The International Journal of Biostatistics

Volume 6, Issue 1

2010

Article 4

Lack of Fit in Self Modeling Regression: Application to Pulse Waveforms

Lyndia C. Brumback^{*}

Douglas Tommet[†]

Richard Kronmal[‡]

^{*}University of Washington, lynb@u.washington.edu

[†]Hebrew SeniorLife, dtommet@myuw.net

[‡]University of Washington, kronmal@u.washington.edu

Lack of Fit in Self Modeling Regression: Application to Pulse Waveforms*

Lyndia C. Brumback, Douglas Tommet, and Richard Kronmal

Abstract

Self modeling regression (SEMR) is an approach for modeling sets of observed curves that have a common shape (or sequence of features) but have variability in the amplitude (y-axis) and/or timing (x-axis) of the features across curves. SEMR assumes the x and y axes for each observed curve can be separately transformed in a parametric manner so that the features across curves are aligned with the common shape, usually represented by non-parametric function. We show that when the common shape is modeled with a regression spline and the transformational parameters are modeled as random with the traditional distribution (normal with mean zero), the SEMR model may surprisingly suffer from lack of fit and the variance components may be over-estimated. A random effects distribution that restricts the predicted random transformational parameters to have mean zero or the inclusion of a fixed transformational parameter improves estimation. Our work is motivated by arterial pulse pressure waveform data where one of the variance components is a novel measure of short-term variability in blood pressure.

KEYWORDS: functional data, nonlinear mixed effects models, self-modeling

*The authors thank the reviewers for helpful comments. The authors also thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at www.mesa-nhlbi.org. This research was supported in part by contracts N01-HC-95159 and N01-HC-95168 from the National Heart, Lung, and Blood Institute, and Career development funds from the University of Washington, Department of Biostatistics.

1 Introduction

The arterial pulse pressure waveform (or pulse waveform) represents a continuous measure of blood pressure throughout the entire cardiac cycle. It is a combination of a pressure wave initiated by the pumping action of the heart and pressure waves reflected from multiple sites located in the arterial system. The maximum and minimum of the waveform represent diastolic and systolic blood pressure, which are predictors of cardiovascular disease and targets of treatment. The possibility that other features of the waveform add information about the cardiovascular disease process has not been fully explored (Duprez and Cohn, 2006).

Pulse waveform data have recently been collected from approximately 6000 participants of the Multi-Ethnic Study of Atherosclerosis (MESA), a prospective study designed to investigate the prevalence, correlates and progression of subclinical cardiovascular disease (Bild *et al.*, 2002). Pulse waveforms from one of the MESA participants are shown in Figure 1; there are 26 observed waveforms or curves. The curves have a similar shape but contain variability in amplitude (y -direction) and phase or time (x -direction). For example, each observed curve has a large peak around 0.2 seconds but the pressure at the peak and the width or timing of the peak vary across curves. Pulse waveform data from four more MESA participants are shown in Figure 2. Within a participant, the observed curves have a similar shape but exhibit variability in amplitude and phase. It appears that some participants have more variability in amplitude compared to others. Researchers are interested in whether features of the common shape, or the variability in amplitude or timing of features among pulse waveforms within a participant predict cardiovascular disease or are associated with risk factors for cardiovascular disease. In this article, we estimate the common shape and the variability among pulse waveforms from MESA participants using self modeling regression.

Self modeling regression (SEMOR) is a method for analyzing sets of observed curves. SEMOR is based on the relatively simple assumption that the set of curves has a common shape and the x and y axes can be separately transformed, often in a parametric manner, for each curve so that the features across curves are aligned with the common shape. The common shape is usually represented by a non-parametric function. SEMOR is attractive because it accounts for the variability in the timing and amplitude of features among curves during estimation of the common shape. In contrast, a cross-sectional average of unaligned data may not be representative of the underlying com-

mon shape. The assumption of the SEMOR model that all curves share a common shape allows pooling of information and improves efficiency. In various applications, the common shape has been modeled with regression splines (Lawton *et al.*, 1972; Ladd and Lindstrom, 2000; Brumback and Lindstrom, 2004; Beath, 2007), kernel smoothers (Kneip and Gasser, 1988), free-knot regression splines (Lindstrom, 1995), smoothing splines (Wang *et al.*, 2003), and penalized splines (Altman and Villarreal, 2004; Coull and Staudenmayer, 2004; Telesca and Inoue, 2008). Typically, the parameters that align the curves are treated as random.

Figure 1: Example of pulse waveform data from one participant.

Figure 2: Pulse waveform data and results from 4 MESA participants. Each participant has approximately 30 observed curves. Grey circles represent observed data. The black line represents the estimate of the common shape from **DF** with 20 knots. Estimates of the common shape from the other models with 20 knots are similar (not shown). Within a participant, estimates of amplitude shift variability σ_a from **DF** and **DZ** are the same, regardless of the number of knots, and are relatively similar to the empirical estimates of variability. For participant D, σ_a from **DF** and **DZ** is 2.1 mmHg, and the standard deviation of the pressures at time 0 and the standard deviation of the maximum pressures are 1.8 and 1.7. Estimates from **DT** vary with the number of knots and are often larger than estimates from **DF** or **DZ** and the empirical estimates. For participant D, they are 14.7, 10.9, and 4.9 mmHg when the model contains 15, 20, and 25 knots, respectively.

