State Space Models with Longitudinal Data

Zach Baucom

Introduction

- State Space Models have been primarily used for time series data with a large number of time points and only a small number of chains observed.
- We are working to apply these models to a small number of time points and a large number of subjects.
  - Small t and large n are typically what we see in observational data.
- We wish to show that the State Space Model can be more flexible and robust than the commonly used mixed effect models (Laird and Ware, 1983; Diggle, Liang and Zeger, 1994).

Computation Consideration

- State space models can be computationally intensive.
- We will compare different state space model estimation methods to find the best balance of computational efficiency and accuracy.
  - State space model in matrix form.
  - Partitioned state space model.
  - Bayesian state space model.

State Space Model

A general linear state space model can be denoted as:

$$
\begin{aligned}
y_t &= F_t \mu_t + v_t \\
\mu_t &= G_t \mu_{t-1} + w_t
\end{aligned}
$$

where at time t,

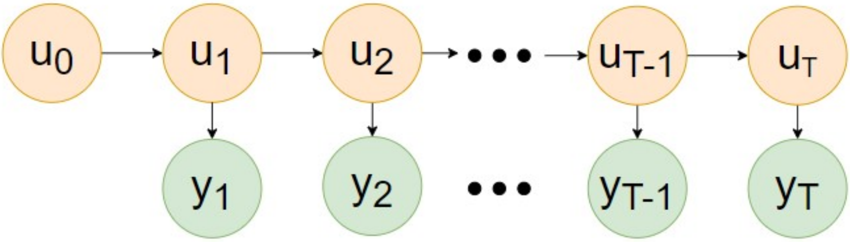
- y_t is the an $n \times 1$ observation vector.
- μ_t is the $q \times 1$ latent state vector, where q is the number of latent states.
- F_t is the $n \times q$ observation matrix.
- G_t is the $q \times q$ state transition matrix.

We assume v_t and w_t are independent identically distributed with distributions $v_t \sim N(0, V)$ and $w_t \sim N(0, W)$ respectively (Harvey, 1990; Durbin and Koopman, 2012) .

State Space Model Illustration

General Model:

$$y_t = F_t \mu_t + v_t$$
$$\mu_t = G_t \mu_{t-1} + w_t$$



Proposed Model

We wish to model the data according to the following,

$$
\begin{aligned}
y_t &= \alpha_t + X_t\beta_t + \varepsilon_t \\
\alpha_t &= \alpha_{t-1} + \eta_t \\
\beta_t &= \beta_{t-1}
\end{aligned}
$$

Where $\alpha_0 \sim N(a_0, P_0)$, $\beta_0 \sim N(\beta, 0)$, $\varepsilon_t \sim N(0, \sigma_\varepsilon^2 I_n)$, and $\eta_t \sim N(0, \sigma_\eta^2 I_n)$.

- y_t is an $n \times 1$ observation vector where n indicates the number of subjects.
- α_t is an $n \times 1$ latent state vector.
    - Variation in α_t over time creates a dynamic moving average auto-correlation between observations y_t.
- X_t is an $n \times p$ matrix of time varying covariates (can be $X_t = t * X$ where X are baseline covarties).

What is α_t

Consider the model,

$$
\begin{aligned}
y_t &= \alpha_t + X_t\beta_t + \varepsilon_t \\
\alpha_t &= \alpha_{t-1} + \eta_t \\
\beta_t &= \beta_{t-1}
\end{aligned}
$$

We can think of α_t as the underlying cognitive state not accounted for by the baseline covariates X.

