Model-based variance estimation for aggregated covariance structure models[1]

Abstract

Covariance structure models comprise a wide class of models, popular in the social and behavioral sciences and often applied to complex sample surveys. Variance estimation for such models has received relatively little attention. This short note introduces a model-based variance estimator under complex sampling for aggregated parameters of covariance structure models. This variance estimator can be used for three purposes: to assess sampling variance when the model is thought to be correct; as a working covariance matrix in GEE estimation; or to estimate "design" (or "misspecification") effects of nonnormality and clustering separately. A small simulation study indicates that the proposed estimator can accurately recover sampling variance, while an example confirmatory factor analysis demonstrates its use.

Key words: Structural equation modeling, covariance structure modeling, complex sampling, variance estimation, design effects.

[1] Daniel L. Oberski, Department of Methodology and Statistics, Tilburg University, 5000LE Tilburg, The Netherlands. E-mail: doberski@uvt.nl.

1. Introduction

Covariance structure modeling comprises a very general class of models, including but not limited to (multivariate) regression, factor analysis, path analysis, structural equation models, heritability models for twin data, growth curve models, and cross-lagged panel models with possible latent variables (Bollen, 1989; Kline, 2011). Such models are often applied to complex sample survey data, and it well-known that in such cases the standard variance estimators need adjustment to allow for correct inference. For an overview of these and other issues arising in covariance structure modeling of complex samples, see Bollen, Tueller, & Oberski (2013).

Under multi-stage sampling, parameters that are either aggregated (marginal) or disaggregated (conditional) over clusters may be of interest (Skinner, Holt, & Smith, 1989, pp. 8-10). Muthén & Satorra (1995) discussed design-based variance estimation for aggregated covariance structure model parameters under complex sampling (see also Skinner, Holt, & Smith, 1989, chapter 3) as well as model-based variance estimation for disaggregated parameters. Currently, however, no model-based variance estimators for aggregated parameters of covariance structure models are available. This short note provides such a variance estimator.

There are three motivations for providing a model-based variance estimator for aggregated parameters.

First, such an estimator is likely to be more accurate when the model holds (Wu & Kwok, 2012). Although such advantages will depend on model

correctness (Skinner & de Toledo Vieira, 2007), they may be considerable when the model is reasonable, particularly in small samples.

Second, even when the model does not hold, the variance estimator presented here can be used as a working covariance matrix in GEE estimation. Asymptotically optimal estimation weight matrices are already available, but known to yield a high mean square error in finite samples (de Toledo Vieira & Skinner, 2006; Oberski, 2013a); on the other hand, the independence working matrix does not take clustering effects into account at all. The variance estimator developed here compromises between these two extremes.

Third, a variance estimator that assumes normality but also accounts for clustering is needed when evaluating misspecification ("design") effects (Skinner et al., 1989, chapter 2) of nonnormality and clustering separately (Oberski, 2013b). Currently available variance estimators cannot separate these two effects on covariance structure model parameters from one another, since they account for both clustering and nonnormality or for neither.

Section 2 describes covariance structure analysis on data with normally distributed cluster effects. Section 3 then presents the standard nonparametric estimator for the variance of the model parameter estimates as well as the model-based estimator introduced by this note. A small simulation study demonstrating the validity of this estimator under the model assumptions is conducted in Section 4, after which an example application is presented in Section 5.

2. Definitions

Suppose a vector y of variables is observed on a multistage sample with C clusters (primary sampling units) from a population, where the *i*-th observation in the *c*-th cluster is assumed to follow the model

$$y_{ic} = \mu_c + \epsilon_{ic},$$

with $\mu_c \sim N\left(\mu, \Sigma^{(\mathrm{b})}\right)$, and $\varepsilon_{ic} \sim N(0, \Sigma^{(\mathrm{w})})$. For nonzero within- and between-cluster covariance matrices $\Sigma^{(\mathrm{w})}$ and $\Sigma^{(\mathrm{b})}$, observed data y_i follow a known but nonnormal distribution, with pooled (aggregated) population covariance matrix Σ, say. Although in some applications the "disaggregated" parameters of $\Sigma^{(\mathrm{w})}$ and $\Sigma^{(\mathrm{b})}$ may be of interest, we treat the case in which the "aggregated" parameters of the overall covariance matrix Σ are of interest.

