The Bayesian Covariance Structure Growth Model for Intensive and Longitudinal Data

Twan ten Hoopen

Cognition, Data and Education (CODE), University of Twente

202000384: BSc Thesis Psychology

First Supervisor: Dr. ir. G.J.A. Fox

Second Supervisor: Dr. R.H.J. van der Lubbe

June 25, 2024

Introduction

Growth Models

Big data is everywhere, and its benefits are numerous and multifaceted. It finds its application in various disciplines, such as economics, healthcare, and social sciences (Einav, L., & Levin, J., 2014; Batko, K., & Ślęzak, A., 2022; Hesse et al., 2015). One way to analyze big data is through the use of growth models. Growth models are theoretical frameworks that describe the mechanisms and factors driving the increase or decrease of an entity over time. Through data-driven insights, growth models can use big data to identify patterns, predict trends, and optimize strategies. Furthermore, growth models allow for a more detailed characterization of the trajectory in big and regular data compared to more traditional data analytic approaches, such as repeated measures analysis of variance (RM-ANOVA) or multiple regression analysis.

To elaborate, unlike traditional data analysis methods, growth models do not view individual variance simply as error variance. Instead, growth models interpret it as reflecting the extent of change occurring at the individual level. To elaborate, as each individual has its own growth trajectory, any variance can be explained by combining or examining individual growth trajectories. This allows for greater insight into the data observed. Also, methods such as RM-ANOVA, as part of traditional data analytics, work with strong analytic assumptions, such as the sphericity assumption. However, growth models do not rely as much on robust analytical premises as traditional data analytics. For example, a growth model might employ a Bayesian estimation framework to iteratively update probabilities and parameters based on new data and prior beliefs, allowing for flexible adaptation to changing conditions without requiring strict adherence to assumptions like normality or equal variance. Therefore, growth models can offer more flexibility and robustness than conventional data methodology.

The Linear Mixed Effect Model

To analyze intensive longitudinal data, multiple approaches can be used. One of them is the linear mixed effect (LME) model. The LME model employs maximum likelihood estimation. However, this model does not allow the random factor variance to get close to zero (Klotzke & Fox, 2019a). When this happens, a multitude of problems arise, one of them being numerical instability. To elaborate, let's take the example of EEG data. When dealing with this form of data, it might be speculated that the relevant (brain) signal has too little variance. When analyzing EEG data, researchers typically aim to detect and analyze relevant brain activity against a background of noise, such as electrical interference. In this context, the low amplitude of the EEG data's relevant brain signal presents a challenge. Due to its insufficient variance, the LME model may struggle to accurately capture or account for these minimal variations, potentially leading to an overestimation of the relevant signal. As a result, the LME model can induce false positives or type I errors (Murphy et al., 2024).

Secondly, the fit of the LME model can be hampered by convergence problems in the estimation method, which in practice often leads to the conclusion that the random factors with small variances are not relevant. To elaborate, convergence problems happen when the algorithm of the LME model fails to estimate the model's parameters as the LME model has trouble dealing with low variance, as explained before. As a result, when retaking the example of EEG data, the LME model is not capable of accurately estimating the appropriate, relevant signal.

Lastly, growth models often work with random effects to accurately capture individual differences in growth trajectories, account for the hierarchical structure of the data, and improve the precision and generalizability of the model's estimates by allowing for subject-specific variations. However, when dealing with big data, these random effects enlarge the computational demands for fitting the model. To elaborate, random effects add extra parameters to estimate, which can prolong the computational duration of the LME model and raise its memory demands, especially when dealing with big data.

