Bayesian shape invariant model for longitudinal growth curve data

Mohammad Alfrad Nobel Bhuiyan, Heidi Sucharew, Rhonda Szczesniak, Marepalli Rao, Jessica Woo, Jane Khoury, Md Monir Hossain

December 24, 2019

Abstract

Growth curve modeling should ideally be flexible, computationally feasible and allow for the inclusion of co-variates for better predictability and mechanistic explanations. The original shape invariant growth curve model, motivated by epidemiological research on the evolution of pubertal heights over time, fits the underlying shape function for height over age and estimates subject-specific deviations from this curve in terms of size, tempo, and velocity using maximum likelihood. This approach is effective in subsequent applications but the estimation method does not provide uncertainty estimates for unknown parameters, and predictive ability has been largely unexplored. A more recent Bayesian implementation undertaken for multivariate SIM modeling and this approach requires multiple longitudinal outcomes and has added a computational burden. Modern growth curve studies of height measurements from children with attention-deficit hyperactivity disorder (ADHD) have gained importance in epidemiological research due to potentially adverse effects from stimulant medications. Motivated by a particular longitudinal study on the heights of 197 pediatric ADHD patients who began stimulant treatment at varying ages, we describe a Bayesian extension of the original SIM model. We incorporate co-variate effects, assess mixing properties and examine different spline formulations to model the underlying growth shape. As demonstrated by the real data application and simulation study, the Bayesian SIM approach provides a natural, computationally feasible way to generate uncertainty estimates for treatment-outcome associations. We discussed the future extensions to the approach.

Keywords: Bayesian inference, functional data analysis, growth curves, shape invariant

1 Introduction

The analysis of childhood growth curves has played a vital role in estimating the growth trajectory of populations as well as identifying critical factors corresponding to various shapes of those trajectories, such as sex. Indeed, early origins of growth curve modeling utilized cross-sectional growth curves, in which the population data were used to derive the growth patterns for various age and gender groups. The two widely known references of using cross-sectional data are the CDC growth chart [Kuczynski, 2002] and WHO growth standards [Organization et al., 2006], which are mainstays in clinical care. More recently, growth curve analyses based on longitudinal data have allowed more accurate identification of growth patterns, since the longitudinal data allows incorporating within- and between-subject effects simultaneously [Willemsen et al., 2015].
Earlier work to fit growth curves based on one of two parametric assumptions namely logarithmic or exponential. Logarithmic curves assume a quick growth increase at the beginning, but the gains slowly disappear as time passes. Logarithmic growth curves broadly applied in bacterial growth [Zwietering et al., 1990, Barry, 1995], biodegradation [Schmidt et al., 1985], fitness and strength training, and learning ability. By contrast, the exponential curve assumes that growth is slower at the beginning with gains that are more rapid over time. [Karlberg, 1987] explained growth curve modeling through exponential curves. Jenss and Bayley [1937] proposed a different model adding a linear term which was fitted by [Berkey, 1982] and found a poor fit of the data based on the systematic variation of the residual [Beath, 2007]. Moreover, [Berkey and Laird, 1986] also analyzed the model proposed by Count [Count, 1943] and found an exponential model as a better model. Different parametric growth curve models proposed over the last two decades, such as the linear model, reciprocal model, logistic model, Gompertz model, and the Weibull model. [Milani et al., 1989], [Wingerd, 1970] explored other approaches of growth curves and found a poor fit. [Wishart, 1938] Explained growth data analysis using polynomial regression. One limitation of polynomial regression that it requires a higher degree of polynomial to provide an adequate fit with the resulting coefficients without having any significant interpretation. Alternative approaches to analyze growth curve data were proposed by [Geva et al., 1993], [Ong et al., 2002] [Newell et al., 2003], [Rao, 1958]. Nonparametric models using regression splines to model the underlying shape function have been shown to decrease bias, thereby improving the estimation of subject-specific effects [Viele et al., 2006]. Furthermore, regression splines, such as natural splines, have been shown to provide a better-localized fit to the mean response, compared to global polynomials [Lambert et al., 2001].

Equally important to finding an appropriate model to depict growth patterns is understanding the risk factors that contribute to adverse growth. The growth model is essential for designing an intervention trial or increase public health awareness surrounding potential benefits and adverse effects of various environmental exposures and growth patterns. Furthermore, having interpretable estimates for growth characteristics, such as peak height velocity, can ameliorate confounding in epidemiologic studies. [Simpkin et al., 2017]. Other examples, which have motivated statistical developments ranging from hierarchical linear models to multivariate analysis methods, include psychological change over time [Hertzog and Nesselroade, 2003], cardiovascular study [Llabre et al., 2004], associations between adolescent moderate-vigorous physical activity and depressive symptoms in young adulthood [Brunet et al., 2013], associations among timing of sexual victimization and timing of drinking behavior [Griffin et al., 2013], examine longitudinal associations among cognition, function, and depression in Alzheimer’s Disease patients [Zahodne et al., 2013], describing change in personality trait [Jackson et al., 2012].