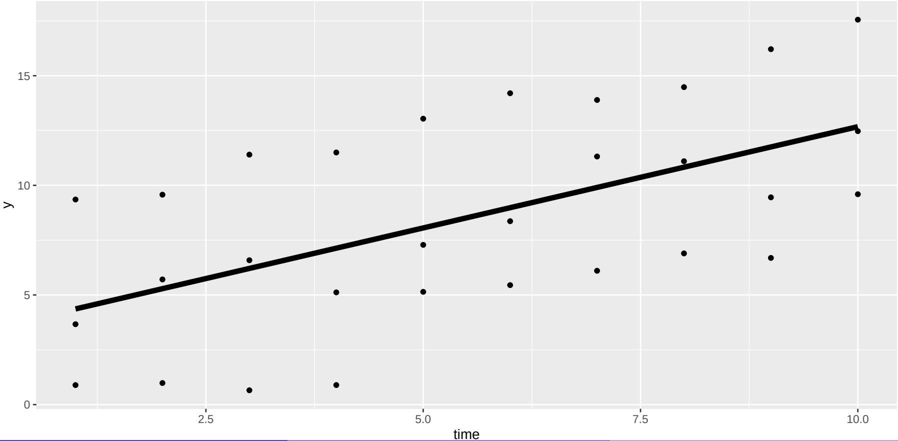
Notice $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \sigma_\eta^2)$. This means our next underlying cognitive state will be centered at the previous underlying cognitive state.

Remember $E(\alpha_t) = E(\alpha_{t-1} + \eta_t) = E(\alpha_{t-1})$. If we iterate all the way down $E(\alpha_t) = a_0$. So $\alpha_t > \alpha_0$ represents an up phase and $\alpha_t < \alpha_0$ represents a down phase.

LME with Random Intercept

Consider the model: $y_{it} = b_{i0} + t * \beta + \epsilon_{it}$ where $b_{i0} \sim iid\ N(0, \sigma_b^2)$ and $\epsilon_{it} \sim iid\ N(0, \sigma^2)$.

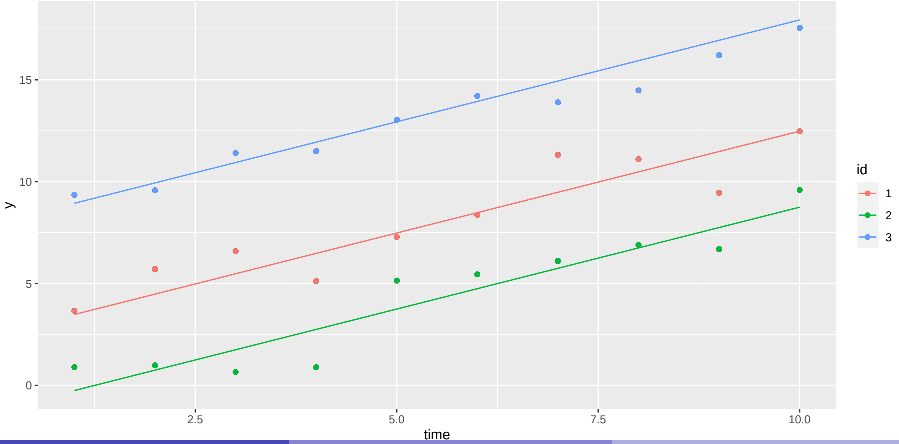
Let $\beta = 1$, $\sigma_b^2 = 10$, and $\sigma^2 = 1$.