Bayesian Covariance Structure Model

When considering the limitations of the LME model, it becomes imperative to utilize a new method, capable of accounting for the shortcomings of the LME model. This model is called the Bayesian Covariance Structure Model (BCSM), as it introduces prior distributions on the parameters of the covariance matrix (Klotzke & Fox, 2019a). This model has numerous answers to the problems discussed in the LME growth model. Firstly, the BCSM can include dependences implied by random effects without relying on random effect parameters, as it models the covariance structure directly (Klotzke & Fox, 2019b). To elaborate, modeling the covariance matrix directly means that instead of modeling individual variables or their means separately, the focus is on the relationships and dependencies between the variables as captured by the covariance matrix. This makes the approach significantly more efficient in terms of the number of parameters and the associated computational demands. Also, modeling the covariance structure directly is necessary to capture dependences among variables within and across clusters, referred to as clustering. This is important as clustering improves model accuracy and enhances interpretability. Clustering refers to grouping data points so that data in the same group (or cluster) are more similar than those in other groups or clusters.

Secondly, the BCSM allows for more flexibility compared to the LME model, as it allows for negative and close to zero covariances, due to the covariance matrix being modeled directly. To elaborate, one critical assumption of the LME models is that observations within clusters are positively correlated. However, relationships within clusters can be negatively correlated, meaning that an increase in one observation is associated with a decrease in another (Nielsen et al., 2021).

Thirdly, the BCSM does not rely on normally distributed random factors, as with the LME model (Fox et al., 2020). This is especially important since the assumption of normality can be problematic when dealing with big data (Ghasemi & Zahediasl, 2011). To elaborate, random effects may deviate from normality due to various factors such as heterogeneity, skewness, or heavy tailing in the data. Since the BCSM does not incorporate random effects, its model parameters remain unaffected by the number of data trials or participants. This feature is another advantage of the BCSM compared to the LME model.

Fourthly, as the BCSM does not rely on random effect parameters, the computational demands of running the model are reduced. To elaborate, through the use of Markov Chain Monte Carlo (MCMC), the BCSM can allow for parallel computing. This is especially useful for big data as it allows for the distribution of the computational load across multiple processors or machines. As a result, the BCSM scales well with data size and is better than the LME model that does not employ the MCMC. For regular data, the decreased computational demands of the BCSM still apply, although to a lesser extent than with bigger data. To elaborate, parallel computing, facilitated by MCMC, can still be beneficial for regular-sized data, enabling more efficient use of computational resources and model fit. Furthermore, the BCSM does not estimate an intercept and slope for each individual data point. Instead, it estimates a single intercept and slope for the entire dataset based on the underlying probability distributions and the data.

Lastly, when it comes to big data, the BCSM offers another advantage compared to the LME model. The BCSM is capable of handling missing data more naturally by integrating missing values. This is done through the framework of Bayesian statistics. To elaborate, in Bayesian analysis, missing data are treated as unobserved variables, and the model estimates their distributions along with the parameters of interest based on the observed data and prior information. This is important as big data often has missing values, leading to biased analyses and compromised decision-making if not handled appropriately (Basiri & Brunsdon, 2022; Emmanuel et al., 2021). Therefore, it is important to test the viability of the BCSM on big data, considering the advantage the BCSM has when it comes to missing data often present in big data.

Simulated Data

One way to compare the BCSM and LME model, is to test them against simulated data that includes both an intercept and a slope. The intercept is the value at which the regression line crosses the y-axis when all predictor variables are zero. The slope is the rate of change in the response variable, for a one-unit increase in the predictor variable. Incorporating a slope and intercept enables the measurement of the effects of lower levels of random factor variance, thereby allowing the flexibility and performance of both models to be evaluated. By controlling and manipulating the variances of these random effects in simulations, one can systematically explore how well each model captures this variability.

The Aim of This Thesis

The current study aims to test the BCSM against an already existing framework; the LME model. It is hypothesized that the BCSM is more capable of handling decreasing levels of random factor variance than the LME model, considering all the advantages the BCSM has over the LME model as mentioned in this introduction. Therefore, growth modeling applications of the BCSM and the LME model are examined in the light of decreasing random factor variance.

Furthermore, growth modeling applications of the BCSM on big data are also examined in the light of decreased random factor variance.