In recent years, shape invariant modeling (SIM) has become an active area of research for non-parametric growth curve modeling, where a single function (or, curve) is transformed by scaling and shifting it to fit each subject usually through affine transformations. [Lawton et al., 1972]. Who first proposed SIM called it self-modeling regression; in their approach, the function for the underlying shape illustrated for various parametric functions. Later, [Beath, 2007] developed a model to explain longitudinal growth patterns and extended the SIM to include time-dependent covariates. [Cole et al., 2010]. Extended the model by changing the sign of the velocity parameter and named it SITAR (Superimposition by Translation and Rotation). As a type of SIM, the regression spline expressed as a basis function consisting of a different set of knots; the resulting structure fitted as a nonlinear mixed effects model and parameters are typically estimated using maximum likelihood. This allows estimating the parameters for the between-subjects variation. Based on the underlying pattern of the data, various shape in-
variant models have proposed. When the data has logarithmic growth, SIM proposed for the log-transformed data.

Bayesian growth curve modeling have also seen similar progress with many applications to real datasets as well as longitudinal growth datasets [Barry, 1995],[Arjas et al., 1997], [Dunson, 2001], [Pinheiro and Bates, 2000], [Shohoji et al., 1991]. One of the main advantages of a Bayesian approach is that it generates the uncertainty estimates (i.e., the estimate for the variance) for all unknown parameters naturally since each parameter explained by a probability distribution. Other advantages include the use of prior probability distributions to assimilate information from previous studies or experts opinion and allows to control confounding; having posterior probabilities as easily interpretable alternatives to p-values; in hierarchical modeling, incorporating latent variables such as an individual’s true disease status in the presence of a diagnostic error. Moreover, MCMC methodology facilitates the implementation of Bayesian analyses of complex data sets containing missing observations and multidimensional outcomes [Dunson, 2001]. Due to this flexibility and better prediction of the exposure-outcome relationship, researchers are becoming more interested in Bayesian modeling. Notable work in Bayesian growth curve modeling includes the multivariate extension by [Willemsen et al., 2015] the original SIM model.

A brief outline of the paper is as follows. In Section 2, a brief description of the original SIM model and the interpretation of various model parameters provided. In a subsection, we provided some descriptions for the spline function used in SIM and how this model connected to GAM. Section 3 described the Bayesian implementation of the SIM model with and without subject-specific covariates, and also the DAG representations of these models. Section 4 illustrated the MCMC implementation; the specification for the prior distributions; the full conditional distribution and the posterior distribution for each model parameter; and how the assessment of model performance. The full derivation of posterior distributions given in the Appendix. Applications with the real data are provided in Section 5, and with simulated data in Section 6. The Final Section includes the discussion and some proposals to future extension.

2 Real Data Example: ADHD Children

Longitudinal data from retrospective chart review for heights and weights for 197 ADHD (Attention-deficit Hyperactivity Disorder) children who visited community-based pediatric primary care practices in Cincinnati (Ohio, USA) collected. The children were in the age range: 1-17 years, and under stimulant medications to effectively reduce symptoms of ADHD. The objective of the original study was to evaluate how the age at start of stimulant medication may have impacted child growth trajectories. The original ADHD data had 6,134-time points recorded on 197 patients, among them only 3084 height measures were available for the analysis. The longitudinal study design is appeared to be unbalanced. Figure 3 shows the growth heights of each patient at various ages. The study enrolled 138 males (70%). For the 197 unique patients the age range was (1.37, 16.76) with the mean age 9.3 years, and the height range (76.2, 183.2) with the mean height 134.4 cm (Table 1). Each patient has prescribed a stimulant medication at a certain visit. The mean age at stimulant medication start was 7.9
years and ranged from 4.2 to 12.3 years.

Table 1: Summary statistics for the 197 ADHD patients

| Variable                              | Min  | 1st.Q | Median | Mean  | 3rd.Q | Max  | SD   |
|---------------------------------------|------|-------|--------|-------|-------|------|------|
| Age                                   | 1.363| 7.249 | 9.443  | 9.263 | 11.369| 16.764| 3.001|
| Height                                | 76.2 | 122.6 | 134.6  | 134.4 | 147.3 | 183.2| 18.86|
| Weight                                | 9.072| 21.32 | 28.12  | 31.71 | 38.33 | 114.3| 31.51|
| Start age of stimulant medication     | 4.17 | 6.75  | 7.86   | 7.949 | 9.120 | 12.330| 1.6  |

We applied the Bayesian SIM model to the ADHD data and sought to describe the size, tempo, and velocity parameters relevant to the height growth of ADHD patients. And simultaneously, we describe the model fitting procedure, and its relative performances in comparison to the frequentist SIM model in details.

