Tutorial Example for R MixTVEM

In order to use the MixTVEM function, we have to use the R `source` command in the usual way:

```
source("MixTvem.r");
```

This reads the code for the MixTVEM function into R.

Of course, we also need data. The tutorial example consists of two files. They are simulated (fake) data created to have similar appearance to the Shiffman et al (1996, 1997) data as analyzed by Dziak, Li, Tan, Shiffman, and Shiyko (2015), which described the experiences of smokers trying to quit, including their self-rated negative affect and urge to smoke. One of the data files, "MixTvemSampleObservationLevel.txt," has time-varying covariates for each subject listed in a longitudinal ("tall" or "stacked") format, and most of the subjects have more than one line of data.

```
ID    NegAffect   Time  Urge
1     1.4   0.9   3.79
1     1.57  3.72  6.07
1     1.27  4.76  3.35
...
200   3.47  2.97  10
```

When reading it in, make sure that it is in the R working directory, or else add the path to the data files to the file name below.

```
obsData <- read.table("MixTvemSampleObservationLevel.txt", sep="\t",
header=TRUE);
```

The other file, "MixTvemSampleSubjectLevel.txt," has subject-level, baseline data, arranged with one line per subject.

```
ID    MeanPrequitNegAffect    BaselineCigarettesPerDay
      MinutesToFirstCigarette RelapseAtOneMonth
1     1     28    19    0
2     1     23    13    0
3     1     32    9     0
4     2     30    14    1
```

John Dziak (jjd264@psu.edu)

We can read it in in the same way as the other file, although we won't need it immediately.

```
subData <- read.table("MixTvemSampleSubjectLevel.txt",sep="\t", header=TRUE);
```

One more piece of data preparation is needed: creating a column of ones to serve as the regressor for the intercept coefficient. This is easy:

```
obsData$Intercept <- 1;
```

We can now fit MiXTVEM models. Below is code for an intercept-only model with three classes and six interior knots. Because multiple random starts are requested, it may take a minute or so to run. However, using multiple random starts is worthwhile because it helps us judge how likely it is that we have found the global maximum likelihood. In practice, we recommend at least 20 and preferably 50 starts.

```
model1 <- TVEMMixNormal(dep=obsData$Urge,
                        id=obsData$ID,
                        numInteriorKnots=6,
                        numClasses=2,
                        numStarts=10,
                        tcov=obsData$Intercept,
                        time=obsData$Time);
```

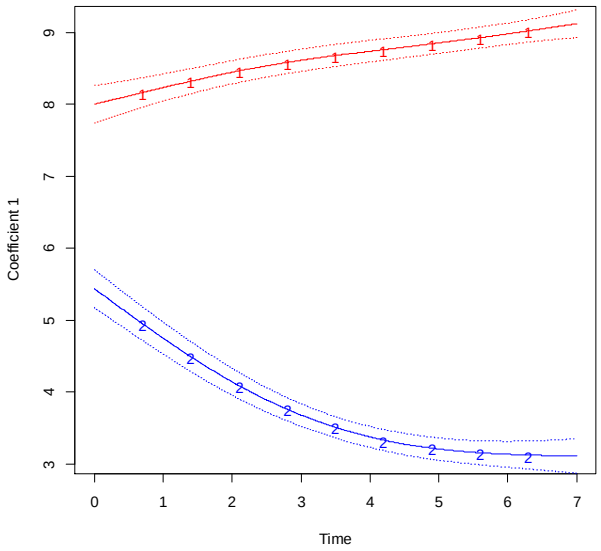
The first screen output is not very substantively interesting: the function lists which seeds and tuning parameter strengths ("lambdas") it is trying. This output is only useful for judging how quickly the estimation is progressing, and doesn't affect the interpretation of the final answer.

```
[1] "Trying seed= 96754566"
[1] "Trying seed= 65753533"
[1] "Trying seed= 66303049"
[1] "Trying seed= 44182993"
[1] "Trying seed= 14603612"
[1] "Trying seed= 90466104"
[1] "Trying seed= 39758284"
[1] "Trying seed= 50747970"
[1] "Trying seed= 7941648"
[1] "Trying seed= 25437984"
[1] "Trying lambda= 0.00307"
[1] "Trying lambda= 0.03067"
[1] "Trying lambda= 0.30666"
[1] "Trying lambda= 3.0666"
[1] "Trying lambda= 30.66599"
```

```
[1] "Trying lambda= 306.65988"
[1] "Trying lambda= 3066.59876"
[1] "Trying lambda= 7.70295"
[1] "Trying lambda= 15.3694"
[1] "Trying lambda= 30.66599"
[1] "Trying lambda= 61.18669"
[1] "Trying lambda= 122.0835"
[1] "Trying lambda= 77.02948"
[1] "Trying lambda= 96.97437"
[1] "Trying lambda= 122.0835"
[1] "Trying lambda= 153.69402"
[1] "Trying lambda= 193.4893"
```