Methods

Introduction Linear Mixed Effects Model

When referring to the LME model, the following formula is used: $Y_{ti} = \beta_0 + \beta_1(TIME_{ti}) + b_{0i} + b_{1i}(TIME_{ti}) + \varepsilon_{ti}$ *with* $\varepsilon_{ti} \sim N(0,\sigma^2)$. Furthermore, $b_{0i} \sim N(0, \tau_0)$ and $b_{1i} \sim N(0,\tau_1)$. To elaborate, Y_{ti} stands for the response or dependent variable for individual i at time t, β_0 stands for the fixed intercept, which is the value of the response variable when time is zero, β_1 is the fixed slope indicative of the change in the response variable per unit change in *TIME*, $TIME_{ti}$ is the variable time for individual i at time t. To elaborate, $TIME_{ti}$ denotes a particular time point or observation within the time series t for individual i. It represents a specific measurement in time chosen to extract an individual's dependent variable.

When looking at the random effects, b_{0i} is the random intercept for individual i, and b_{1i} is the random slope for individual i. It allows each individual to have their own change in the response variable while considering $TIME_{ti}$. For the random effect b_{0i}, assumed is a normal distribution with a mean of zero and a random factor variance denoted as τ_0. The random effect b_{1i} is also assumed as a normal distribution with a mean of zero and a random factor variance denoted as τ_1. Lastly, ε_{ti} is the residual error for individual i at time t. It is assumed to be normally distributed with a mean of zero and variance σ^2. The variance of the residuals represents the individual-specific random errors.

The Bayesian Covariance Structure Model

Unlike the LME model, the key difference and core idea of the BCSM is to model the covariance matrix directly, thereby excluding the need to include random effects in the mean part of the model. However, although random effects are not part of the model's framework, these effects can still be incorporated through the covariance matrix. Without random effects, an increase in the level of participants does not increase the model's complexity in the BCSM, thereby reducing the risk of overfitting.

Simulation Study Design

First Simulation Study

The simulation study design consisted of two parts: simulation study one and simulation study two. In the first simulation study, the BCSM was compared to the LME model through a simulated dataset consisting of n = 1,000, number of data clusters, and k = 10, number of responses per subject. The number of data replications was set to 1,000 for both simulation studies. In both simulation studies, τ_0 represented the random intercept variance and τ_1 represented the random slope variance. The σ^2 in both simulation studies was equal to 1, τ_0 to 0.5, and τ_1 to 0.2. Summary statistics extracted from the first data simulation study included the mean, the bias, and the root-mean-square error (RMSE). The bias refers to the systematic error in parameter estimation, indicating the average difference between estimated and true values. Furthermore, the RMSE is a measure of the average deviation between estimated and true values, calculated by taking the square root of the average of squared differences between estimated and true values. For the BCSM, the posterior mean was also extracted, just as the bias and the RMSE. For the bias and the RMSE, the lower the value, the better the parameter estimate.

Second Simulation Study

The sole performance of the BCSM was evaluated on a larger sample size. The dataset had n = 1,000 clusters with cluster size k = 1,000. The covariance parameters were estimated by the posterior mean. Furthermore, the RMSE and bias were also used to evaluate the parameter estimate. As with the first simulation study, for both the RMSE and bias, the lower the value, the better the estimates. Next to random effect variances, fixed effects were also estimated for both the β_0 and β_1 coefficients. The bias and RMSE were examined for each fixed effect coefficient and corresponding τ value.

Simulation Procedure

For both the first and second data simulations, values of τ_0 were reduced starting from 0.5 to 0.2, 0.1, 0.05, and 0.01. When doing so, τ_1 was set to 0.2. For τ_1, also in both data simulations, values were reduced starting from 0.2 to 0.1, 0.05 and 0.01. When doing so, τ_0 was set to 0.5 Then, the estimates of τ_0 and τ_1 were reported for each simulation study and each τ value. Lastly, for the big data simulation study, fixed effects estimates were also reported. The software used to analyze both simulated datasets was R version 4.3.3[1].