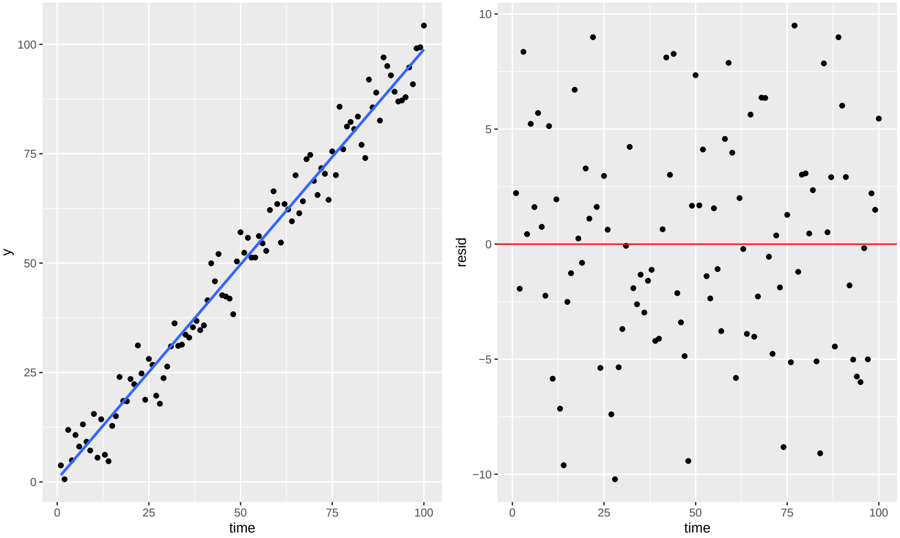
LME with Random Intercept

Consider the model: $y_{it} = b_{i0} + t * \beta + \epsilon_{it}$ where $b_{i0} \sim iid\ N(0, \sigma_b^2)$ and $\epsilon_{it} \sim iid\ N(0, \sigma^2)$.

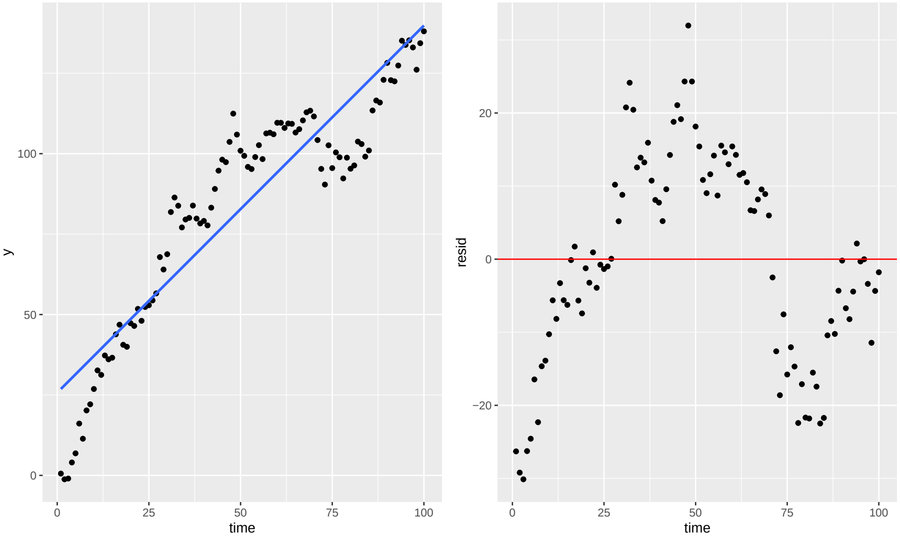
Let $\beta = 1$, $\sigma_b^2 = 10$, and $\sigma^2 = 1$.



Single observation from a LMEM



Single observation from a SSM



Auto-correlation

The correlation between observations at any two time points is called the auto-correlation.

Our proposed SSM model has the following auto correlation structure.

$$corr(y_{it}, y_{i(t+\tau)}) = \frac{t\sigma_\eta^2}{\sqrt{\sigma_\varepsilon^2 + t\sigma_\eta^2}\sqrt{\sigma_\varepsilon^2 + (t+\tau)\sigma_\eta^2}}$$

This is equivalent to a dynamic moving average covariance structure which is very flexible. If $\sigma_\eta^2 = 0$ then auto-correlation is 0 and our proposed model boils down to a LMEM.

$$y_t = \alpha_0 + X_t\beta + \varepsilon_t$$

Summary

Consider the model,

$$
\begin{aligned}
y_t &= \alpha_t + X_t\beta_t + \varepsilon_t \\
\alpha_t &= \alpha_{t-1} + \eta_t \\
\beta_t &= \beta_{t-1}
\end{aligned}
$$

We can think of α_t as the underlying cognitive state not accounted for by the baseline covariates X.