3 Shape invariant model

Shape invariant model (SIM) is a non-parametric model where a single function is transformed by shifting and scaling to fit each subject. In SIM a regression spline is used as the function, with log transformation of the data. For this study we followed the notation from [Beath, 2007], the SIM model can be expressed as:

\[ y_{ij} = \gamma_{i2} + h(t_{ij} - \gamma_{i1}) + \epsilon_{ij}; i = 1, \ldots N, \text{ and } j = 1, \ldots T_i. \]  (1)

Here, the \( y_{ij} \) is the growth measure of \( i^{th} \) child at the \( j^{th} \) time points which corresponds to age (in years) in our motivating example. Subject-specific coefficients \( \gamma_i = (\gamma_{i1}, \gamma_{i2}, \gamma_{i3}) \) enable each individual’s growth trajectory to be aligned to a common growth curve, \( h(\cdot) \), via transformations to the \( x- \) and \( y- \) axes. In this formulation, we will estimate \( h(\cdot) \) using a spline function and let \( \epsilon_{ij} \) be measurement error.

The goal is to estimate the subject-specific vector \( \gamma_i \) such that the corresponding individual growth curve from deviations from the average curve \( h(\cdot) \). Following previously described work by Cole and others in equation (1), \( \gamma_{i2} \) is termed as Size. \( \gamma_{i2} \) can also be interpreted as subject-specific shift up or down in the spline curve along the response axis. \( \gamma_{i2} \) is a random intercept term; when the response measure is height, \( \gamma_{i2} \) is larger for taller children and smaller for shorter children. \( \gamma_{i1} \) termed Tempo, which is a random time intercept and corresponds to differences in the timing of the growth spurt. This subject-specific left-right shift in the growth curves positive for late puberty and negative for early. The scaling factor within the spline function, \( \gamma_{i3} \) is termed Velocity and corresponds to differences in the duration of the growth spurt between individuals. The Velocity parameter shrinks or stretches the time scale [Cole et al., 2010].

4 Bayesian Implementation of Shape invariant model

The above SIM model presented in equation (1) can be written with basis representation as follows:

\[ y_{ij} = \gamma_{i2} + Z_{ij}^T \beta_{(\kappa+2)} + \epsilon_{ij}; i = 1, \ldots N, \text{ and } j = 1, \ldots T_i. \]  (2)
Figure 1: Schematic representation of shape invariant model with Horizontal shift($\gamma_1$), vertical shift($\gamma_2$) and stretch($\gamma_3$)

Where,

$$Z_{ij} = B(\exp(\gamma_{i3})(t_{ij} - \gamma_{i1})),$$

$$\gamma_i = (\gamma_{i1}, \gamma_{i2}, \gamma_{i3}), \gamma_i \sim N_3(0, \Sigma_{\gamma}), \text{and} \quad \epsilon_{ij} \sim N(0, \sigma^2).$$

$Z_{ij}$ is the basis of the natural cubic spline, evaluated at $(\exp(\gamma_{i3})(t_{ij} - \gamma_{i1}))$. Thus, $z_{ijk}$ is a vector of length $\kappa + 2$, and $\beta$ is the regression coefficient vector of same length. Here, $\kappa$ is the number of inner knots and 2 is for the boundary knots. so natural cubic spline has $\kappa + 2$ independent coefficients. Subject-specific $\gamma_i$ is assumed to have multivariate normal of order 3 with 0 mean vector and $\Sigma_{\gamma}$ variance-covariance matrix. We assume that $\epsilon_{ij}$ is independently normally distributed with mean 0 and variance $\sigma^2$, and also independent of $\gamma_i$.

The other distributions for $\epsilon_{ij}$ such as Student’s-t can also be considered depending on the type of growth data.

With subject-specific covariates: In many growth curves analyses, the needs for the inclusion of covariates for better predictability, as well as for better explanation of growth mechanisms are essential. For example, it may be of interest to know how the gender difference affects the growth patterns, or how the medication at early age for a specific disease condition affects the growth at later ages specifically by size, tempo, and velocity. The subject-specific covariates can be included in the model specified for $\gamma_i$ with a non-zero mean vector. If we assume $(p-1)$ subject-specific covariates, the mean and the variance of $\gamma_i$ becomes,

$$\gamma_i \sim N_3(AX_i, \Sigma_{\gamma}),$$

where,

$$A = \begin{pmatrix}
\alpha_{11} & \alpha_{12} & \ldots & \alpha_{1p} \\
\alpha_{21} & \alpha_{22} & \ldots & \alpha_{2p} \\
\alpha_{31} & \alpha_{32} & \ldots & \alpha_{3p}
\end{pmatrix}, \quad X_i = \begin{pmatrix} x_{i1}, & x_{i2}, & \ldots, & x_{ip} \end{pmatrix}^T, \quad \text{and} \quad \Sigma_{\gamma} = \begin{pmatrix}
\sigma^2_1 & \sigma_{12} & \sigma_{13} \\
\sigma_{21} & \sigma^2_2 & \sigma_{23} \\
\sigma_{31} & \sigma_{32} & \sigma^2_3
\end{pmatrix}.$$

The first column of the regression coefficient matrix $A$ is for the intercept, and the remaining columns are for each $(p - 1)$ covariate coefficients. Similarly, the first column of the design matrix $X$ contains the value 1.