We apply SEMOR to pulse waveform data from each MESA participant separately and assume that the features across curves within a participant are aligned by simply shifting the y axis and scaling the x axis. We model the common shape with a regression spline, and we model the shift and scale parameters as random using three different distributions that have been proposed for SEMOR. We show that one of the distributions can lead to poor estimation of the variance components. This is important for application to the MESA pulse waveform data where one of the variance components is a novel measure of average short term variability in blood pressure as well as other applications of SEMOR where the variance components may be of interest.

The article is organized as follows. In section 2, we review the SEMOR models. We apply the models to pulse waveform data from MESA participants in section 3. In section 4, we discuss our results and future work.

2 Self modeling regression (SEMOR)

A SEMOR model for a set of observed curves consists of: a common shape, parameters which align the observed curves with the common shape, and distributions for random parameters.

2.1 Common shape: Natural cubic regression spline

In this article, we model the common shape with a natural cubic regression spline. A regression spline is a piecewise polynomial with continuity constraints. The polynomial pieces are joined at a specified sequence of knots $(\gamma_1, \gamma_2, \dots, \gamma_Q)$. A *cubic* regression spline is made up of cubic polynomials and is constrained to have two continuous derivatives at the knots. A *natural* cubic regression spline is further constrained to have its second derivative equal to zero at the end knots (γ_1 and γ_Q), which forces the spline to be linear outside the end knots and often results in nice behavior of the spline at the end knots.

A natural cubic regression spline with Q distinct knots requires Q basis functions, and thus Q parameters or spline coefficients to be estimated. If we let γ denote the set of knots $(\gamma_1, \gamma_2, \dots, \gamma_Q)$, $\beta \equiv (\beta_1, \dots, \beta_Q)^T$ denote the length Q vector of spline coefficients to be estimated, and $\mathbf{B}(t, \gamma) \equiv [B_1(t, \gamma) \ B_2(t, \gamma) \ \dots \ B_Q(t, \gamma)]$ denote the $1 \times Q$ matrix of basis functions evaluated at t , then a natural cubic regression spline evaluated at t can be

written as:

$$g(t, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{B}(t, \boldsymbol{\gamma}) \boldsymbol{\beta}.$$

Eubank (1988) and de Boor (1978) describe various properties of natural cubic regression splines and basis functions. Appendix A reviews properties that we use in this article.

2.2 SEMOR model with amplitude-shift and time-scale parameters

In this article, we assume that features of the curves are aligned after simply shifting the curves in amplitude and scaling them in time, although more flexible transformations are possible within the SEMOR framework. The SEMOR model with amplitude-shift and time-scale (or phase-scale) parameters can be written as:

$$y_{j,k} = a_{j,\text{sh}} + g(\exp(p_{j,\text{sc}}) t_{j,k}, \boldsymbol{\beta}, \boldsymbol{\gamma}) + e_{j,k}, \quad j = 1, \dots, M; \quad k = 1, \dots, n_j \quad (1)$$

where $y_{j,k}$ is the k^{th} response for the j^{th} curve evaluated at $t_{j,k}$; and $a_{j,\text{sh}}$ shifts the j^{th} curve in amplitude and $p_{j,\text{sc}}$ scales the j^{th} curve in time so that it is aligned with the common shape g . Here, g is modeled by a natural cubic regression spline with coefficients $\boldsymbol{\beta}$ and knots $\boldsymbol{\gamma}$. We model the errors, $e_{j,k}$, as normal with mean 0 and constant variance σ^2 , $e_{j,k} \sim \mathcal{N}(0, \sigma^2)$, although more flexible distributions can be accommodated within the model framework.

The transformational parameters that align the curves and/or the shape function of SEMOR models usually need to be constrained to ensure identifiability. Lindstrom (1995), Kneip and Gasser (1998), and Ke and Wang (2001) discuss identifiability of various SEMOR models. In model 1, exponentiation is used to ensure that $\exp(p_{j,\text{sc}})$ are positive and to avoid compensation of a negative scaling of each of $p_{j,\text{sc}}$ with a negative scaling of the knots and a transformation of the spline coefficients (property P2 of Appendix A). It is necessary to further constrain the set of time scale parameters to avoid compensation of a positive scaling of $p_{j,\text{sc}}$ with a positive scaling of the knots (property P1). It is also necessary to constrain the set of amplitude shift parameters to avoid compensation of a shift to each of $a_{j,\text{sh}}$ with a transformation of the set of spline coefficients (property P3).