Results

For the first simulation study, the LME approach and the BCSM were applied. For the second simulation study, solely the BCSM was applied to analyze the dataset. The estimation method for the LME model was maximum likelihood estimation and for the BCSM it was a Bayesian estimation.

First Simulation Study

[1] The R code used originated from numerous R libraries and was part of the BCSM code developed by G.J.A. Fox

To start, the results of the LME model and the BCSM of the first data simulation study are depicted in Table 1.

Table 1

The τ_0 Estimation Under the LME and BCSM of the First Simulation Study

τ_0	Linear Mixed Effects Model			Bayesian Covariance Structure Model		
	Mean	Bias	RMSE	Mean	Bias	RMSE
0.5	0.500	0.000	0.029	0.503	0.003	0.038
0.2	0.200	0.000	0.016	0.202	0.002	0.025
0.1	0.101	0.001	0.012	0.102	0.002	0.021
0.05	0.051	0.001	0.009	0.051	0.001	0.018
0.01	0.011	0.001	0.007	0.011	0.001	0.017

Note. $n = 1{,}000$, $k = 10$, $\tau_1 = 0.2$

When examining the LME results in Table 1, when τ_0 decreased, the bias remained more or less the same. Furthermore, for LME, the RMSE decreased when the τ_0 parameter value also decreased. Unlike LME, reducing τ_0 in BCSM led to more accurate estimations of the true τ_0 value through the posterior mean. Consequently, for BCSM, the bias decreased. Furthermore, for BCSM, the RMSE also decreased when lowering the τ_0 parameter value. However, the RMSE of BCSM decreased less rapidly than the RMSE of the LME.

Lastly, it is also important to take a look at the absolute numbers. For LME, lower absolute values in RMSE were observed compared to BCSM. Furthermore, posterior mean estimates of BCSM were at $\tau_0 = 0.05$ and onward just as capable of estimating τ_0 as the maximum likelihood estimate of LME.

Table 2

The τ_1 Estimation Under the LME and BCSM of the First Simulation Study

τ_1	Linear Mixed Effects Model			Bayesian Covariance Structure Model		
	Mean	Bias	RMSE	Mean	Bias	RMSE
0.2	0.196	-0.004	0.046	0.202	0.002	0.064
0.1	0.097	-0.003	0.042	0.102	0.002	0.060
0.05	0.050	-0.001	0.037	0.052	0.002	0.058
0.01	0.020	0.010	0.027	0.012	0.002	0.057

Note. $n = 1.000$, $k = 10$, $\tau_0 = 0.5$

Looking at the LME results in Table 2, when τ_1 was reduced, the more likely the maximum likelihood estimate could estimate τ_1 up to $\tau_1 = 0.01$. There, for the LME, the bias was five times higher than for the BCSM. When decreasing τ_1, for BCSM, the bias was more consistent than for LME. As a result, the posterior mean was more consistently capable of estimating lowering levels of τ_1. Lastly, for both LME and BCSM, RMSE decreased when τ_1 also decreased.

When looking at absolute numbers, the posterior mean estimate of τ_1 was more accurate than the maximum likelihood estimate of LME, especially noticeable for $\tau_1 = 0.01$. However, for LME, absolute RMSE values were lower than for BCSM.

Second Simulation Study

Now we look at the estimates of the second simulation study as provided in Tables 3 and 4.

