Simulation Results on the Performance of MixTVEM Semiparametric Models

John J. Dziak

The Methodology Center, Penn State University

October 7, 2016

Abstract

In this simulation study [1] the performance of the MixTVEM model proposed by Dziak, Li, Tan, Shiffman, and Shiyko (2015) is explored. Estimation performance is fairly good, although oversmoothing (underfitting), caused by the penalty function, sometimes occurs. Model-based standard errors may outperform sandwich standard errors for this method.

Intensive longitudinal data, in which many observations are taken over time on each subject, offer many new opportunities for researchers.

[1] **Acknowledgments**. This project was supported by Awards P50 DA010075 and R21 DA024260 from the National Institute on Drug Abuse and Award R03 CA171809-01 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse, the National Cancer Institute, or the National Institutes of Health.

Sometimes, richer models may be required in order to fully exploit the potential of this data. Dziak, Li, Tan, Shiffman, and Shiyko (2015) proposed MixTVEM, a flexible finite mixture semiparametric model with supporting software, which can be used to provide a rich exploratory analysis of intensive longitudinal datasets. MixTVEM is an combination of finite mixture modeling (see, e.g., McLachlan & Peel, 2000; Muthén & Muthén, 2000; Nagin, 1999) with time-varying effects modeling (TVEM; Hastie & Tibshirani, 1993; Shiyko, Lanza, Tan, Li, & Shiffman, 2012; Tan, Shiyko, Li, Li & Dierker, 2012). As in finite mixture modeling, it assumes that the population is divided into latent classes, and that a different regression model applies within each class for the mean response variable as a function of time and other predictors. However, within each class it allows the regression coefficients in the model to vary as functions of time, as in TVEM. This combination was first proposed by Lu and Song (2012) in the Bayesian context. The frequentist approach of Dziak et al. (2015) may be easier for many researchers to use, especially since it is accompanied by a versatile R function and SAS macro for fitting the model. Dziak, Li, Tan, Shiffman, and Shiyko (2015) model the coefficient functions using penalized B-splines (Eilers & Marx, 1996), and model within-subject correlation using a modified autoregressive covariance structure with nugget effect. Overfitting is avoided by applying a quadratic penalty on the second-order differences among B-spline coefficients, which provides regularization by biasing estimated coefficient functions towards

a straight line.

The article by Dziak, Li, Tan, Shiffman, and Shiyko (2015) did not include any simulation of the performance of the method. Therefore, in this supplementary report we provide the results of a simulation study on the performance of MixTVEM.

1 Methods

Simulations were done in R (R Core Team, 2015), using the MixTVEM.r function provided by Dziak and colleagues (2015) for analysis. One thousand random datasets were generated, in each of three sizes:

- n =100 subjects with a mean of $\bar{m}$ =25 observations each
- n =100 subjects with a mean of $\bar{m}$ =75 observations each
- n =300 subjects with a mean of $\bar{m}$ =25 observations each

Within a given dataset, data was generated as follows. There are two equally sized latent classes, and each subject belongs randomly to one class or the other. The number of observations m_i for a given subject i is selected by generating a uniformly distributed random variable from 0 to $2\bar{m}$ and then rounding up. The measurement times $t_1, ..., t_{m_i}$ are assigned as m_i sorted random draws from a Uniform(0,1) distribution. A single time-varying covariate X_{ij} and a single time-varying outcome Y_{ij} are observed at each outcome. The covariate and

outcome are related as follows:

$$Y_{ij} = \beta_{0|c}(t) + \beta_{1|c}(t)X_{ij} + e_{ij} \tag{1}$$

Both the covariate X_{ij} and the error e_{ij} are generated as Gaussian random processes with mean zero and a modified autoregressive structure with nugget effect (see Dziak, Li, Tan, Shiffman, and Shiyko 2015) :

$$\begin{aligned}\mathrm{Cov}(X_{ij}, X_{ij'}) =& \sigma^2_{1X}\rho_X^{|t_{ij}-t_{ij'}|} + \sigma^2_{2X} \\ \mathrm{Cov}(e_{ij}, e_{ij'}) =& \sigma^2_{1e}\rho_e^{|t_{ij}-t_{ij'}|} + \sigma^2_{2e}.\end{aligned}$$