Directed acyclic graph (DAG): DAG is a graphical representation of a hierarchical model which shows how the observed data and the unobserved parameters are conditionally depending on each others. In the graph, the circle indicates the stochastic node (or, the unobserved
parameters that need to be estimated), and the rectangle indicates observed data or the hyper-
parameters where they were assigned to fixed values apriori. When the Bayesian hierarchical
model has a complex dependency structure, DAG helps in better visualization of the model
as a whole, and the derivation of the posterior distribution for each stochastic node. Figure 2
shows the DAG of SIM model without covariates (a), and with subject-specific covariates (b).

![DAG of SIM model](image)

Figure 2: Graphical representation of the model without(a) and with subject-specific covari-
ates(b).The subject specific covariates are shown in red.

## 5 MCMC Implementation

### 5.1 Prior distributions:

The prior distribution for all model parameters are assumed to be independent apriori, and
follow uninformative flat probability distribution in general. The residual variance was assumed
to follow inverse gamma distribution with fixed shape and scale parameters such that \( \sigma^2 \sim IG(0.001, 0.001) \). Alternative prior distributions for residual variance parameter can also be
used following [Gelman, 2004]. For covariate coefficients and basis coefficient, we have assumed
\( \alpha \sim N(0, 1000) \) and \( \beta \sim N(0, 1000) \), respectively. The vector \( \alpha \) was defined after stacking
the A matrix. The variance-covariance matrix for \( \gamma_i \) is assumed to follow an inverse Wishart
distribution with 3 degree of freedom and 0.01 scale parameter, \( \Sigma_{\gamma} \sim IW(3, 0.01) \).

### 5.2 Full conditional distributions:

In Bayesian framework the joint posterior distribution is proportional to the product of the
likelihood function and the prior distributions. Therefore, the full posterior distribution for the
model in (2) with subject-specific covariates is as follows;

\[
p(\alpha, \beta, \sigma^2, \gamma, \Sigma_\gamma | y) \propto \prod_{i=1}^{N} \prod_{j=1}^{T_i} N(y_{ij} | \gamma_{i2} + z_{ij} (\gamma)^T \beta, \sigma^2) \prod_{i=1}^{N} N(\gamma_i | Ax_i, \Sigma_\gamma) \\
\times \prod_{i=1}^{3p} N(\alpha | 0, \sigma_{\alpha} I_{np}) \times \prod_{k=1}^{k+2} N(\beta_k | 0, \sigma_{\beta_k}^2) \times IG(\sigma^2 | \alpha_{\sigma}, \beta_{\sigma}) \times IW(\Sigma_\gamma | \delta, \psi) \tag{3}
\]
We follow the block update procedures in MCMC iterations whenever the posterior distribution has a standard form. It improves the convergence and mixing as well, and also saves computational time. Following the Gibbs sampling procedure, the updates for each parameters are as shown below. For the sake of simplicity, covariate effects and subject-specific notation \((i)\) are omitted for portions of the updates.

**Updating \(\alpha\):**

\[
p(\alpha | \beta, \sigma^2, \gamma, \Sigma_{\gamma}, y) \propto N(\bar{\Sigma}_{\alpha}[(X^T \otimes \Sigma_{\gamma})^{-1}]_{\gamma_{vec}}, [X^T X \otimes \Sigma_{\gamma}^{-1} + I_{n_p} \sigma^2_{\alpha}]^{-1})
\]

where, \(\bar{\Sigma}_{\alpha} = [X^T X \otimes \Sigma_{\gamma}^{-1} + I_{n_p} \sigma^2_{\alpha}]\), and \(\gamma_{vec}\) is a vector of stacked \(\gamma^T\).

**Updating \(\beta\):**

\[
p(\beta | \alpha, \Sigma_{\gamma}, \Sigma, y) \propto N(\sigma^{-2}\bar{\Sigma}_{\beta}Z^T(y - \gamma_2), \sigma^{-2}Z^T Z + \sigma^{-2}_{\beta} I_{k+2})
\]

where, \(\bar{\Sigma}_{\beta} = [\sigma^{-2}Z^T Z + \sigma^{-2}_{\beta} I_{k+2}]\).