Table 3

The τ_0 and β_0 Estimation Under the BCSM of the Second Simulation Study

τ_0	BCSM			β_0 Coefficient	
	Mean	Bias	RMSE	Bias	RMSE
0.5	0.499	-0.001	0.023	-0.001	0.023
0.2	0.200	0.000	0.010	-0.001	0.015
0.1	0.099	0.000	0.005	0.000	0.011
0.05	0.050	0.000	0.003	0.000	0.008
0.01	0.010	0.000	0.001	0.000	0.004

Note. $n = 1{,}000$, $k = 1{,}000$, $\tau_1 = 0.2$

When looking at the estimates in Table 3, the posterior mean estimates were close to the true τ_0 parameter values. Moreover, as τ_0 decreased, the RMSE also decreased, resulting in the posterior mean becoming more accurate in estimating τ_0. When looking at the β_0 coefficient, as with τ_0, β_0 estimates were also close to the true β_0 values. Furthermore, for β_0, when τ_0 decreased, the RMSE and bias decreased also.

Table 4

The τ_1 and β_1 Estimation Under the BCSM of the Second Simulation Study

τ_1	BCSM			β_1 Coefficient	
	Mean	Bias	RMSE	Bias	RMSE
0.2	0.200	0.000	0.020	0.000	0.014
0.1	0.100	0.000	0.014	0.000	0.010
0.05	0.050	0.000	0.010	0.000	0.008
0.01	0.010	0.000	0.006	0.000	0.005

Note. $n = 1{,}000$, $k = 1{,}000$, $\tau_1 = 0.5$

Looking at Table 4, the posterior mean estimated the differing τ_1 values extremely accurately. No bias was observed regarding the different τ_1 estimates in BCSM when rounding off to three decimals. Also, as the variance of the τ_1 parameter decreased, the RMSE also decreased. Lastly, the BCSM showed no difficulty estimating the fixed effect of the β_1 coefficient.

When comparing the second simulation study to the first simulation study, the next notable differences were found. First of all, the posterior mean of BCSM in simulation study two was better capable of estimating the parameter values of τ_0 and τ_1 than the posterior mean was in the first simulation study. The same result was observed for RMSE between the different simulation studies and τ estimations.

Discussion

Thesis Conclusion

This study examined the performance of the BCSM and LME model in light of decreased random factor variances, including fixed effects and big data. It was proposed that the BCSM performs better than the LME model when random factor variance is decreased. When looking at the results, the following conclusions are drawn. The LME model is less effective at estimating the random slope when the variance of the random factor is minimal, compared to the BCSM. Furthermore, the BCSM is more consistent in the estimation of the random slope than the LME model. Additionally, the RMSE of the BCSM decreases more slowly than the RMSE of the LME model. However, the magnitude of the errors generated by the LME model is smaller than those of the BCSM when random factor variance is decreased. Looking at big longitudinal data, the bias and RMSE of the BCSM decreases compared to conventional data. However, the limitations of this study must be considered when interpreting these results.

As expected, the LME model and BCSM perform better with reduced random factor variance, as both models are less sensitive to the variability introduced by random effects. However, for the LME model, compared to the BCSM, the estimation of the random slope variance was less accurate due to convergence problems when the random factor variance was minimal. Lastly, the BCSM also performs adequately when it comes to big data and lower levels of random factor variance.

Limitations of the Study

When looking at the limitations, numerous constraints arise. First of all, for the BCSM, only random factor variances of lower-order polynomial terms were examined. One could wonder how the model is capable of estimating random factor variances in cubic and quadratic polynomial terms (Lim & Ziegler, 2023). This is important as especially growth models tend to include higher-order polynomial terms. A second limitation is that the random factor variance is only lowered to 0.01 due to the number of data replications performed in this study. A higher number of data replications will allow the BCSM to be examined against a random factor variance lower than 0.01. As a result, the performance of the BCSM to estimate even lower random factor variances can be examined, which can be useful for EEG data (Murphy et al., 2024). Lastly, although the BCSM performs better on big data than on regular data, it is important to understand that more data on itself already leads to better estimation of the random factor variance.