The variable β_t is the effect of the covariates X_t. It has the same interpretation as with a LMEM.

Relation to State Space Model

We can rewrite the proposed model to fit the state space model as follows,

$$y_t = \begin{bmatrix} I_n & X_t \end{bmatrix} \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} + \varepsilon_t$$

$$\begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} I_{(n+p)\times(n+p)} \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} \eta_t \\ 0_{p\times 1} \end{bmatrix}$$

- $F_t = \begin{bmatrix} I_n & X_t \end{bmatrix}$
- $v_t = \varepsilon_t$
- $w_t = \begin{bmatrix} \eta_t \\ 0_{p\times 1} \end{bmatrix}$
- $\mu_t = \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix}$
- $G_t = I_{(n+p)\times(n+p)}$

Kalman Filter

The Kalman filter is a recursive algorithm to estimate the unobserved states conditioned on the observed data (Kalman, 1960; Durbin and Koopman, 2012). Let $\hat{\mu}_{i|j} = E(\mu_i | y_{1:j})$ and $P_{i|j} = var(\mu_i | y_{1:j})$.

Predicted state: $\hat{\mu}_{t|t-1} = G_t \hat{\mu}_{t-1|t-1}$

Predicted state covariance: $P_{t|t-1} = G_t P_{t-1|t-1} G_t' + W$

Innovation covariance: $S_t = F_t P_{t|t-1} F_t' + V$

Kalman Gain: $K_t = P_{t|t-1} F_t' S_t^{-1}$

Innovation: $\tilde{f}_t = y_t - F_t \hat{\mu}_{t|t-1}$

Updated state estimate: $\hat{\mu}_{t|t} = \hat{\mu}_{t|t-1} + K_t \tilde{f}_t$

Updated state covariance: $P_{t|t} = (I - K_t F_t) P_{t|t-1}$

Updated innovation: $\tilde{f}_{t|t} = y_t - F_t \hat{\mu}_{t|t}$

Kalman Smoother

Let $J_t = P_{t|t} G'_{t+1} + P^{-1}_{t+1|t}$. We can then calculate $E(\mu_t | y_{1:T})$ and $var(\mu_t | y_{1:T})$ using the following Kalman smoother equations.

$$
\begin{aligned}
E(\mu_t | y_{1:T}) &= \hat{\mu}_{t|t} + J_t(\hat{\mu}_{t+1|T} - \hat{\mu}_{t+1|t}) \\
var(\mu_t | y_{1:T}) &= P_{t|t} - J_t G_{t+1} P_{t|t}
\end{aligned}
$$

Setting Parameters

We assume $\mu_0 \sim N(u_0, P_0)$, however u_0 and P_0 are unknown.

- By initializing $u_0 = 0$ and $P_0 = \infty$ we are essentially putting a flat prior on μ_0.
- It has been shown $\hat{\mu}_{0|T}$ and $P_{0|T}$ quickly converge to u_0 and P_0 respectively for even small T (Kalman, 1960; Durbin and Koopman, 2012).

In our proposed model, $\mu_t = \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix}$.

- $\hat{\beta}_{0|T}$ is then our estimate for β and has variance covariance $P_{\hat{\beta}} = [P_{0|T}]_{(n+1):(n+p),(n+1):(n+p)}$.
- We can then use $\hat{\beta}_{0|T}$ and $P_{\hat{\beta}}$ for inference on β.
  - $\hat{\beta} \overset{\text{asym}}{\sim} N(\beta, P_{\hat{\beta}})$.