The substantive output is as follows.

```
MixTVEM R Function
Number of subjects:                              200
Total number of observations:                   4963
Effect of time between knots treated as cubic
Roughness penalty weight:                      96.97
Weighted RSS statistic:                    15697.22
Weighted GCV statistic:                        3.18
Log-likelihood:                           -9746.22
AIC:                                      19515.04
BIC:                                      19552.30
Proportion of starting values giving approximately the best obtained ...
 ... log-likelihood:                            1.00
 ... weighted sum squared error:                   1.00
Estimated autocorrelation parameter rho:     0.434
Estimated proportion nugget:                 0.687
Count number of parameters:                  25.00
Smoothed number of parameters:               11.30
Class proportions:
0.5158 0.4842
Total standard deviations:
1.7841 1.7718
Logistic regression for class membership:
  Column Class    Estimate SEUncorrected SECorrected SESandwich          z
1     S1     2 -0.06304893     0.1412694   0.1342774  0.1299251 -0.4852715
         p
1 0.6274838
```



Class 1 is estimated to comprise about 52% of the (simulated) population and Class 2 accounts for the other 48%. They are estimated to have about the same standard deviation. The logistic regression predictor S1 is just the intercept in the intercept-only logistic regression model for class membership; its statistical nonsignificance in this context indicates that we don't have evidence against the null hypothesis that the class proportions are equal.

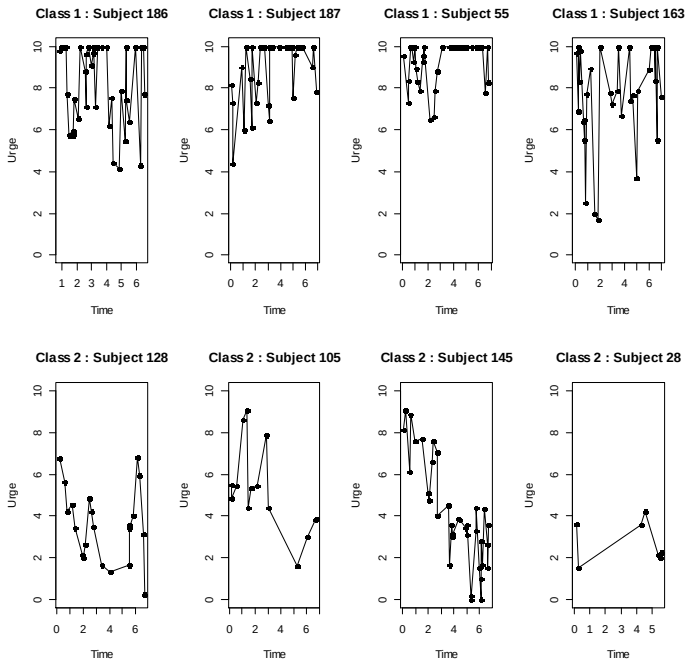
The decision to use six knots and two classes seems arbitrary. However, in practice you could use AIC or BIC, provided in the output, to try many possible choices and select the best.