**Updating \(\sigma^2\):**

\[
p(\sigma^2 | \alpha, \beta, \gamma, \Sigma_{\gamma}, y) \propto \text{Inverse-Gamma} \left( a + \frac{\sum T_i}{2}, b + \frac{\sum \{y_{ij} - \gamma_{i2} - B(\exp(\gamma_{i3})(t_{ij} - \gamma_{i1}))\}^2}{2} \right)
\]

**Updating \(\Sigma_{\gamma}\) (without covariate):**

\[
p(\Sigma_{\gamma} | \alpha, \beta, \gamma, \Sigma, y) \propto \text{Inverse-Wishart} \left( \delta + n, (\psi^{-1} + \gamma^T \gamma)^{-1} \right)
\]

where, \(\delta\) is degrees of freedom, and \(\psi\) is the scale matrix.

**Updating \(\Sigma_{\gamma}\) (with covariate):**

\[
p(\Sigma_{\gamma} | \alpha, \beta, \gamma, \Sigma, y) \propto \text{Inverse-Wishart} \left( \delta + n, (\psi^{-1} + (\gamma - XA)^T(\gamma - XA))^{-1} \right)
\]

where, \(\delta\) is degrees of freedom and \(\psi\) is the scale matrix.

Moreover, for the full conditional of the subject-specific parameter, \(\gamma_i\) is given by,

\[
\gamma_i | y_i, x_i, \alpha, \beta, \sigma^2, \Sigma_{\gamma} \propto N(\gamma_i | Ax_i, \Sigma_{\gamma}) \times \prod_j N(y_{ij} | \gamma_{i2} + B(\exp(\gamma_{i3})(t_{ij} + \gamma_{i1}))\beta, \sigma^2)
\]

For updating \(\gamma_i\), we used a random walk Metropolis-Hastings (M-H) algorithm to generate posterior samples for subject-specific effect. The candidate samples are generated from multivariate Student’s \(- t\) distribution with 5 degrees of freedom and mean at the current value. The variance parameter of multivariate Student’s \(- t\) was appropriately tuned to ensure that the acceptance rate of candidate samples to posterior samples in M-H step is around 20-30%. 

7
5.3 Convergence, Mixing, and Identifiability

Before summarizing the MCMC samples (or, posterior samples) as posterior mean, median or highest posterior region; it is important to check that the posterior samples for each parameter are converging, mixing well and has less auto-correlation. Convergence of MCMC samples indicates how close we are to the true posterior distribution and mixing indicates how well the parameter space explored. There are different ways of checking these namely, trace plot, autocorrelation plot, QQ plot, Brooks plot, Gelman-Rubin test, etc. A simple exploration of the trace plot gives an insight into the characteristics of the MCMC samples. Trace plots are produced for each parameter and checked whether different starting values led to better mixing and convergence, saving the computational time as well.

The SIM model requires to specify the Spline function with a specific number of knots. We used both B-spline and the natural cubic spline functions for checking their relative performances. As for specifying the number of knots, we used five equally spaced knots for both spline function, three interior and two exteriors. The number of knots was determined based on a compromise between optimizing a fit criterion and the computational burden. The prior distributions specified according to Section 4.1. The MCMC implementation followed a block update procedure with a mix of Gibbs and M-H algorithm. In an implementation, the inverse-Wishart distribution for the posterior distribution for \( \Sigma \) was redefined as a scaled inverse-Wishart distribution for correct estimation of correlation matrix as well as for quick convergence.

We ran multiple chain with a relatively longer burn-in period. The model run for 500,000 iterations with 90% burn-in samples. Then we have 50,000 samples for posterior inference. We further reduced the posterior samples to size 5,000 after thinning by parameter 10 for reducing the autocorrelation in posterior samples. All the results reported in this manuscript derived from this 5,000 posterior samples. Convergence was checked by examining the trace plot of the posterior samples for each parameter and by using [Gelman and Rubin, 1992] test. We have also checked the autocorrelation from the autocorrelation plot, and it showed a much faster rate of decreasing towards zero with the increasing lag values. The variance-covariance matrix of the proposal density in the M-H algorithm was tuned accordingly so that the acceptance rate was approximately 20%.

Mixing well of posterior samples is an important property to ensure that there is no specific trend in posterior samples among the parameters and they are independent as much as possible. We randomly selected three patients and their posterior samples for a horizontal shift (\( \gamma_1 \)), and stretch (\( \gamma_3 \)) parameters plotted in Figure 3. The plot shows there is no specific trend in posterior samples for these two parameters and they scattered around the center indicating low correlation.
6 Simulated Data Example

The simulated experiment was designed to assess the relative performances of the Bayesian SIM model concerning its frequentist counterpart. An R package "AGD (Analysis of Growth Data)" Version 0.35 was used to simulate the height data using the LMS method to obtain normalized growth. Using the growth charts data for the reference population, AGD computes heights in cm conditioning on age, sex, and the growth percentile. The current version of 'AGD' includes data for three reference populations: the United States, The Netherlands, and WHO data from multiple countries. We used the United States population as our reference population in simulating heights. To ensure wide variability and some levels of grouping patterns existed in the simulated heights, we generated data from three groups. In the first group, growth percentile range was set to (70 - 90)%; growth percentile range for the second group set to (40 - 60)%; and the third group set to (10 - 30)%. In each group, the proportions for males and females were equal.