Estimation of σ_ε^2 and σ_η^2

- We get proper estimates for β given we have correctly specified our model, including σ_ε^2 and σ_η^2.
- The parameters σ_ε^2 and σ_η^2 are unknown, but can be estimated using Maximum Likelihood Estimation (MLE).

$$\ell(\sigma_\varepsilon^2, \sigma_\eta^2) = -\frac{np}{2} log(2\pi) - \frac{1}{2} \sum_{i=1}^{t} \left(log|\tilde{S}_i| + \tilde{f}_i{`} S_i^{-1} f_i \right)$$

- To maximize the log-likelihood we used a Newton-Raphson method with a limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method (Liu and Nocedal, 1989; Zhou and Li, 2007).

Missing Data

If a subject is missing an observation at time t we can set

- $y^* = W_t y_t$ where W_t is a subset of rows of I_n corresponding to those with observed data.
- $F_t^* = W_t F_t$
- $\varepsilon_t^* = W_t \varepsilon_t$

then carry out the same Kalman filter and smoother replacing y with y^*, Z with Z^*, and ε_t^* with ε_t. Doing this modification still allows us to get the smoothed values for α_t and β_t.

Computational Challenges

For each iteration of the kalman filter we must invert $var(Y_t|y_{1:(t-1)}) = S_t$.

- S_t is non-sparse as calculating $var(Y_t|y_{1:(t-1)})$ is a function of β_{t-1} which is shared between all observations.
- S_t is an $n \times n$, so as n increases there is an exponential increase in computation time.

Solution 1: Partitioning

A solution to solving inversion computational inefficiencies is to partition:

- Partition the subjects into k groups.
- Run the Kalman filter and smoother on each group independently to extract $\hat{\beta}_{0|T}^{(i)}$ and $P_{\beta}^{(i)}$ for i in $1, ..., k$.
- Use the estimate $\bar{\beta} = \frac{\sum_{i=1}^{k} \hat{\beta}_{0|T}^{(i)}}{k}$.
  - $\bar{\beta} \sim N(\beta, \frac{\sum_{i=1}^{k} P_{\hat{\beta}^{(i)}}}{k^2})$

Solution 2: Bayesian Gibb's Sampling Approach

- For the Bayesian approach we use a Gibb's sampler.
- Instead of calculating β in the Kalman filter, we can estimate it separately.
- The model,

$$
\begin{aligned}
y_t &= \alpha_t + X_t\beta + \varepsilon_t \\
\alpha_t &= \alpha_{t-1} + \eta_t
\end{aligned}
$$

Gibb's Sampling

- Gibb's sampling is a method to gain an approximate sample from a posterior distribution for a given variable (Gelfand-Smith, 1990).
- It works by:
    - calculating the distribution of a variable conditioned on all other unknown variables, known as the posterior distribution.
    - sampling from the posterior distribution and assigning the new sample to the variable.
    - calculate the posterior of the next variable and continue to sample, update, and recalculate the other posteriors.
    - The process is commonly repeated thousands of times.
- We need to calculate the posterior for $\alpha_{1:T}, \beta, \sigma_\varepsilon^2, \sigma_\eta^2$.

Posterior of α

- Notice, if we are conditioning on β for the posterior $\alpha_{1:T}|...$ then each y_{it} is independent and we can run the Kalman filter chains independently.
- Let $y_t^* = y_t - X_t\beta$, then the model becomes

$$y_t^* = \alpha_t + \varepsilon_t$$
$$\alpha_t = \alpha_{t-1} + \eta_t$$

- We can then run a forward Kalman filter with a backward sampler to sample from the posterior of $\alpha_{1:T}$ (Fruhwirth-Schnatter, 1994)

Posterior of β

- We let $\beta \sim N(\theta, \sigma_\beta^2)$
- The posterior is $\beta | ... \sim N(\Sigma^{-1}B, \sigma_\varepsilon^2 \sigma_\beta^2 \Sigma^{-1})$ where,
- $B = \sigma_\beta^2 \big(\sum_{t=1}^{T} y_t - \alpha_t \big)' X_t - \sigma_\varepsilon^2 \theta$
- $\Sigma = (\sigma_\beta^2 \sum_{t=1}^{T} X_t' X_t) + \sigma_\varepsilon^2 I_p$