Here we set $\sigma^2_{1X} = \sigma^2_{2X} = \sigma^2_{1e} = \sigma^2_{2e} = \frac{1}{2}$, so that the total marginal variance of X or e is 1, and set $\rho_X = \rho_e = .5$. Lastly, for members of latent class 1, the coefficient functions in (1) were defined as $\beta_{0|1}(t) = 1 - e^{-t}$ and $\beta_{1|1}(t) = 2\sin(\pi t)$. For members of latent class 2, the functions were $\beta_{0|2}(t) = e^{-t}$ and $\beta_{1|2}(t) = 2\cos(\pi t)$. For each dataset, a MixTVEM model was fit using the R function described in Dziak et al (2015). A penalized cubic spline with 10 interior knots was specified for each function, and the tuning parameter required for tuning the model was selected using either an AIC-like (Akaike, 1973) or a BIC-like (Schwarz, 1978) criterion; the performance of these criteria was compared. Specifically, the tuning criteria were $AIC = -2\ell + 2\tilde{p}$ and $BIC = -2 * \ell + \log(n) * \tilde{p}$ where ℓ is the log-likelihood and $\tilde{p}$ is a measure of the effective number of parameters. It was assumed to be known that there were two latent classes, so the step of selecting the

number of latent classes was not simulated. Because the true model was relatively simple and in order to save time, only 5 sets of random starting values were used for fitting each model. Further details on the MixTVEM procedure are given in Dziak, Li, Tan, Shiffman, and Shiyko (2015), and full simulation code is available online.

Bias, mean squared error, and coverage for the estimated coefficient functions $\beta_{0|1}(t)$, $\beta_{1|1}(t)$, $\beta_{0|2}(t)$, and $\beta_{1|2}(t)$ were calculated for each method, averaged across subjects and across an evenly spaced grid of 100 time points. The coverage is for pointwise 95% confidence intervals, either with standard errors based on the robust or sandwich method (as recommended by Dziak et al., 2015) or with model-based standard errors. In each case, a correction for class membership uncertainty is included, as described by Dziak and colleagues (2015). Both kinds of standard errors are provided by the R function. In order to meaningfully define error or coverage, it was necessary to deal with label switching; that is, how to determine which of the two estimated classes should be considered to represent Class 1 and which one represents Class 2. However, because there were only two classes, this was not difficult; the class whose coefficients were the better match (in terms of sum squared error) to the true values for Class 1 was considered to be Class 1, and the class which better matched Class 2 was considered to be Class 2. This is a satisfactory solution because the numerical labels are arbitrary; the theoretical meanings of the classes come from their coefficient estimates.

Table 1: Mean Simulated Bias in Coefficient Functions, Averaged over a Grid of Time Points

Subjects	100	100	300
Obs. per Sub.	25	75	25
AIC Tuning			
$\beta_{0\|1}$ Bias	0.0082	-0.0215	-0.0274
$\beta_{1\|1}$ Bias	-0.0056	0.0009	0.0062
$\beta_{0\|2}$ Bias	-0.0268	-0.0141	0.0555
$\beta_{1\|2}$ Bias	-0.0028	0.0048	0.0062
BIC Tuning			
$\beta_{0\|1}$ Bias	0.0093	-0.0208	-0.0269
$\beta_{1\|1}$ Bias	-0.0068	0.0007	0.0060
$\beta_{0\|2}$ Bias	-0.0267	-0.0148	0.0552
$\beta_{1\|2}$ Bias	-0.0025	0.0048	0.0056

2 Results

Bias (averaged across time points) was generally quite small, as shown in Table 1. However, on some occasions underfitting may have occurred, biasing the estimated functions towards straight lines, and this would not necessarily be recorded as part of the average bias (because it would be upward bias at some time points and downward at others). There was not much noticeable difference between the results of AIC and BIC tuning.

MSE (averaged across time points) depended on sample size, as shown in Table 2. There was not much noticeable difference between the results of AIC and BIC tuning.

Coverage (nominal 95%) depended on the parameter and scenario, as shown in Table 3. Coverage was nominal or conservative for the intercept function, but somewhat below nominal for the time-varying

Table 2: Simulated Mean Squared Error (MSE) in Coefficient Functions, Averaged over a Grid of Time Points

Subjects	100	100	300
Obs. per Sub.	25	75	25
AIC Tuning			
$\beta_{0\|1}$ MSE	0.1523	0.0752	0.0614
$\beta_{1\|1}$ MSE	0.0419	0.0246	0.0335
$\beta_{0\|2}$ MSE	0.0868	0.0707	0.0855
$\beta_{1\|2}$ MSE	0.0481	0.0347	0.0375
BIC Tuning			
$\beta_{0\|1}$ MSE	0.1509	0.0729	0.0594
$\beta_{1\|1}$ MSE	0.0485	0.0247	0.0356
$\beta_{0\|2}$ MSE	0.0852	0.0694	0.0843
$\beta_{1\|2}$ MSE	0.0654	0.0460	0.0498

covariate. The sometimes low coverage for β_1 can be explained in terms of bias. The penalty function in the MixTVEM software is based on second order differences and tends to bias estimated parameter functions towards straight lines. Accordingly, one can explain much of the difference between datasets in mean pointwise coverage by departure from nonlinearity. To show this, for each dataset in the BIC-tuned, $n = 300$, $\bar{m} = 25$ scenario, a measure of nonlinearity was calculated for the estimated covariate function. This measure was defined as the sum squared errors, across grid points, of the best fit linear regression of the **estimated** function on time. Thus, if the estimated function was perfectly linear, the nonlinearity measure would be zero. If it was far from zero, the nonlinearity measure would be higher. For each of the two classes, Figure 1 shows a scatterplot of nonlinearity versus average pointwise coverage for the 1000 datasets