The steps involved in the data generation process are:

1. Generate the height, $y_{ij}$ using AGD package in R with the United States as a reference population. Generate 150 subjects such that: a) 50 from each group, and b) 25 males and 25 females in each group
2. All subjects height data were simulated for ages 5-20 years with six months increase
3. The simulated data in steps 1 and 2 was used in SIM package in R to get estimates of $\gamma^{SIM}$, $\beta^{SIM}$, $\sigma^{2, SIM}$, and $\Sigma^{SIM}$
4. Using these estimates, a new $y_{ij}$ for height was generated following the equation:

$$y_{ij}^{new} = \gamma_{i2}^{SIM} + BS(exp(\gamma_{i3}^{SIM})(t_{ij} - \gamma_{i1}^{SIM}))\beta^{SIM} + noise(0, \sigma^{2, SIM}),$$

Where, the random noise generated from a normal distribution with mean 0 and variance $\sigma^{2, SIM}$, and 'BS' is for the B-spline function with 5 knots. This random noise can interpreted as a measurement error or the errors that are still unaccounted for by the SIM parameters. Step 4 can be run many times depending on the desired number of realizations achieved for the simulated dataset. We run it five times to generate five realizations. Figure ** a) and b) plots the data generated from the AGD package and one realization after adding the random noise, respectively.
In the evaluation of model performance for the simulated datasets, we will use the estimates from SIM package at Step 3: $\gamma^{SIM}$, $\beta^{SIM}$, $\sigma_{\gamma}^{SIM}$, and $\Sigma^{SIM}$, as true estimates of all model parameters.

As similar to the real data example, we implemented both the original SIM model and the Bayesian SIM model with Two spline functions, B-spline and natural cubic spline, on the simulated data. The number of knots was assigned to five. We checked the DIC and the RMSPE values for the most parsimonious model and for checking the predictive model ability, respectively. It appeared that the Bayesian SIM model with natural cubic spline was performing better in both of these criteria.

7 Results:

In this section, we present the results of our analysis. We compared the frequentist SIM model with Bayesian SIM model applying on ADHD patient data. Here, age at peak velocity (APV) is 12.38 years, and peak velocity(PV) is 8.76 cm.

![Figure 4: Comparison between Age at peak velocity in bayesian and frequentist method](image)

In below we reported the results from the Bayesian SIM model with basis spline. Subject-specific posterior mean estimates of $\gamma$ plotted in Figure 5. It indicates that the correlation between the horizontal and vertical shift is negative and the correlation between the horizontal effect and the stretch effect are negative.

We were interested to see the pattern of the growth velocity of the ADHD patient children and the effect of medication on their growth. The scatter plot of the centered posterior mean effect, which indicates the deviation from the centered effect.
Figure 5: posterior mean with 95% confidence interval for subject-specific parameter (a) Tempo ($\gamma_1$), (b) Size ($\gamma_2$) and (c) Velocity ($\gamma_3$).

Figure 6: Scatter plot of the posterior mean of the subject-specific effect, gamma_1 is the horizontal shift, gamma_2 is the vertical shift and gamma_3 is the horizontal shift.
8 Model Selection and Predictive ability:

We implemented both the original SIM model and the Bayesian SIM model on the ADHD data. Two spline functions, B-spline and natural cubic spline, used in Bayesian SIM model. In all spline functions, five knots used. We checked the DIC values for the two spline functions for the Bayesian SIM model to check the model adequacy. Smaller the DIC is indicative of a parsimonious model. We calculated the root mean square prediction error (RMSPE) for both the SIM model and the Bayesian SIM model to check their predictive ability. It appeared that the Bayesian SIM model with natural cubic spline was performing better in both of these criteria (Table 2).

For checking the predictive ability of Bayesian SIM model, we adopted the cross-validation method where four patients with two early start ages of stimulant medication from each sex and two late start ages from each sex were randomly left out from the analysis as a validation sample. The remaining 193 patients constructed the derivation sample were analyzed using the Bayesian SIM model to generate the mean predictive curve. Figure 6 plots the mean predictive curve with 95% credible interval from the derivation sample. The four patients from the validation sample plotted in the same figure with red color for females and blue for males. The results ensure that the randomly picked four patients with various start ages of stimulant medication and sex to lie within 95% credible interval of the mean predictive curve.

Figure 7: Predicted mean growth curve for the 194 ADHD patients with 95% credible interval. The remaining four patients were plotted with red color for females and blue for males

8.1 Sensitivity Analysis

In Bayesian inference, it is recommended to check the validity of the posterior estimates for the choice of other prior distributions, or other hyper-parameters values. Sensitivity analysis ensures this purpose. Even though we assigned noninformative flat prior distributions to
all parameters, we checked the sensitivity of the estimates by changing the values for hyper-
parameters. We also did the sensitivity analysis following the method proposed by [Kass et al.,
1989] called perturbation function. We perturbed the model with different spline function (with
and without covariates) and analyzed the sensitivity of the model. Two spline functions: basis
spline and natural cubic spline functions considered for this purpose.