Posterior of β

For each iteration of the Gibb's sampler we must calculate, $\Sigma^{-1} = ((\sigma_\beta^2 \sum_{t=1}^T X_t' X_t) + \sigma_\varepsilon^2 I_p)^{-1}$. As σ_ε^2 is updated each iteration, Σ^{-1} will be different for each iteration as well. However, $(\sigma_\beta^2 \sum_{t=1}^T X_t' X_t)$ remains constant. By calculating the eigenvalue decomposition before the Gibb's sampler we can increase computation speed.

$$
\begin{aligned}
((\sigma_\beta^2 \sum_{t=1}^T X_t' X_t) + \sigma_\varepsilon^{2(i)} I) &= (Q \Lambda Q' + \sigma_\varepsilon^{2(i)} I) \\
&= (Q \Lambda Q' + \sigma_\varepsilon^{2(i)} QQ') \\
&= Q(\Lambda + \sigma_\varepsilon^{2(i)} I) Q'
\end{aligned}
$$

then,

$$
((\sigma_\beta^2 \sum_{t=1}^T X_t' X_t) + \sigma_\varepsilon^{2(i)} I)^{-1} = Q(1/(\Lambda + \sigma_\varepsilon^{2(i)} I)) Q'
$$

Posterior of σ_ε^2 and σ_η^2

- Let,

$$\sigma_\eta^2 \sim IG(a_0/2, b_0/2)$$
$$\sigma_\varepsilon^2 \sim IG(c_0/2, d_0/2)$$

- Then

$$\sigma_\eta^2|... \sim IG(\frac{nT + a_0}{2}, \frac{\sum_{t=1}^T(\alpha_t - \alpha_{t-1})^2 + b_0}{2})$$
$$\sigma_\varepsilon^2|... \sim IG(\frac{nT + c_0}{2}, \frac{d_0 + \sum_{t=1}^T(y_t - X_t\beta - \alpha_t)^2}{2})$$

The Gibbs Sampling Algorithm

1. Select prior parameters for $\theta, \sigma_\beta^2, a_0, b_0, c_0, d_0$.
2. Let $\beta^{(0)} = \theta$, $\sigma_\eta^{2(0)} = \frac{d_0/2}{1+c_0/2}$, and $\sigma_\varepsilon^{2(0)} = \frac{b_0/2}{1+a_0/2}$.
3. Run a forward-filtering backward sampling procedure as described above conditioning on $\beta^{i-1}, \sigma_\eta^{2(i-1)}, \sigma_\varepsilon^{2(i-1)}$ and set the samples equal to $\alpha^{(i)}$ for the i^{th} iteration.
4. Sample σ_η^{2*} from $IG(\frac{nT+a_0}{2}, \frac{\sum_{t=1}^{T}(\alpha_t^{(i)}-\alpha_{t-1}^{(i)})^2+b_0}{2})$ and set $\sigma_\eta^{2(i)} = \sigma_\eta^{2*}$.
5. Sample σ_ε^{2*} from $IG(\frac{nT+c_0}{2}, \frac{d_0+\sum_{t=1}^{T}(y_t-X_t\beta^{(i-1)}-\alpha_t^{(i)})^2}{2})$ and set $\sigma_\varepsilon^{2(i)} = \sigma_\varepsilon^{2*}$.
6. Sample β^* from $N(\Sigma^{-1}B, \sigma_\varepsilon^2\sigma_\beta^2\Sigma^{-1})$ where $\alpha = \alpha^{(i)}, \sigma_\eta^2 = \sigma_\eta^{2(i)}, \sigma_\varepsilon^2 = \sigma_\varepsilon^{2(i)}$ and set $\beta^{(i)} = \beta^*$.
7. Repeat steps 3-6 for i in 1, 2, . . . , M.