Both the LME model and the BCSM perform well, showing acceptable levels of bias and RMSE. The lower RMSE observed in the LME model as random factor variance decreases could be misleading due to convergence issues. To elaborate, negative random factor variances are assumed to be zero leading to the underestimation of the RMSE in the LME model. The BCSM, on the other hand, shows higher RMSE because it accounts for these negative random factor variances, thereby assuming more variance in the data. This explains the overall performance differences between the two models when it comes to the RMSE.

When looking at the BCSM. when random factor variance decreases, the probability of incorrectly rejecting the null hypothesis when it is true is reduced (Murphy et al., 2024). This is because when random factor variance decreases, less variability in the random effects is captured by the model leading to a more reliable estimation of the fixed and random effects. The BCSM does not suffer from the same convergence problems as the LME model. Therefore, in the BCSM, the estimations of the random factor variance become both more accurate and consistent when the random factor variance is lowered. Lastly, when looking at fixed effects, the decrease in RMSE can be explained by the fact that uncorrelated data is more informative than correlated data. Therefore, as random factor variance decreases, data becomes more informative, resulting in smaller RMSE for the fixed coefficients.

Suggestions for Future Research

When looking at future studies, the BCSM needs to be tested against higher-order polynomial terms and see how well these polynomial terms can be estimated in the light of decreasing random factor variance. Furthermore, a higher number of data replications can generate a higher precision, thereby extending the accuracy of random factor variance estimations to third decimals, and if necessary, even further than that. By doing this, data characterized by even lower random factor variance can still be accurately examined, as can also be the case with EEG data (Murphy et al., 2024).

In the end, this report states that the BCSM is more capable than the LME model when dealing with lowering degrees of random factor variance. Moreover, one can also say that the BCSM is a capable model for estimating fixed and lowering degrees of random factor variance when dealing with intensive longitudinal data.

References

Basiri, A., & Brunsdon, C. (2022). Missing data as data*.* *Patterns, 3*(9), 100587. https://doi.org/10.1016/j.patter.2022.100587

Batko, K., & Ślęzak, A. (2022). The use of Big Data Analytics in healthcare. *Journal of Big Data, 9*(1), 1-24. https://doi.org/10.1186/s40537-021-00553-4

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy, 14*(1), 1-24. https://doi.org/10.1086/674019

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data, 8*(1), 1-37. https://doi.org/10.1186/s40537-021-00516-9

Fox, J.P.., Koops, J., Feskens, R., & Beinhauer, L. (2020). Bayesian covariance structure modeling for measurement invariance testing. *Behaviormetrika 47*, 385–410. https://doi.org/10.1007/s41237-020-00119-3

Ghasemi, A., & Zahediasl, S. (2011). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486-489. https://doi.org/10.5812/ijem.3505

Hesse, B. W., Moser, R. P., & Riley, W. T. (2015). From Big Data to Knowledge in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science*. https://doi.org/10.1177/0002716215570007

Klotzke, K., & Fox, J. P. (2019a). Bayesian Covariance Structure Modeling of Responses and Process Data. *Frontiers in Psychology, 10*, 1675. https://doi.org/10.3389/fpsyg.2019.01675

Klotzke, K. & Fox, J. P. (2019b). Modeling Dependence Structures for Response Times in a Bayesian Framework. *Psychometrika, 84*, 649–672. https://doi.org/10.1007/s11336-019-09671-8

Lim, D., & Ziegler, M. (2023). Degrees of Second and Higher-Order Polynomials. *ArXiv*. https://doi.org/10.48550/arXiv.2305.03439

Murphy, M., Wang, J., Jiang, C. et al. (2024) A Potential Source of Bias in Group-Level EEG Microstate Analysis. *Brain Topogr 37*, 232–242. https://doi.org/10.1007/s10548-023-00992-7

Nielsen, N. M., Smink, W. A. C., & Fox, J.-P. G. J. A. (2021). Small and negative correlations among clustered observations: Limitations of the linear mixed effects model. *Behaviormetrika*, *48*(1), 51-77. https://doi.org/10.1007/s41237-020-00130-8