9 Discussion

The original SIM model is effective in explaining subject-specific deviations in terms of size,
tempo, and velocity from the underlying shape curve. An R package "SIM" was made available
to implement this model. The model is currently unable to produce standard error estimates
for subject-specific deviations for size, tempo, and velocity parameters. Although bootstrap
methods are often used to estimate the standard error, this is a post-hoc analysis and often
fails to estimate the true uncertainty specifically when the model has a complex structure,
i.e., a non-linear mixed effects model. Estimating the standard error for estimates of subject-
specific deviations for size, tempo, and velocity in Bayesian SIM model comes naturally from
the posterior samples for the subject-specific parameter ($\gamma$). Although we considered vague
prior distributions in our analyses, the model is flexible to assimilate information from previous
studies or experts opinion through specifying informative prior distributions.

We applied the Bayesian SIM model with two spline functions and with an optimum number
of knots (i.e., 5) to real and simulated data sets, and compared the results with the original
SIM model. We observed that for both data sets, Bayesian SIM model with natural cubic
spline function has better predictive ability than the original SIM model. With the real data
application.

In our future works, we plan to extend the Bayesian SIM model in finding clustering patterns
in shape invariant parameters tempo ($\gamma_1$), size ($\gamma_2$), and velocity ($\gamma_3$). In many applications
of shape invariant models in growth curve modeling, finding the group of subjects with similar
growth patterns in terms of size, tempo and velocity have real significance. Another extension
can include but not limited to, extending the model to free knot natural cubic spline, where
we will utilize the Bayesian adaptive regression spline (BARS). It shown earlier that BARS
provide a parsimonious fits [DiMatteo et al., 2001].

10 Acknowledgement

The project described was supported by the National Center for Advancing Translational Sci-
ences of the National Institutes of Health, under Award Number 5UL1TR001425-04. The
content is solely the responsibility of the authors and does not necessarily represent the official
views of the NIH. Coauthor (RS) supported by grant K25 HL125954 from the National Heart,
Lung and Blood Institute of the National Institutes of Health.
References

Elja Arjas, Liping Liu, and Niko Maglaperidze. Prediction of growth: a hierarchical bayesian approach. *Biometrical journal*, 39(6):741–759, 1997.

Daniel Barry. A bayesian model for growth curve analysis. *Biometrics*, pages 639–655, 1995.

Ken J Beath. Infant growth modelling using a shape invariant model with random effects. *Statistics in medicine*, 26(12):2547–2564, 2007.

Catherine S Berkey. Comparison of two longitudinal growth models for preschool children. *Biometrics*, pages 221–234, 1982.

Catherine S Berkey and Nan M Laird. Nonlinear growth curve analysis: estimating the population parameters. *Annals of human biology*, 13(2):111–128, 1986.

Jennifer Brunet, Catherine M Sabiston, Michael Chaiton, Tracie A Barnett, Erin O’Loughlin, Nancy CP Low, and Jennifer L O’Loughlin. The association between past and current physical activity and depressive symptoms in young adults: a 10-year prospective study. *Annals of epidemiology*, 23(1):25–30, 2013.

Tim J Cole, Malcolm DC Donaldson, and Yoav Ben-Shlomo. Sitar?a useful instrument for growth curve analysis. *International journal of epidemiology*, 39(6):1558–1566, 2010.

Earl W Count. growth patterns of the human physique: an approach to kinetic anthropometry: part i. *Human Biology*, 15(1):1–32, 1943.

Ilaria DiMatteo, Christopher R Genovese, and Robert E Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001.

David B Dunson. Commentary: practical advantages of bayesian analysis of epidemiologic data. *American journal of Epidemiology*, 153(12):1222–1226, 2001.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models. Technical report, EERI Research Paper Series, 2004.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

Diklah Geva, Lidush Goldschmidt, David Staffer, and Nancy L Day. A longitudinal analysis of the effect of prenatal alcohol exposure on growth. *Alcoholism: Clinical and Experimental Research*, 17(6):1124–1129, 1993.

Melissa J Griffin, Jeffrey D Wardell, and Jennifer P Read. Recent sexual victimization and drinking behavior in newly matriculated college students: A latent growth analysis. *Psychology of Addictive behaviors*, 27(4):966, 2013.

Christopher Hertzog and John R Nesselroade. Assessing psychological change in adulthood: an overview of methodological issues. *Psychology and aging*, 18(4):639, 2003.

Joshua J Jackson, Patrick L Hill, Brennan R Payne, Brent W Roberts, and Elizabeth AL Stine-Morrow. Can an old dog learn (and want to experience) new tricks? cognitive training increases openness to experience in older adults. *Psychology and aging*, 27(2):286, 2012.
Rachel M Jenss and Nancy Bayley. A mathematical method for studying the growth of a child. *Human Biology*, 9(4):556, 1937.