Estimating β

- After throwing out a number of initial samples from the Gibb's sampler we can estimate β by taking the mean of the posterior samples.
- We create a 95 credibility interval (as a pseudo-confidence interval) by calculating the 97.5^{th} and 2.5^{th} percentiles of the posterior draws.

Simulation

- We sampled from the model,

$$\begin{aligned} y_t &= \alpha_t + t * X\beta + \varepsilon_t, & \varepsilon_t &\sim N(0, \sigma_\varepsilon^2 I_n) \\ \alpha_t &= \alpha_{t-1} + \eta_t, & \eta_t &\sim N(0, \sigma_\eta^2 I_n) \end{aligned}$$

We simulated 100 subjects at 6 time points. X was simulated from a $U(0, 20)$ distribution and $\beta = (4 \quad 2 \ -1)'$.

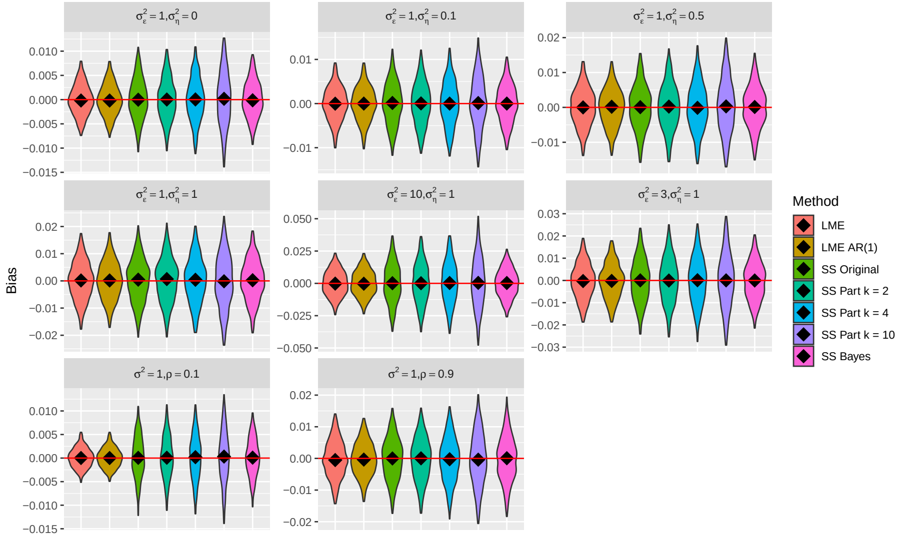
The variables σ_ε^2 and σ_η^2 varied between simulations. Recall, $\sigma_\eta^2 = 0$ corresponds to a lmem with a random intercept.

We compared 95% CI coverage, CI length, and estimate variance between 1. LMEM with a random intercept, 2. LMEM with a random intercept and AR(1) error correlation structure, the matrix formulation of the state space model, the Bayesian estimated state space model, the a state space model partitioned into 2, 4, and 10 groups.