This aggregated covariance matrix is parameterized by a covariance structure model Σ = Σ(θ). Given a consistent sample estimate of the pooled covariance matrix, S, say, this model is fitted by minimizing a fitting function $F\left(\mathrm{S}, \Sigma(\theta)\right)$, yielding sample parameter estimates, $\hat{\theta} = \mathrm{argmin}_{\theta}\, F\left(\mathrm{S}, \Sigma(\theta)\right)$. Note that the estimation of S may involve sampling weights. The most commonly used point estimator for θ is obtained by minimizing the normal-theory maximum likelihood fitting function F_{ML} under multivariate normality, but other choices are also possible. Notably, weighted least squares F_{WLS} with an appropriately chosen weight matrix will yield asymptotically optimal estimates (Browne, 1984) that nevertheless have performed badly in finite sample applications (e.g. Vieira & Skinner, 2008). Regardless of the choice of

fitting function, important matrices are the Jacobian $\Delta = \partial\sigma/\partial\theta$ and the hessian matrix $V = \partial^2 F/2\partial\sigma\partial\sigma'$, where σ is the half-vectorization of Σ, i.e., $\sigma = \text{vech}(\Sigma)$.

3. Variance estimation for aggregated covariance structure models

In general the asymptotic variance of the parameter estimates has a familiar "sandwich" form,

$$\text{avar}(\hat{\theta}) = (\Delta' V \Delta)^{-1} \Delta' V \Gamma V \Delta (\Delta' V \Delta)^{-1},$$

where $\Gamma = \text{var}(s)$, and s the half-vectorization of S (Satorra, 1989). This result can be derived by noting that $\partial\hat{\theta}/\partial s = [\Delta' V \Delta + o(\sqrt{n})]^{-1} \Delta' V$ (Oberski, 2013a) and applying the linearization variance formula of Wolter (2007, eq. 6.2.2). A consistent sample estimate of the variance can be obtained by replacing the parameter values by their sample estimates in V and Δ, expressions that may be obtained for a very general class of covariance structure models from Neudecker & Satorra (1991).

The sampling variance Γ of the non-redundant covariances then remains to be estimated. Commonly, the nonparametric estimator is used (Muthén & Satorra, 1995, p.. 285-7; Skinner et al., 1989, p. 47-9),

$$\hat{\Gamma}_{\text{clus, NP}} = n^{-1} \sum_{c=1}^{C} \frac{C}{(C-1)} (d_c - s)(d_c - s)',$$

where n is taken to be the sum of the weights instead of the sample size, $d_c = \sum_{i=1}^{n_c} w_i \, \text{vech}[(y_i - \overline{y})(y_i - \overline{y})']$, and $\overline{y} = n^{-1} \sum_{i=1}^{n} w_i y_i$. This estimator

accounts for both the complex sampling design and any nonnormality. Under the model described above, however, it is possible to derive an alternative variance estimator that accounts for the complex sampling design but is derived under the normality assumptions outlined above.

Under the model, it follows from standard results on the normal distribution that the variance of the observed covariances will equal

$$\Gamma_{\text{clus, NT}} = 2\mathrm{D}^{+}\left[C^{-1}\left(\Sigma^{(\mathrm{b})}\otimes\Sigma^{(\mathrm{b})}\right) + n^{-1}\left(\Sigma^{(\mathrm{w})}\otimes\Sigma^{(\mathrm{w})}\right)\right]\mathrm{D}^{+\prime},$$

where $D^{+}$is the Moore-Penrose inverse of the duplication matrix (Magnus & Neudecker, 2007). A consistent (maximum-likelihood) sample estimator can be obtained by estimating $\Sigma^{(\mathrm{w})}$ and $\Sigma^{(\mathrm{b})}$as

$$S^{(w)} = C^{-1}\sum_{c=1}^{C}(s_c - \overline{s}_{\cdot})^2 \qquad \text{and} \qquad S^{(b)} = C^{-1}\sum_{c=1}^{C}s_c,$$

where $s_c = {n_c}^{-1}\sum_{i=1}^{n_c}(y_{ic} - \overline{y}_{\cdot c})^2$, and $\overline{y}_{\cdot c}$ is a design-consistent estimate of $E(y)$ that may involve sampling weights. Note that it is possible to replace ${n_c}^{-1}$ by $(n_c - 1)^{-1}$ and C^{-1} by $(C-1)^{-1}$ in these estimators to obtain unbiased sample estimates, although the maximum likelihood divisors used here are more common to covariance structure analysis.

4. Small simulation study

The performance of the proposed variance estimator depends entirely on whether it adequately estimates the variance under the model specified. A small simulation study was therefore conducted to evaluate this performance. In the design of this simulation, the following factors were manipulated:

- Number of primary sampling units (PSU's) or clusters:
  $C \in \{10, 25, 50, 100\}$;
- Number of observations per cluster (2SU's):
  $n_c \in \{2, 5, 10, 25, 50, 100\}$;
- Amount of between-cluster heterogeneity, $\Sigma^{(b)} = \sigma^2$ I, with variance
  $\sigma^2 \in \{0, 0.1, 0.5, 1\}$.