We model the amplitude-shift and time-scale parameters as random. In the next section, we review three distributions that have been proposed for

SEMOR and describe their effect on identification and estimation of model parameters.

2.3 Random shift and scale parameters

2.3.1 Traditional distribution: DT

It is most common to model the distribution of the random transformational parameters of a SEMOR model as multivariate normal with mean zero. In our case (model 1), this means the amplitude-shift and phase-scale parameters are modeled as:

$$(a_{j,\text{sh}}, p_{j,\text{sc}})^T \sim \mathcal{N}_2(\mathbf{0}, \mathbf{V}) \quad (2)$$

where $\mathbf{0}$ is a vector of zeros; and $\mathbf{V}$ is a positive-definite covariance matrix:

$$\mathbf{V} = \begin{bmatrix} \sigma_a^2 & \sigma_{ap}^2 \\ \sigma_{ap}^2 & \sigma_p^2 \end{bmatrix} \quad (3)$$

The variance component, σ_a , represents the variability of amplitude shifts. It is the focus of analyses in Sections 3.

The distributional assumption in equation 2 can be written equivalently as

$$\mathbf{DT} : \text{vec}([\mathbf{a}_{\text{sh}} \ \mathbf{p}_{\text{sc}}]) \sim \mathcal{N}_{2M}(\mathbf{0}, \mathbf{V} \otimes \mathbf{I}_M) \quad (4)$$

where $\mathbf{a}_{\text{sh}} \equiv (a_{1,\text{sh}}, \dots, a_{M,\text{sh}})^T$, $\mathbf{p}_{\text{sc}} \equiv (p_{1,\text{sc}}, \dots, p_{M,\text{sc}})^T$, $\text{vec}([\mathbf{a}_{\text{sh}} \ \mathbf{p}_{\text{sc}}])$ stacks $\mathbf{a}_{\text{sh}}$ and $\mathbf{p}_{\text{sc}}$ into a single vector, $\otimes$ denotes the Kronecker product, and $\mathbf{I}_M$ is a dimension M identity matrix. We denote the distributional assumption in equation 4 as **DT**. Equation 4 is more convenient than equation 2 for comparing **DT** to the distributions in sections 2.3.2 and 2.3.3.

Model 1 with **DT** is a nonlinear mixed effects model with Q (the number of knots for the spline) fixed effects: β ; random effects: $a_{j,\text{sh}}$ and $p_{j,\text{sc}}$; and variance components: σ_a , σ_p , σ_{ap} . The parameters can be estimated by maximum likelihood using existing software. For the application in this paper, we use R (R Development Core Team, 2008) with the **splines** package and the **nlme** package (Pinheiro *et al.*, 2008) that implements the approximate maximum likelihood approach described by Lindstrom and Bates (1990).

Model 1 with **DT** is identifiable. If the common shape is correctly specified (by a natural cubic regression spline with knots γ), then the mean of the predicted amplitude-shift parameters, $\mathbf{1}^T \hat{\mathbf{a}}_{\text{sh}}/M$ where $\mathbf{1}$ is a vector of

ones, is zero. The mean of the predicted time-scale parameters is also zero so $\exp(\mathbf{1}^T \hat{\mathbf{p}}_{\text{SC}}/M)$ is approximately one. Thus, the estimated shape $g(t, \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma})$ is approximately the average curve.

Despite its apparent flexibility, model 1 with **DT** may suffer from lack of fit. If the common shape is poorly approximated by a regression spline with the specified knots, then the variance component for the time scale parameters, σ_p , will be over-estimated. Furthermore, if the time scale parameters are allowed to be correlated with the amplitude shift parameters ($\mathbf{V}$ is unstructured), the variance component for the amplitude shift parameters, σ_a , will also be over-estimated. This happens in simulations (not shown) and appears to happen in application to the pulse waveform data from MESA (section 3).

While predicted random effects are restricted to have mean zero (across experimental units) in usual mixed effects models (Searle, 1997), the predicted random time-scale parameters in model 1 with **DT** are not restricted to have mean zero; however, a well-chosen set of knots will ensure that the mean is approximately zero. If the knots are poorly specified, then the set of time-scale parameters in model 1 with **DT** can move the knots by having a non-zero mean (properties P1 and P2 of Appendix A) as well as align the curves. To see this, suppose the true common shape can be specified more accurately using knots $c\boldsymbol{\gamma}$, for some $c > 0$, rather than the specified knots $\boldsymbol{\gamma}$. The non-parametric component of the SEMOR model can then be appropriately modeled as

$$g(\exp(p_{j,\text{SC}}^*) t_{j,k}, \boldsymbol{\beta}, c\boldsymbol{\gamma}) \quad (5)$$

where $p_{j,\text{SC}}^*$ is normal with mean 0. By P1, equation 5 is equal to

$$g(\exp(p_{j,\text{SC}}^* - \log c) t_{j,k}, \boldsymbol{\beta}, \boldsymbol{\gamma}). \quad (6)$$

The model actually implemented has the non-parametric component

$$g(\exp(p_{j,\text{SC}}) t_{j,k}, \boldsymbol{\beta}, \boldsymbol{\gamma})$$

where $p_{j,\text{SC}}$ is assumed normal with mean zero. But, according to equation 6, the correct mean is not 0 but $-\log c$. This misspecification of the mean of the random effects $p_{j,\text{SC}}$ leads to overestimation of their variance.