Johan Karlberg. On the modelling of human growth. *Statistics in medicine*, 6(2):185–192, 1987.

Robb E Kass, Luke Tierney, and Joseph B Kadane. Approximate methods for assessing influence and sensitivity in bayesian analysis. *Biometrika*, 76(4):663–674, 1989.

Robert J Kuczmarski. 2000 cdc growth charts for the united states; methods and development. 2002.

Paul C Lambert, Keith R Abrams, David R Jones, Aidan WF Halligan, and Andrew Shennan. Analysis of ambulatory blood pressure monitor data using a hierarchical model incorporating restricted cubic splines and heterogeneous within-subject variances. *Statistics in medicine*, 20(24):3789–3805, 2001.

WH Lawton, EA Sylvestre, and MS Maggio. Self modeling nonlinear regression. *Technometrics*, 14(3):513–532, 1972.

Maria M Llabre, Susan Spitzer, Scott Siegel, Patrice G Saab, and Neil Schneiderman. Applying latent growth curve modeling to the investigation of individual differences in cardiovascular recovery from stress. *Psychosomatic Medicine*, 66(1):29–41, 2004.

S Milani, A Bossi, and E Marubini. Individual growth curves and longitudinal growth charts between 0 and 3 years. *Acta Paediatrica*, 78(s350):95–104, 1989.

Marie-Louise Newell, Mario-Cortina Borja, and Catherine Peckham. Height, weight, and growth in children born to mothers with hiv-1 infection in europe. *Pediatrics*, 111(1):e52–60, 2003.

Ken KL Ong, Michael A Preece, Pauline M Emmett, Marion L Ahmed, and David B Dunger. Size at birth and early childhood growth in relation to maternal smoking, parity and infant breast-feeding: longitudinal birth cohort study and analysis. *Pediatric research*, 52(6):863, 2002.

World Health Organization et al. *WHO child growth standards: length/height for age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age, methods and development*. World Health Organization, 2006.

José C Pinheiro and Douglas M Bates. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56, 2000.

C Radhakrishna Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17, 1958.

Steven K Schmidt, Stephen Simkins, and Martin Alexander. Models for the kinetics of biodegradation of organic compounds not supporting growth. *Applied and Environmental Microbiology*, 50(2):323–331, 1985.

Takao Shohoji, Kouji Kanefuji, Takahiro Sumiya, and TAO Qin. A prediction of individual growth of height according to an empirical bayesian approach. *Annals of the Institute of Statistical Mathematics*, 43(4):607–619, 1991.
Andrew J Simpkin, Adrian Sayers, Mark S Gilthorpe, Jon Heron, and Kate Tilling. Modelling height in adolescence: a comparison of methods for estimating the age at peak height velocity. *Annals of human biology*, 44(8):715–722, 2017.

Kert Viele, Mark Lancaster, and Robin L Cooper. Self-modeling structure of evoked postsynaptic potentials. *Synapse*, 60(1):32–44, 2006.

Sten P Willemsen, Paul HC Eilers, Régine PM Steegers-Theunissen, and Emmanuel Lesaffre. A multivariate bayesian model for embryonic growth. *Statistics in medicine*, 34(8):1351–1365, 2015.

John Wingerd. The relation of growth from birth to 2 years to sex, parental size and other factors, using rao’s method of the transformed time scale. *Human biology*, pages 105–131, 1970.

John Wishart. Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika*, 30(1/2):16–28, 1938.

Laura B Zahodne, Devangere P Devanand, and Yaakov Stern. Coupled cognitive and functional change in alzheimer’s disease and the influence of depressive symptoms. *Journal of Alzheimer’s Disease*, 34(4):851–860, 2013.

MH Zwietering, Il Jongenburger, FM Rombouts, and K Van’t Riet. Modeling of the bacterial growth curve. *Applied and environmental microbiology*, 56(6):1875–1881, 1990.
Normality Assumption: check the normality assumption of the subject-specific parameter, we checked the Q-Q plot. The Q-Q plot indicates the normality assumption is satisfied.

Autocorrelation:

Figure 8: normality assumptions plot for Basis spline for the subject specific parameters (a) $\gamma_1$ (b) $\gamma_2$ and (c) $\gamma_3$.

Autocorrelation:

Figure 9: Auto correlation plot for Basis spline for the subject specific parameters (a) $\gamma_1$ (b) $\gamma_2$ and (c) $\gamma_3$.

Trace plot: Trace plot indicates the convergence of the model parameters. We randomly selected a person and checked the trace plot, which indicates the model parameters converged.

Figure 10: Auto correlation plot for Basis spline for the subject specific parameters (a) $\gamma_1$ (b) $\gamma_2$ and (c) $\gamma_3$. 