For each of the resulting $4 \times 6 \times 4 = 96$ conditions, 1000 datasets were simulated from the model using R 3.0.1 (R Core Team, 2013). The 21 unique elements of the observed 6×6 variance-covariance matrix were then calculated together with their model-based asymptotic variance matrix $\hat{\Gamma}_{\text{clus, NT}}$ as proposed in the previous section. For comparison, the usual nonparametric asymptotic variance matrix $\hat{\Gamma}_{\text{clus, NP}}$ was also calculated for each sample.

Figure 1 shows the ratio of estimated to observed variance of the variances and covariances for different conditions. The number of PSU's (clusters) is plotted on the horizontal axis, the colors of the points indicate the amount of between-PSU heterogeneity, and the shape of the points

corresponds to the number of observations per PSU. Figure 1 shows that with a small number of PSU's and a large amount of heterogeneity, the standard nonparametric variance estimator somewhat underestimates the variance: the ratio of estimated to observed variance is 0.90. The model-based variance estimator performs much better in this condition, with a ratio close to unity. Increasing the number of PSU's this ratio remains stable near unity for the model-based estimator, although there appears to be a slight (3.5%) overestimation in some conditions. Overall, under the model assumptions, the model-based estimator appears to accurately reproduce the sampling variance of the variances and covariances.

To compare the relative performance of the nonparametric and model-based estimators when the model is indeed correct, Figure 2 reproduces the ratio of the root mean squared errors of both estimators. The horizontal axis of Figure 2 corresponds to the amount of cluster heterogeneity while the colors of the points correspond to the number of second-stage units. These two factors were chosen because a preliminary analysis indicated that they were the most influential on the root mean squared error ratio. The points are averages within these conditions. Figure 2 shows that the relative root mean squared error ranges from 5 to 1.2, meaning that, under model correctness, the model-based estimator is between 20% and 400% more accurate than the nonparametric estimator.

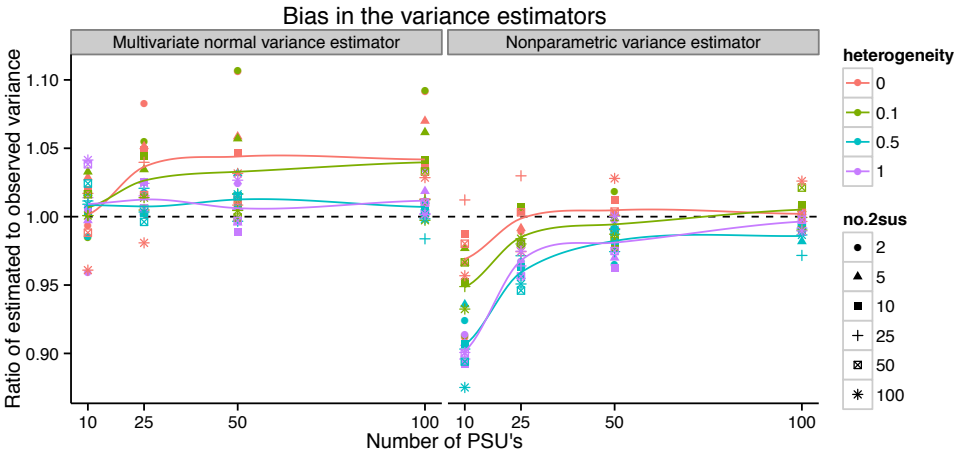

Figure 1: Ratios between estimated and simulation variance of estimates for different numbers of primary and secondary sampling units and between-cluster heterogeneity.