Unfortunately, it is not possible to analytically show the effect of lack of fit on estimation of the variance components or predicted random effects from the SEMOR model because the SEMOR model does not have closed form estimators. However, closed form estimators exist for mixed effects models

with a similar form of lack of fit. In Appendix B, we use these similar models to analytically show that lack of fit can lead to a set of predicted random effects that does not have mean zero and over-estimation of the variance component.

2.3.2 Distribution with a fixed time-scale parameter: DF

One approach to improve estimation is to add a fixed time-scale parameter to the model to unlink the alignment of the curves from the movement of the knots: the random parameters align the curves and the fixed time-scale parameter determines the locations of the knots of the common shape function. A similar method was used by Ladd and Lindstrom (2000) for modeling two-dimensional response curves. With this approach, the transformational parameters are modeled as

$$(a_{j,\text{sh}}, p_{j,\text{sc}})^T \sim \mathcal{N}_2((0, \phi_{\text{sc}})^T, \mathbf{V})$$

or equivalently,

$$\mathbf{DF} : \text{vec}([\mathbf{a}_{\text{sh}} \ \mathbf{p}_{\text{sc}}]) \sim \mathcal{N}_{2M}((0, \phi_{\text{sc}})^T \otimes \mathbf{1}_M, \mathbf{V} \otimes \mathbf{I}_M) \quad (7)$$

where $\mathbf{V}$ is a positive-definite covariance matrix with dimension 2×2 as before in equation 4, and ϕ_{sc} is the fixed time transformation (scale) parameter. Like model 1 with $\mathbf{DT}$, model 1 with $\mathbf{DF}$ is a nonlinear mixed effects model that can be estimated by maximum likelihood using existing software.

2.3.3 Distribution that restricts the predicted random effects to have mean zero: DZ

Another approach to unlink the alignment of curves from the movement of knots is to restrict the predicted random shift and scale parameters to have mean zero. A similar method was suggested by Brumback and Lindstrom (2004) for self modeling with flexible time transformations. For this approach, each set of M transformational parameters is defined as a special function of $M - 1$ unconstrained parameters. Specifically, we let $\mathbf{a}_{\text{sh}} = \mathbf{Z}\mathbf{v}_{\text{sh}}$ and $\mathbf{p}_{\text{sc}} = \mathbf{Z}\mathbf{u}_{\text{sc}}$ where $\mathbf{v}_{\text{sh}}$ and $\mathbf{u}_{\text{sc}}$ are each length $M - 1$ vectors of unconstrained parameters and $\mathbf{Z}$ is an $M \times (M - 1)$, full column-rank matrix whose columns each sum to zero. Thus, each set of M transformational parameters has mean zero.

Similar to Brumback and Lindstrom, we obtain $\mathbf{Z}$ from the singular value decomposition of $\mathbf{Z}\mathbf{Z}^T \equiv -1/(M-1)\mathbf{1}_M\mathbf{1}_M^T + M/(M-1)\mathbf{I}_M$ and we model the unconstrained parameters as:

$$\text{vec}([\mathbf{v}_{\text{sh}} \ \mathbf{u}_{\text{sc}}]) \sim \mathcal{N}_{2M-2}(\mathbf{0}, \mathbf{V} \otimes \mathbf{I}_{M-1})$$

where $\mathbf{V}$ is as before. Thus, the transformational parameters

$$\text{vec}([\mathbf{a}_{\text{sh}} \ \mathbf{p}_{\text{sc}}]) \equiv (\mathbf{I}_2 \otimes \mathbf{Z}) \text{vec}(\mathbf{v}_{\text{sh}} \ \mathbf{u}_{\text{sc}})$$

have distribution

$$\mathbf{DZ}: \text{vec}([\mathbf{a}_{\text{sh}} \ \mathbf{p}_{\text{sc}}]) \sim \mathcal{N}_{2M}(\mathbf{0}, \mathbf{V} \otimes (\mathbf{Z}\mathbf{Z}^T)) \quad (8)$$

and are correlated within a curve as well as across curves, unlike in equations 4 and 7. The correlation between parameters across curves gets closer to zero as the number of curves increases.

Model 1 with $\mathbf{DZ}$ is still a nonlinear mixed effects model and can be estimated by maximum likelihood. However, the R `nlme` package and other existing software packages assume that random effects are independent among experimental units (curves). For our application, we modified the R `nlme` `pd-Kron` covariance class to create a random effect covariance structure of the form $\mathbf{V} \otimes \mathbf{I}_{M-1}$ and used a dummy group variable equal to one for all curves to force `nlme` to fit random effects that may be correlated among experimental units.