Table 3: Mean Simulated Coverage of Pointwise 95% Confidence Intervals for Coefficient Functions, Averaged over a Grid of Time Points

Subjects	100	100	300
Obs. per Sub.	25	75	25
AIC Tuning, Model-based Standard Errors			
$\beta_{0\|1}$ Coverage	0.7725	1.0000	0.9812
$\beta_{1\|1}$ Coverage	1.0000	0.9800	0.9325
$\beta_{0\|2}$ Coverage	0.9562	0.9800	0.8125
$\beta_{1\|2}$ Coverage	0.9550	0.9275	0.9025
BIC Tuning, Model-based Standard Errors			
$\beta_{0\|1}$ Coverage	0.7800	1.0000	0.9950
$\beta_{1\|1}$ Coverage	0.9588	0.9775	0.8912
$\beta_{0\|2}$ Coverage	0.9525	0.9838	0.8062
$\beta_{1\|2}$ Coverage	0.9125	0.7988	0.6900
AIC Tuning, Sandwich Standard Errors			
$\beta_{0\|1}$ Coverage	0.7887	1.0000	0.9775
$\beta_{1\|1}$ Coverage	0.9738	0.9450	0.8625
$\beta_{0\|2}$ Coverage	0.9538	0.9912	0.8462
$\beta_{1\|2}$ Coverage	0.9312	0.8338	0.8550
BIC Tuning, Sandwich Standard Errors			
$\beta_{0\|1}$ Coverage	0.7875	1.0000	0.9950
$\beta_{1\|1}$ Coverage	0.8825	0.9262	0.7950
$\beta_{0\|2}$ Coverage	0.9462	1.0000	0.8362
$\beta_{1\|2}$ Coverage	0.8250	0.6412	0.6050

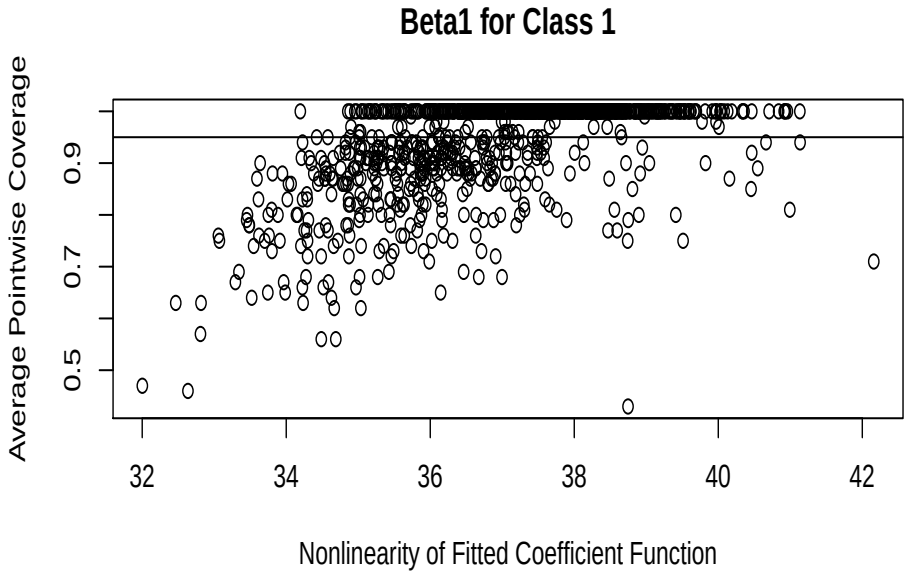
in the scenario. Figure 1 shows that in samples for which coverage is poorer than usual, it is usually the case that the estimated function is closer to a straight line than usual. In other words, the method tends to err on the side of parsimony.