Plots are automatically created. However, we can also create additional plots using the model information stored in the model1 results object. For instance, we can plot observed data from random subjects assigned (by highest posterior probability) to each class. This will help us diagnose how closely the observed data matches the smoothed averaged trajectories.

```
set.seed(42); # random seed;
class1 <- which(model1$bestFit$postProbsBySub[,1]>
                              model1$bestFit$postProbsBySub[,2]);
   # Gets the indexes (as integers from 1 through #subjects, in data order)
   # of all the subjects estimated to be more likely to belong to class 1
than
   # to class 2.
class1exemplars <- sample(class1,4);
class2 <- which(model1$bestFit$postProbsBySub[,2]>
                              model1$bestFit$postProbsBySub[,1]);
   # Gets the indexes (as integers from 1 through #subjects, in data order)
   # of all the subjects estimated to be more likely to belong to class 1
than
   # to class 2.
class2exemplars <- sample(class2,4);
par(mfrow=c(2,4));
for (i in class1exemplars) {
plot(model1$time[which(model1$intId==i)],
     model1$dep[which(model1$intId==i)],type="l",
     main=paste("Class",1,": Subject",i),
     ylim=c(0,10),xlab="Time",ylab="Urge");
points(model1$time[which(model1$intId==i)],
     model1$dep[which(model1$intId==i)],pch=16);
}
for (i in class2exemplars) {
plot(model1$time[which(model1$intId==i)],
     model1$dep[which(model1$intId==i)],type="l",
     main=paste("Class",2,": Subject",i),
     ylim=c(0,10),xlab="Time",ylab="Urge");
points(model1$time[which(model1$intId==i)],
     model1$dep[which(model1$intId==i)],pch=16);
```

```
}
```

Using the code above, we get the following plot.



The data were simulated to behave like the empirical data in the paper. That data was highly variable due to the many changing internal and external circumstances that determined each participant's urge to smoke at a given time. Thus, individual subjects do not always stay close to the average trajectory for their class. However, the urge ratings for class 1 typically stay high and those for class 2 typically fall, as the characteristics of the class as a whole would suggest.

We can add a covariate such as negative affect, in order to predict urge more precisely. We could let this covariate have time-varying or time-invariant effects, or we could fit both and compare the AIC or BIC values. Following the paper, we center negative affect at its grand mean.

The code for the model which assumes a constant regression coefficient for negative affectis as follows:

```
model1b <- TVEMMixNormal(dep=obsData$Urge,
                         id=obsData$ID,
                         numInteriorKnots=6,
                         numClasses=2,
                         numStarts=10,
                         tcov=obsData$Intercept,
                         xcov=obsData$CenteredNegAffect,
                         time=obsData$Time);
```

The resulting fit criteria are:

```
Weighted GCV statistic:                    2.66
Log-likelihood:                         -9257.30
AIC:                                    18541.55
BIC:                                    18586.02
```

The AIC and BIC are better (smaller) than the previous model, suggesting that the predictor is significant. However, we can continue further. The code for the model which assumes a time-varying regression coefficient is:

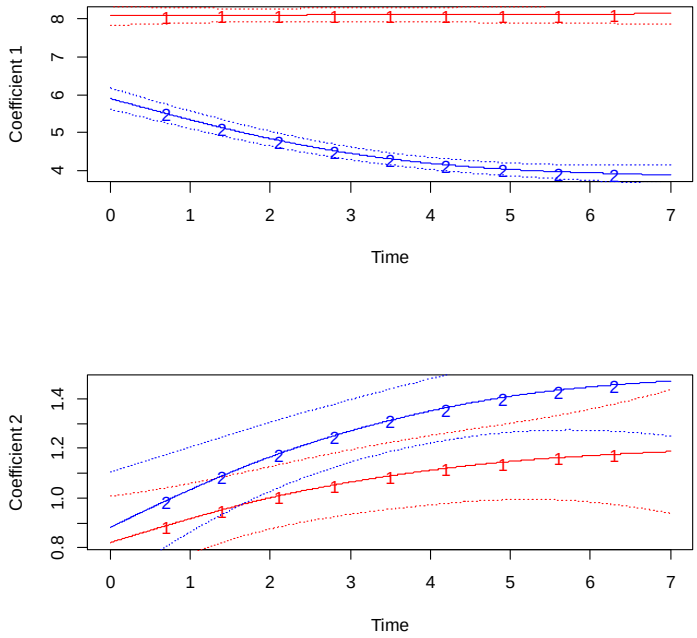
```
model2 <- TVEMMixNormal(dep=obsData$Urge,
                         id=obsData$ID,
                         numInteriorKnots=6,
                         numClasses=2,
                         numStarts=10,
                         tcov=cbind(obsData$Intercept,
                                    obsData$CenteredNegAffect),
                         time=obsData$Time);
```