The estimators of model 1 with distributional assumption $\mathbf{DZ}$, like those of model 1 with $\mathbf{DT}$ and $\mathbf{DF}$, do not have closed form. In Appendix B, we analytically show that restricting predicted random effects to have mean zero can improve estimation over the traditional distribution in a simple mixed effects model with lack of fit.

3 Application to pulse waveforms

3.1 Methods

To illustrate differences in the performance of these three SEMOR variants, we apply each to pulse waveform data collected from all participants at one of the MESA centers (954 participants) during their baseline exam (Bild *et*

al., 2002). Each participant has approximately 30 observed waveforms that were obtained by first sampling pressure values every 0.005 seconds from the participant's right radial artery for a total of 30 seconds and calibrating the values to blood pressure with the HDI/PulseWave™ CR-2000 Research CardioVascular Profiling Instrument, and then segmenting the data into observed waveforms that correspond to the cardiac cycle. The focus of our analysis is estimation of the amplitude shift variability from the set of observed waveforms for each participant.

For our application, we include an index i to model 1 and the distributional assumptions to represent data and parameters for the i^{th} participant, $i = 1, \dots, 954$. We choose to place the knots γ_i for the common shape for the i^{th} participant equally spaced in the range of the observed time values for that participant. Visual inspection of fits to data from several participants suggests that 20 equally spaced knots adequately captures the features (peaks and valleys) of the data. More equally spaced knots do not seem necessary for the critical points and fewer than 20 equally spaced knots do not seem enough.

To investigate the sensitivity of estimation of the amplitude shift variability to the number of knots, we also fit models with 15 and 25 equally spaced knots in the range of the observed time values for each participant. Thus, we fit a total of 9 models to data from each participant. Starting values for the spline coefficients are based on a natural cubic regression spline fit to data from the participant. The starting value for the fixed time-scale parameter for **DF** is set at 0. No other starting values are provided.

In addition to estimating the variability in amplitude by SEMOR, we estimate the variability empirically. For each participant, we estimate the standard deviation of pressures at time 0 and the standard deviation of the maximum pressures.

3.2 Results

Table 1 summarizes the number of participants for which the different models converge based on initial starting values. Convergence is very good for **DF**; the model with 20 knots converges for 953 of the 954 participants, and the models with 15 and 25 knots converge for 952 and 950 participants, respectively. Convergence for **DT** is also reasonably good, but convergence for **DZ** is too unreliable for comfort with these data. The latter model converges for 690-722 (72% to 76%) of the 954 participants. For the one participant where **DF** with 20 knots does not converge based on initial starting values, fitting

that model with a diagonal structure for $\mathbf{V}$ first and using the results as starting values for fitting the model with an unstructured $\mathbf{V}$ yields convergence.

Distribution	Knots	N
DT	15	944
DT	20	950
DT	25	952
DF	15	952
DF	20	953
DF	25	950
DZ	15	722
DZ	20	702
DZ	25	690

Table 1: Convergence of models fit to pulse waveform data from 954 participants. N is the number of participants where the models converge based on initial starting values. The model with **DF** and 20 knots converges for 953 participants. All 9 models converge for 518 participants. The models with the traditional assumption **DT** and the models that include a fixed time-scale parameter **DF** (6 models) converge for 934 participants.

Estimates of the amplitude shift variability from **DT**, **DF**, and **DZ** and with 15, 20, and 25 knots are compared in Figures 3 and 4. The estimates from **DF** and **DZ** are similar while estimates from **DT** are often greater, regardless of the number of knots (Figure 3). Estimates from **DT** vary with the number of knots while estimates from **DF** and **DZ** are relatively constant with 15, 20, and 25 knots (Figure 4). Results for the time scale variability are similar; plots of estimates of σ_p (not shown) have the same patterns as σ_a in Figures 3 and 4.

The empirical and model-based estimates of variability are compared in Figure 5. The empirical estimates are highly correlated with estimates from **DF** with 20 knots (which are similar to estimates from **DF** with 15 and 25 knots and also from **DZ** with 15, 20, and 25 knots) compared to estimates from **DT**.

Figure 3: Comparison of estimates of amplitude shift variability σ_a from MESA pulse waveform data from **DT**, **DF**, and **DZ**. Left panels show estimates from models with 15 knots, middle panels show estimates from models with 20 knots, right panels show estimates from models with 25 knots. Top panels compare estimates from **DT** and **DF**, and bottom panels compare estimates from **DF** and **DZ**. The number of estimates per panel is 518, the number of participants where all models converge. Estimates from **DF** and **DZ** are similar while estimates from **DT** are often greater, regardless of the number of knots.