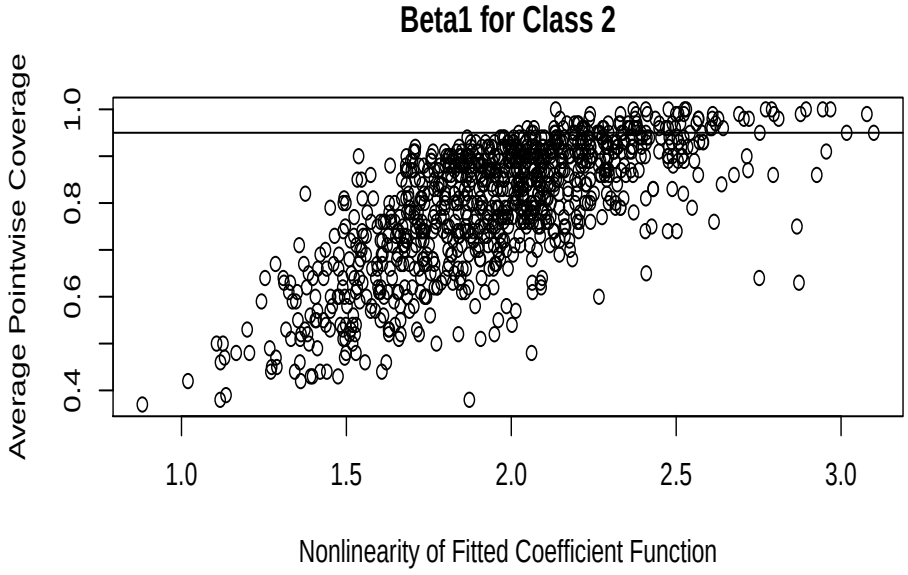
It appears that better coverage is obtained by model-based standard errors than by sandwich standard errors. This appears to contradict the recommendation by Dziak and colleagues (2015), who suggested using sandwich standard errors. However, it may be that this finding is incomplete because in this simulation the working covariance structure (autoregressive with nugget) was correct. It is possible that sandwich standard errors might provide greater robustness, especially for larger sample sizes (see Liang and Zeger, 1986).

3 Discussion

The performance of the methods was generally good, although confidence interval coverage was sometimes lower than desired. When lower coverage occurred, this was usually related to oversmoothing (underfitting). In many applications, particularly in the social sciences, parsimony and interpretability are highly valued, so it is better to oversmooth than to undersmooth. It is unfortunately a very widespread phenomenon that the necessity of choosing a model introduces additional uncertainty into parameter estimates which is not accounted for by usual methods for estimating standard errors (see, e.g., Buckland, Burnham and Augustin, 1997); choosing a tuning pa-

Figure 1: Average Pointwise Confidence Interval Coverage as Related to Nonlinearity of Estimated Function





rameter is really just a special case of that old problem. It is possible that bootstrapping might be used to create confidence intervals which adjusted for bias, but bootstrapping can be especially challenging in mixture models.

The current implementation in R is slow. The amount of time taken in the three conditions was

- 100 subjects, 25 observations: 69 hours for 1000 simulations with 5 random starts each, or about 4.1 minutes per simulation
- 100 subjects, 75 observations: 42 hours for 1000 simulations with 5 random starts each, or about 2.5 minutes per simulation
- 300 subjects, 25 observations: 156 hours for 1000 simulations, or about 9.4 minutes per simulation

This was with 5 random starts each. In a real data application it would be better to use at least 50 random starts, although 5 was enough for most datasets in these simulations.

Fortunately, the implementation could be made much faster; speed is probably not an inherent limitation of the method.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), Second international symposium on information theory (pp. 267?281). Budapest, Hungary: Akademai Kiado.

Buckland, Burnham, and Augustin (1997). Model selection: An integral part of inference. *Biometrics, 53,* 603-618.

Dziak, J. J., Li, R., Tan, X., Shiffman, S., & Shiyko, M. P. (2015). Modeling intensive longitudinal data with mixtures of nonparametric trajectories and time-varying effects. *Psychological Methods, 20*: 444-469.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*, 89–121.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B, 55*, 757–796.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22. http://dx.doi.org/10.1093/biomet/73.1.13

Lu, Z., & Song, X. (2012). Finite mixture varying coefficient models for analyzing longitudinal heterogenous data. *Statistics in Medicine, 31*, 544–560.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York, NY: Wiley.

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*, 882– 891.

Nagin, D. S. (1999). Analyzing developmental trajectories: A Semi-parametric, group-based approach. *Psychological Methods, 4*, 139–157.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Shiyko, M. P., Lanza, S. T., Tan, X., Li, R., & Shiffman, S. (2012). Using the time-varying effect model (TVEM) to examine dynamic associations between negative affect and self confidence on smoking urges: Differences between successful quitters and relapsers. *Prevention Science, 13*, 288–299.

Tan, X., Shiyko, M. P., Li, R., Li, Y., & Dierker, L. (2012). A time-varying effect model for intensive longitudinal data. *Psychological Methods, 17*, 61–77.