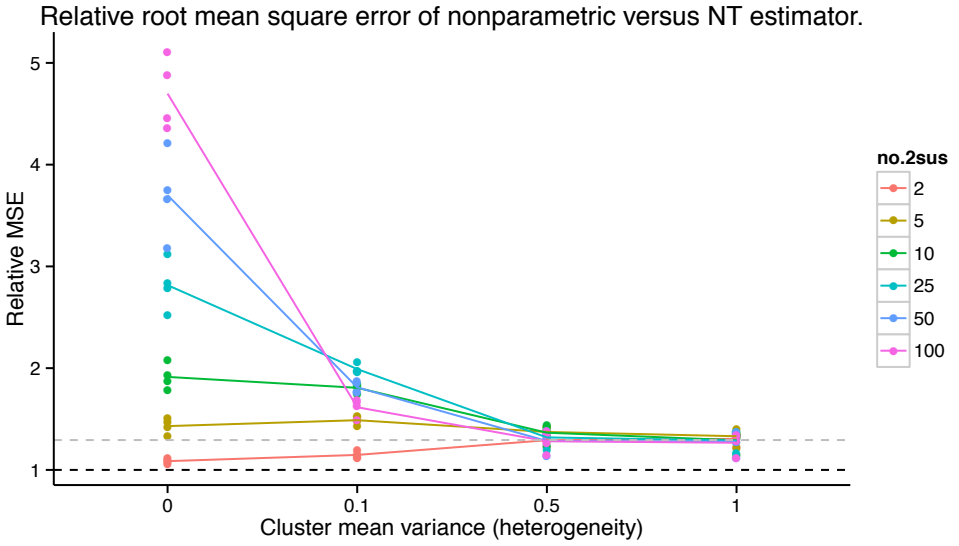

Figure 2: Comparison between the model-based and nonparametric variance estimators under model correctness, for different second-stage sample sizes and cluster heterogeneity values. Higher values indicate that the nonparametric variance estimator has a higher root mean square error.

5. Example application

Roosma, Gelissen, & Oorschot (2012) discuss a confirmatory factor analysis of respondents' attitudes towards the welfare state. They postulate a two-factor model for nine observed variables from the 2008 European Social Survey (see http://europeansurveyresearch.org for precise descriptions, variable names, and data). The original authors did not take into account the complex sampling design of the European Social Survey. Oberski (2013a) reanalyzed these data with the lavaan package in R, taking account of the complex sampling design in the UK data from the European Social Survey (R Core Team, 2013; Rosseel et al., 2013). Here we demonstrate the use of the variance estimator discussed above on these data to separately estimate clustering and nonnormality misspecification effects.

Table 1 gives the parameter estimates for the loadings, error variances, and factor (co)variances of this model. Standard errors accounting (or not accounting) for clustering and nonnormality are also given. The normal-theory standard errors that do take clustering into account are those introduced here.

Using all four combinations, "conditional misspecification effects" (Oberski, 2013b) can be calculated as the theoretical increase in standard error relative to the hypothetical situation without clustering or nonnormality. These estimates are shown in Table 2. The Table shows the average of conditional misspecification effects on standard errors for the three types of parameters. Of these types, it is well-known that loading and factor (co)variance estimates are highly correlated, while error variance parameters are independent of the other

types (Skrondal & Rabe-Hesketh, 2004). It is therefore unsurprising that the conditional misspecification effects are similar for loadings and factor (co)variances. Under normality, clustering increases standard errors by about 10% while without clustering, nonnormality increases standard errors by only about 5% for these parameters. For factor loadings, the misspecification effects are stronger: clustering under normality increases standard errors by 19% whereas nonnormality without clustering increases them by 25% -- this large relative increase is less, however, when clustering is already taken into account. Another way of seeing this is that when clustering is ignored, calculating standard errors that are robust to nonnormality will remove some, but not all, of the clustering effects.

Table 1: Parameter estimates and four types of standard errors for two-factor analysis of nine indicators of attitudes to the welfare state. NT: normal-theory, NP: nonparametric, clust.: clustered.

Parameter	Est.	s.e., *iid,* NT	s.e., *iid,* NP	s.e., clust., NT	s.e., clust., NP
Loadings					
range→gvhlthc	0.59	0.0323	0.0391	0.0391	0.0444
range→gvslvol	0.68	0.0342	0.0397	0.0408	0.0427
range→gvslvue	0.82	0.0470	0.0425	0.0466	0.0409
range→gvcldcr	0.91	0.0482	0.0463	0.0500	0.0474
range→gvpdlwk	0.87	0.0469	0.0468	0.0512	0.0481
goals→sbeqsoc	1.32	0.1218	0.1265	0.1361	0.1237
goals→sbcwkfm	0.92	0.0796	0.0857	0.0881	0.0897
Error variances					
gvjbevn↔gvjbevn	4.64	0.1650	0.1815	0.1878	0.1933
gvhlthc↔gvhlthc	1.39	0.0506	0.1016	0.0867	0.1144
gvslvol↔gvslvol	1.13	0.0465	0.0698	0.0598	0.0700
gvslvue↔gvslvue	3.32	0.1169	0.1406	0.1274	0.1488
gvcldcr↔gvcldcr	2.81	0.1059	0.1332	0.1220	0.1326
gvpdlwk↔gvpdlwk	2.84	0.1047	0.1552	0.1348	0.1704
sbprvpv↔sbprvpv	0.58	0.0251	0.0281	0.0291	0.0290
sbeqsoc↔sbeqsoc	0.52	0.0341	0.0358	0.0377	0.0376
sbcwkfm↔sbcwkf	0.49	0.0212	0.0230	0.0225	0.0222
Factor (co)variances					
range↔range	1.96	0.1664	0.1624	0.1804	0.1710
goals↔goals	0.19	0.0234	0.0255	0.0268	0.0293
range↔goals	-0.11	0.0218	0.0240	0.0233	0.0298

Table 1: Mean conditional misspecification effects of clustering and nonnormality, and their interaction.