Figure 4: Comparison of estimates of amplitude shift variability σ_a from MESA pulse waveform data from models with 15, 20 or 25 knots. Top panels show estimates from **DT**, and bottom panels show estimates from **DF**. Left panels compare estimates from models with 15 and 20 knots, middle panels compare estimates from models with 15 and 25 knots, and right panels compare estimates from models with 20 and 25 knots. The number of estimates per panel is 934, the number of participants where **DT** and **DF** converge. Estimates from **DT** vary with the number of knots while estimates from **DF** are relatively constant across 15, 20, and 25 knots.

Figure 5: Comparison of empirical and model-based estimates of variability in blood pressure from MESA pulse waveform data. Estimates of σ_a from **DF** with 20 knots are relatively similar to the standard deviation of the maximum pressures and the standard deviation of the pressures at time 0 (top left and middle panels). The standard deviation of the pressures at time 0 and the standard deviation of the maximum pressures are also relatively similar (top right panel). Estimates of σ_a from **DT** with 15, 20, and 25 knots are often greater than the standard deviation of the maximum pressures (bottom row).

Results for four MESA participants are displayed along with the observed data in Figure 2. The estimates of the common shape appear reasonable in that they mimic the shapes of the observed curves. The estimates of amplitude shift variability from **DF** and **DZ** are the same, regardless of the number of knots, and are relatively similar to the empirical estimates of variability. When the SEMOR model adequately describes the data, σ_a is a more efficient estimate of variability than the empirical estimates. Comparison of the data and estimates of amplitude shift variability from **DT** shows that **DT** can grossly overestimate the variability.

4 Discussion

In this article, we show that the traditional SEMOR modeling assumptions can result in poor performance of the variance component estimators in a typical application. Simple modification of the assumptions relieves the problem. We recommend including a fixed time-scale parameter in SEMOR models that include random time-scale parameters. The approach is simple to implement and yields interpretable variance components. For SEMOR models that include more than 1 or 2 random time transformational parameters, restricting the predicted random transformational parameters to have mean zero may be preferable.

More careful selection of the knots for the common shape based on extrema or inflection points or some other optimal procedure (Wold, 1974) may also alleviate the problem. However, implementing such procedures for large applications (such as 954 sets of curves for this article, or approximately 6000 sets for all of the MESA pulse waveform data) may not be feasible. Using a large number of knots may improve estimation of the variance components but could result in an estimated common shape that is too wiggly. Modeling the common shape with a penalized spline or other non-parametric function might help but natural cubic regression splines are commonly used because they are flexible, smooth approximating functions whose parameters are relatively easy to estimate.

Convergence of nonlinear mixed effects models is often an issue. In our application, starting values based on a simple algorithm yield reasonably good convergence except for the model which restricts the predicted random parameters to have mean zero. Our experience suggests that modification of the initial starting values would improve convergence.

In our application, we model the set of curves from participants separately, although the SEMOR framework can accommodate nested curves (multiple curves from multiple participants). Preliminary analyses of estimates of the amplitude shift variability σ_a from MESA participants (not shown) suggest that the measure is associated with diabetes, consistent with results regarding the association between other summary measures of pulse waveform data and diabetes (Haller *et al.*, 2004; and Schram *et al.*, 2004) and between variability in blood pressure (determined from data other than the pulse waveform) and risk factors for cardiovascular disease (Parati, 2005). The estimates of variability and common shape from the MESA pulse waveform data will be stored with other data from MESA participants for future study of their association with cardiovascular disease.

Appendix A: Properties of regression splines

Let $\mathbf{B}(t, \gamma)$ represent a $1 \times Q$ matrix of basis functions for a regression spline with Q distinct knots γ evaluated at time t , and β represent the set of regression spline coefficients.

- P1. A positive scaling of time can be compensated by a positive scaling of knots: When $\phi > 0$, $\mathbf{B}(t, \gamma) \beta = \mathbf{B}(\phi t, \phi \gamma) \beta$.
- P2. A negative scaling of time can be compensated by a negative scaling of knots and a transformation of the set of spline coefficients: When $\phi < 0$, $\mathbf{B}(t, \gamma) \beta = \mathbf{B}(\phi t, \phi \gamma) \mathbf{f}(\beta)$ for some function $\mathbf{f}$.
- P3. A shift in the spline can be compensated by a transformation of the set of spline coefficients: $\mathbf{B}(t, \gamma) \beta = \alpha + \mathbf{B}(t, \gamma) \mathbf{h}(\beta)$ for some function $\mathbf{h}$.

Appendix B: Effect of lack of fit and distributions on variance component estimation

In this appendix, we consider a simple mixed effects model that suffers from lack of fit with a distribution similar to **DT**. The simple model with distributions similar to **DT**, **DF**, and **DZ** in section 2.3 has closed form maximum likelihood estimators and thus provides a heuristic explanation of the effect of

the distributions on parameter estimation. We describe the models and then compare the estimators.