Coverage

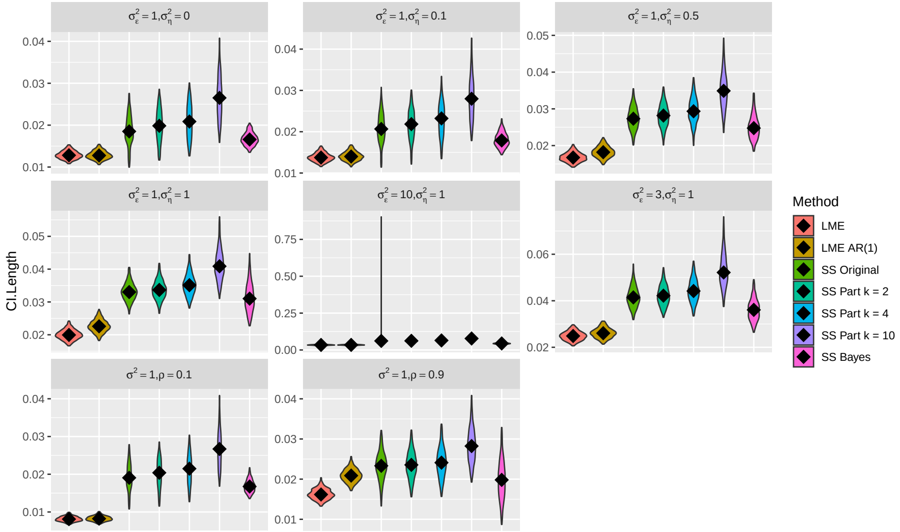
Variance Parameters		Traditional Methods		State Space Methods				
σ_ε^2	σ_η^2	LME	AR(1)	SSM	Bayes	Part2	Part4	Part10
1	0	0.950	0.949	NA	0.956	0.959	0.970	0.960
1	0.1	0.921	0.926	0.944	NA	0.961	0.972	0.948
1	0.5	0.847	0.884	0.953	0.957	0.955	0.957	0.941
1	1	0.817	0.875	0.949	0.953	0.951	0.956	0.941
10	1	0.915	0.919	0.952	0.950	0.949	0.942	0.953
3	1	0.879	0.896	0.954	0.954	0.957	0.956	0.955
1	$\rho = 0.9$	0.809	0.940	NA	0.899	NA	0.894	0.792
1	$\rho = 0.1$	0.936	0.938	NA	0.947	NA	0.964	0.949

Bias



$\sigma_\varepsilon^2 = 1, \sigma_\eta^2 = 0$
$\sigma_\varepsilon^2 = 1, \sigma_\eta^2 = 0.1$
$\sigma_\varepsilon^2 = 1, \sigma_\eta^2 = 0.5$
$\sigma_\varepsilon^2 = 1, \sigma_\eta^2 = 1$
$\sigma_\varepsilon^2 = 10, \sigma_\eta^2 = 1$
$\sigma_\varepsilon^2 = 3, \sigma_\eta^2 = 1$
$\sigma^2 = 1, \rho = 0.1$
$\sigma^2 = 1, \rho = 0.9$
Bias
Method
LME
LME AR(1)
SS Original
SS Part k = 2
SS Part k = 4
SS Part k = 10
SS Bayes

CI Length



$\sigma_\varepsilon^2 = 1, \sigma_\eta^2 = 0$
$\sigma_\varepsilon^2 = 1, \sigma_\eta^2 = 0.1$
$\sigma_\varepsilon^2 = 1, \sigma_\eta^2 = 0.5$
$\sigma_\varepsilon^2 = 1, \sigma_\eta^2 = 1$
$\sigma_\varepsilon^2 = 10, \sigma_\eta^2 = 1$
$\sigma_\varepsilon^2 = 3, \sigma_\eta^2 = 1$
$\sigma^2 = 1, \rho = 0.1$
$\sigma^2 = 1, \rho = 0.9$
CI.Length
Method
LME
LME AR(1)
SS Original
SS Part k = 2
SS Part k = 4
SS Part k = 10
SS Bayes

Key Take-aways

- The state space methods give unbiased estimates while maintaining near 0.95 coverage probability for the 95% CIs.
  - While the LME methods are unbiased, they do not maintain 0.95 coverage probability when auto-correlation is increased.
- If the number of subjects in each partition is reasonable compared to the number of coefficients to estimate, then partitioning returns very similar results to not partitioning.
- Of the state space models, the Bayesian method has the least amount of variability in the estimates, the smallest variability in the estimate variances, all while maintaining 0.95 coverage probability.
- However, the Bayesian method fails to converge when the data generation came from an AR(1) model.

Note

- All the SSM models can handle non standard, unequally spaced, and continuous time observations.

Future Steps

- Apply methods to existing data where the underlying distributions are unknown.
  - National Alzheimer's Coordinating Center (NACC).
  - Run power analysis (Similar to Alicia's).