	Conditional misspecification effect on standard errors		
Parameter type	Clustering	Nonnormality	Clustering × Nonnormality
Error variances	1.19	1.25	0.85
Loadings	1.10	1.05	0.93
Factor (co)variances	1.09	1.05	1.04

6. Concluding remarks

This note introduced a variance estimator for aggregated covariance structure models that accounts for clustering effects but does assume these effects follow a normal distribution. Such an estimator may be useful in at least three scenarios: 1) when the normality assumption is reasonable but aggregated parameters are of interest; 2) as an inverse estimation weight matrix in GEE (WLS) estimation; and 3) when conditional misspecification effects of nonnormality and clustering are of interest separately. A small simulation showed that the variance estimator proposed provides accurate estimates under a variety of conditions, while an example application to a confirmatory factor analysis demonstrated the use of the variance estimator introduced here for the third purpose (misspecification effects estimation).

The purpose of this note was to introduce the model-based variance estimator, point out some possible applications, and demonstrate its feasibility. In the future, more evaluations of this estimator are necessary. Particularly, its robustness, or lack thereof, to violations of the normality assumption for purposes 1 and 2; and its performance in GEE estimation as compared with other complex sampling estimators for covariance structure modeling.

Acknowledgements

This work was supported by the Netherlands Organization for Scientific Research (NWO) [vici grant number 453-10-002].

References

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Bollen, K. A., Tueller, S., & Oberski, D. L. (2013). Issues in the Structural Equation Modeling of Complex Survey Data. In I. S. I. (ISI) (Ed.), *Proceedings of the 59th World Statistics Congress*. Hong Kong.

Browne, M. W. (1984). Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures. *Br J Math Stat Psychol*, *37*, 62–83.

De Toledo Vieira, M., & Skinner, C. J. (2006). *Estimating models for panel survey data under complex sampling*. Southampton, UK: University of Southampton. Retrieved from Retrieved from http://eprints.soton.ac.uk/42001/.

Kline, R. B. (2011). *Principles and practice of structural equation modeling, 3rd ed.* New York: The Guilford Press.

Magnus, J. R., & Neudecker, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics, Third Edition*. New York: John Wiley & Sons.

Muthén, B., & Satorra, A. (1995). Complex Sample Data in Structural Equation Modeling. *Sociological Methodology*, *25*, 267–316.

Neudecker, H., & Satorra, A. (1991). Linear Structural Relations: Gradient and Hessian of the Fitting Function. *Statistics and Probability Letters*, *11*, 57–61.

Oberski, D. L. (2013a). lavaan.survey: An R Package for Complex Survey Analysis of Structural Equation Models. *Journal of Statistical Software*.

Oberski, D. L. (2013b). Conditional Design Effects for Structural Equation Model estimates. In I. S. I. (ISI) (Ed.), *Proceedings of the 59th World Statistics Congress*. Hong Kong. Retrieved from http://daob.nl/wp-content/uploads/2013/04/hk-oberski.pdf

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

Roosma, F., Gelissen, J., & Oorschot, W. van. (2012). The Multidimensionality of Welfare State Attitudes: A European Cross-National Study. *Social indicators research*.

Rosseel, Y., Oberski, D. L., Byrnes, J., Vanbrabant, L., & Savalei, V. (2013). *lavaan: Latent Variable Analysis*.

Satorra, A. (1989). Alternative Test Criteria in Covariance Structure Analysis: A Unified Approach. *Psychometrika*, *54*(1), 131–151.

Skinner, C. J., & de Toledo Vieira, M. (2007). Variance Estimation in the Analysis of Clustered Longitudinal Survey Data. *Survey methodology*, *33*(1), 3–12.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Vieira, M. D. T., & Skinner, C. J. (2008). Estimating Models for Panel Survey Data Under Complex Sampling. *Journal of Official Statistics*, *24*(3), 343–364.

Wolter, K. (2007). *Introduction to Variance Estimation, Second Edition*. New York: Springer-Verlag.

Wu, J. Y., & Kwok, O. M. (2012). Using SEM to Analyze Complex Survey Data: A Comparison between Design-Based Single-Level and Model-Based Multilevel Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 16–35.