Models

Suppose a set of n repeated measures from each of M experimental units is represented by an average intercept, an average slope, and random deviations from the average slope (normal mean 0). The true model for the data can be written as:

$$\mathbf{y}_i = \alpha_F \mathbf{1}_n + (\beta_F + b_i) \mathbf{x} + \mathbf{e}_i \quad i = 1, \dots, M \quad (9)$$

where $\mathbf{1}_n$ is a length n vector of ones; $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,n})^T$ is the set of n repeated measures for experimental unit i evaluated at $\mathbf{x}$; $\mathbf{x} = (x_1, \dots, x_n)^T$ and $\mathbf{x}^T \mathbf{1}_n = 0$; α_F is a fixed parameter for the average intercept; β_F is a fixed parameter for the average slope; b_i is a random deviation from the average slope for the i^{th} experimental unit, $b_i \sim \mathcal{N}(0, \sigma_{b_F}^2)$; and $\mathbf{e}_i \sim \mathcal{N}_n(0, \sigma_F^2 \mathbf{I})$ independent of b_i . If we let $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_M)^T$, $\mathbf{b}_F = (b_1, \dots, b_M)^T$, and $\mathbf{e}^{[F]} = (\mathbf{e}_1, \dots, \mathbf{e}_M)^T$, then model 9 can also be written as:

$$\mathbf{MF} : \quad \mathbf{y} = \alpha_F \mathbf{1}_{Mn} + \beta_F (\mathbf{1}_M \otimes \mathbf{x}) + (\mathbf{I}_M \otimes \mathbf{x}) \mathbf{b}_F + \mathbf{e}^{[F]} \quad (10)$$

where $\mathbf{b}_F \sim \mathcal{N}_M(\mathbf{0}, \sigma_{b_F}^2 \mathbf{I})$ and $\mathbf{e}^{[F]} \sim \mathcal{N}_{Mn}(\mathbf{0}, \sigma_F^2 \mathbf{I})$. We denote the model in equation 10 as **MF**. Since **MF** contains a fixed parameter (β_F) to describe the mean for the set of random effects (b_i , $i = 1, \dots, M$), it can be viewed as analogous to model 1 with distribution **DF**.

Next, consider a model where the fixed parameter which specifies the mean for the set of random slopes is omitted:

$$\mathbf{MT} : \quad \mathbf{y} = \alpha_T \mathbf{1}_{Mn} + (\mathbf{I}_M \otimes \mathbf{x}) \mathbf{b}_T + \mathbf{e}^{[T]} \quad (11)$$

where α_T is the only fixed parameter; $\mathbf{b}_T$ is a random vector of length M , $\mathbf{b}_T \sim \mathcal{N}_M(\mathbf{0}, \sigma_{b_T}^2 \mathbf{I})$; and $\mathbf{e}^{[T]} \sim \mathcal{N}_{Mn}(\mathbf{0}, \sigma_T^2 \mathbf{I})$. Model **MT** suffers from lack of fit relative to **MF** when the true average slope is not zero and thus can be viewed as analogous to model 1 with distribution **DT**.

Last, consider a model where the fixed parameter which specifies the mean for the set of random slopes is omitted and the set of predicted random slopes is restricted to have mean zero

$$\mathbf{MZ} : \quad \mathbf{y} = \alpha_Z \mathbf{1}_{Mn} + (\mathbf{I}_M \otimes \mathbf{x}) \mathbf{Z} \mathbf{b}_Z + \mathbf{e}^{[Z]} \quad (12)$$

where α_Z is the only fixed parameter; $\mathbf{b}_Z$ is a random vector of length $M - 1$, $\mathbf{b}_Z \sim \mathcal{N}_{M-1}(\mathbf{0}, \sigma_{b_Z}^2 \mathbf{I})$; and $\mathbf{Z}$ is an $M \times (M - 1)$, full column-rank matrix whose columns each sum to zero; and $\mathbf{e}^{[Z]} \sim \mathcal{N}_{Mn}(\mathbf{0}, \sigma_Z^2 \mathbf{I})$. As in section 2.3.3, we construct $\mathbf{Z}$ from the singular value decomposition of $\mathbf{Z}\mathbf{Z}^T = M/(M - 1)\mathbf{I}_M - 1/(M - 1)\mathbf{1}_M\mathbf{1}_M^T$. Thus, $\mathbf{Z}\mathbf{b}_Z$ is a random vector whose mean is restricted to be zero. Model **MZ** can be viewed as analogous to model 1 with distribution **DZ**.

Comparison of variance component estimators

It is straightforward to show that the difference between the maximum likelihood estimator for the variance component from the model with lack of fit, **MT**, and from the model without lack of fit, **MF** is:

$$\hat{\sigma}_{bL}^2 - \hat{\sigma}_{bF}^2 = \hat{\beta}_F^2 \quad (13)$$

where $\hat{\cdot}$ denotes a maximum likelihood estimator. Thus, the estimator for the variance component from the model with lack of fit is larger than the estimator from the model without lack of fit except when the estimator of the fixed parameter for the average slope $\hat{\beta}_F$ is zero. Furthermore, the over-estimation converges in probability to the square of true average slope, β_F^2 . It is also straightforward to show that the over-estimation is:

$$\hat{\sigma}_{bT}^2 - \hat{\sigma}_{bF}^2 = \left(\frac{\hat{\sigma}_{bT}^2 \mathbf{1}_M^T \hat{\mathbf{b}}_T}{\hat{\mathbf{b}}_T^T \hat{\mathbf{b}}_T} \right)^2$$

which is equal to 0 when the mean of $\hat{\mathbf{b}}_T$ is zero.

