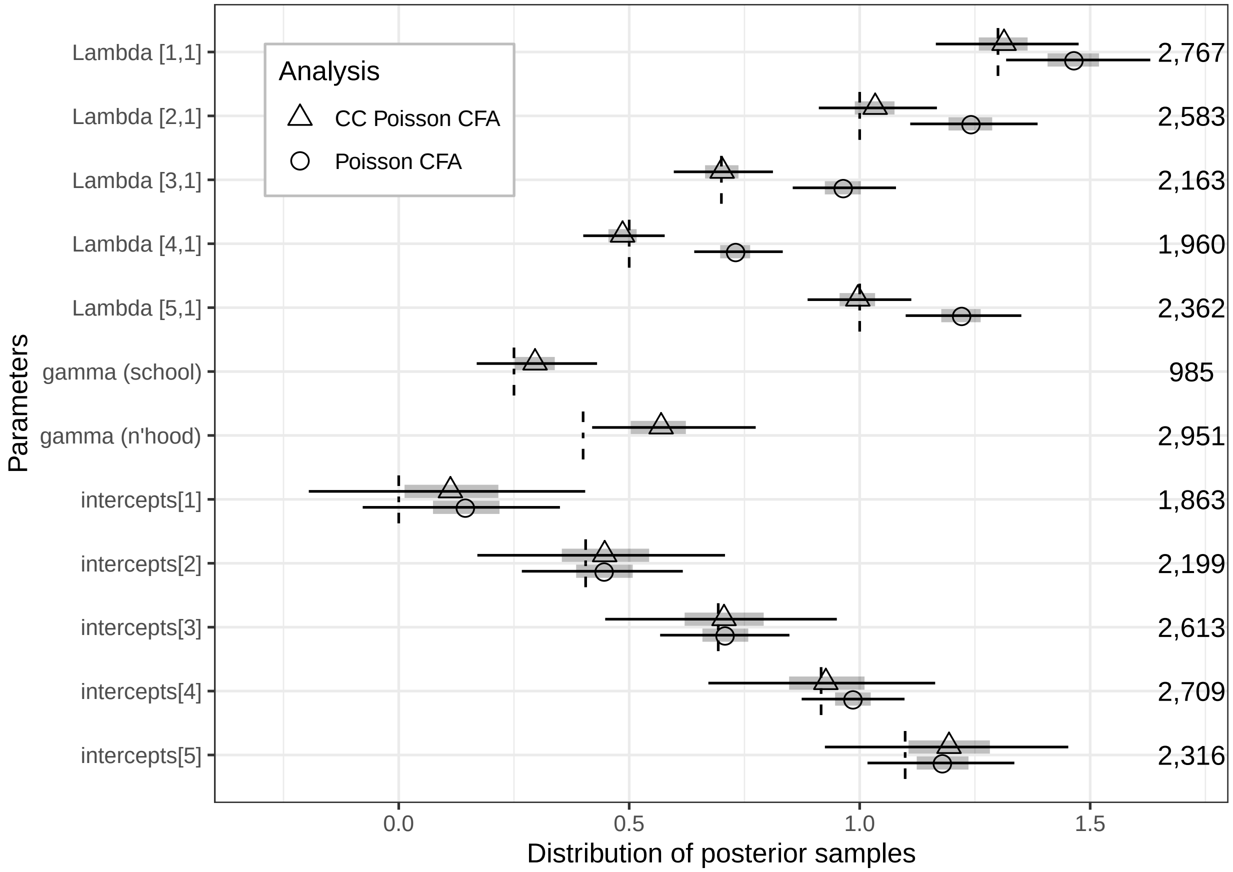

Figure 4. Result of single level and cross-classified (CC) Bayesian Poisson CFA applied to simulated data. The dashed vertical lines are population parameters. The thick lines are the interquartile range, and the thin lines are the middle 95% of posterior samples. There are no estimates for gamma (school) and gamma (n'hood) for the single level approach because the apparoach ignores these clusters. The numbers of the right hand side of the plot were the effective sample sizes for the parameters from the CC Poisson CFA. We have not reported the posterior samples of sigma (lambda) as its posterior intervals were fairly wide.