The difference between the maximum likelihood estimator for the variance component from the model with lack of fit and that restricts the predicted random effects to have mean zero, **MZ**, and from the model without lack of fit, **MF**, is:

$$\hat{\sigma}_{bZ}^2 - \hat{\sigma}_{bF}^2 = \frac{n(\mathbf{x}^T \mathbf{x})^{-1}}{M(n - 1) + 1} \hat{\sigma}_F^2 - \frac{M - 1}{M(n - 1) + 1} \hat{\beta}_F^2. \quad (14)$$

The difference is a function of the maximum likelihood estimator of the fixed parameter for the average slope and the error variance under the model without lack of fit, and converges in probability to zero as the number of observations per experimental unit (n) goes to infinity.

REFERENCES

- Altman, N.S. and Villarreal J.S. (2004). Self-modeling regression for longitudinal data with time-invariant covariates. *The Canadian Journal of Statistics* **32**, 251–268.
- Beath, K.J. (2007). Infant growth modelling using a shape invariant model with random effects. *Statistics in Medicine* **26**, 2547–2564.
- Bild, D., Bluemke, D.A., Burke, G.L., Detrano, R., Roux, A.V.D., Folsom, A.R., Greenland, P., Jacobs Jr., D., Kronmal, R., Liu, K., Nelson, J.C., O’Leary, D., Saad, M.F., Shea, S., Szklo, M., and Tracy, R.P. (2002). The multi-ethnic study of atherosclerosis (MESA): Objectives and design protocol. *American Journal of Epidemiology* **156**, 871–881.
- Brumback, L.C. and Lindstrom, M.J. (2004). Self modeling regression with flexible, random time transformations. *Biometrics* **60**, 461–470.
- Coull, B.A. and Staudenmayer, J. (2004). Self-modeling regression for multivariate curve data. *Statistica Sinica* **14**, 695–711.
- de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.
- Duprez, D.A. and Cohn, J.N. (2006). Monitoring vascular health beyond blood pressure. *Current Hypertension Reports* **8**, 287–291.
- Eubank, R. (1988). *Spline smoothing and nonparametric regression*. New York: Marcel Dekker, Inc.
- Haller, M.J., Samyn, M., Nichols, W.W., Brusko, T., Wasserfall, C., Schwartz, R.F., Atkinson, M., Shuster, J.J., Pierce, G.L., and Silverstein, J.H. (2004). Radial artery tonometry demonstrates arterial stiffness in children with type 1 diabetes. *Diabetes Care* **27**, 2911–2917.
- Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications (with discussion). *Journal of the American Statistical Association* **96**, 1272–1298.

- Kneip, A. and Gasser, T. (1988). Convergence and consistency results for self-modeling nonlinear regression. *The Annals of Statistics* **16**, 82-112.
- Ladd, W.M. and Lindstrom, M.J. (2000). Self modeling regression for two-dimensional response curves. *Biometrics* **56**, 97-105.
- Lawton, W.H., Sylvestre, E.A., and Maggio, M.S. (1972). Self modeling nonlinear regression. *Technometrics* **14**, 513-532.
- Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673-687.
- Lindstrom, M.J. (1995). Self-modeling with random shift and scale parameters and a free-knot spline shape function. *Statistics in Medicine* **14**, 2009-2021.
- Parati, G. (2005). Blood pressure variability: its measurement and significance in hypertension. *Journal of Hypertension* **23**, S19-S25.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and the R Core team. (2008). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-89.
- R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Schram, M.T., Henry, R.M.A., van Dijk, R.A.J.M., Kostense, P.J., Dekker, J.M., Nijpels, F., Heine, R.J., Bouter, L.M., Westerhof, N. and Stehouwer, C.D.A. (2004). Increased central artery stiffness in impaired glucose metabolism and Type 2 diabetes. *Hypertension* **43**, 176-181.
- Searle, S.L. (1997). Built-in restrictions on best linear unbiased predictors (BLUP) of random effects in mixed models. *The American Statistician* **51**, 19-21.
- Telesca, D. and Inoue, L.Y.T. (2008). Bayesian hierarchical curve registration. *Journal of the American Statistical Association* **103**, 328-339.

- Wang, Y., Ke, C. and Brown, M. B. (2003). Shape-invariant modeling of circadian rhythms with random effects and smoothing spline ANOVA decompositions. *Biometrics* **59**, 804–812.
- Wold, S. (1974). Spline functions in data analysis. *Technometrics* **16**, 1–11.