The resulting fit statistics are:

```
Weighted GCV statistic:                    2.66
Log-likelihood:                         -9252.05
AIC:                                    18536.35
BIC:                                    18589.54
```

There is little evidence (in this model for this simulated dataset) of a better fit if negative affect is allowed to have a time-varying effect than if it is not. As often occurs, AIC favors the larger model and BIC favors the smaller one. However, because this is a demonstration, let's continue to examine the output from this time-varying-effect model. The full output for this model is:

```
MixTVEM R Function
Number of subjects:                           200
Total number of observations:                4963
Effect of time between knots treated as cubic
Roughness penalty weight:                  193.49
Weighted RSS statistic:                 13092.72
Weighted GCV statistic:                     2.66
Log-likelihood:                         -9252.05
AIC:                                    18536.35
BIC:                                    18589.54
Proportion of starting values giving approximately the best obtained ...
 ... log-likelihood:                         1.00
 ... weighted sum squared error:                1.00
Estimated autocorrelation parameter rho:    0.438
Estimated proportion nugget:                0.681
Count number of parameters:                 45.00
Smoothed number of parameters:              16.13
Class proportions:
0.5107 0.4893
Total standard deviations:
1.6285 1.6192
Logistic regression for class membership:
  Column Class   Estimate SEUncorrected SECorrected SESandwich          z
1     S1     2 -0.0428189     0.1402191   0.1392194  0.1374766 -0.3114632
         p
1 0.7554485
```



The main findings are that the class with a lower intercept function has a higher coefficient function for negative affect, and that for both classes there is a trend for a higher coefficient for negative affect as time goes on. (These simulated findings are based on the real findings discussed in the Dziak et al paper.)

We can also fit models to predict what kinds of people will be in each class. For example, we can use two baseline covariates – minutes to first cigarette in the morning, and average cigarettes smoked per day – to predict class membership. We can use the code

```
obsData <- merge(obsData,subData,by="ID");
model3 <- TVEMMixNormal(dep=obsData$Urge,
                        id=obsData$ID,
                        numInteriorKnots=6,
                        numClasses=2,
                        numStarts=10,
                        tcov=cbind(obsData$Intercept,
obsData$CenteredNegAffect),
                        scov=cbind(obsData$BaselineCigarettesPerDay,
obsData$MinutesToFirstCigarette),
```

```
                        time=obsData$Time);
```

The output, shown below, suggests a significant relationship of membership in class 2 with minutes to first cigarette. The plots are not shown because they were very similar to those generated without the subject-level covariates. (It is always important to check the plots, though, because adding covariates can sometimes change the meaning of classes in a mixture model.)

```
MixTVEM R Function
Number of subjects:                          200
Total number of observations:               4963
Effect of time between knots treated as cubic
Roughness penalty weight:                 193.49
Weighted RSS statistic:                13109.74
Weighted GCV statistic:                    2.66
Log-likelihood:                        -9178.97
AIC:                                   18394.18
BIC:                                   18453.97
Proportion of starting values giving approximately the best obtained ...
 ... log-likelihood:                       1.00
 ... weighted sum squared error:               1.00
Estimated autocorrelation parameter rho:    0.443
Estimated proportion nugget:                0.683
Count number of parameters:                 47.00
Smoothed number of parameters:              18.13
Class proportions:
0.5156 0.4844
Total standard deviations:
1.6345 1.6144
Logistic regression for class membership:
  Column Class    Estimate SEUncorrected SECorrected SESandwich         z
1     S1     2 -3.99891997    1.32322660  1.58361696 1.34030367 -2.983592
2     S2     2 -0.09266771    0.04905445  0.04892375 0.04571231 -2.027194
3     S3     2  0.38580884    0.05140577  0.02556141 0.02216858 17.403409
           p
1 0.00284886
2 0.04264259
3 0.00000000
```

In the output above, S1, S2, and S3 refer to the intercept and the subject-level covariates BaselineCigarettesPerDay and MinutesToFirstCigarette. The "class" column shows that membership in class 2, versus class 1 as a reference class, is being predicted. The estimated regression coefficient is shown along with three different standard errors. The first is not corrected for class membership uncertainty. The second is corrected for class membership uncertainty. The third uses a robust (sandwich) formula to also attempt to correct for uncertainty about

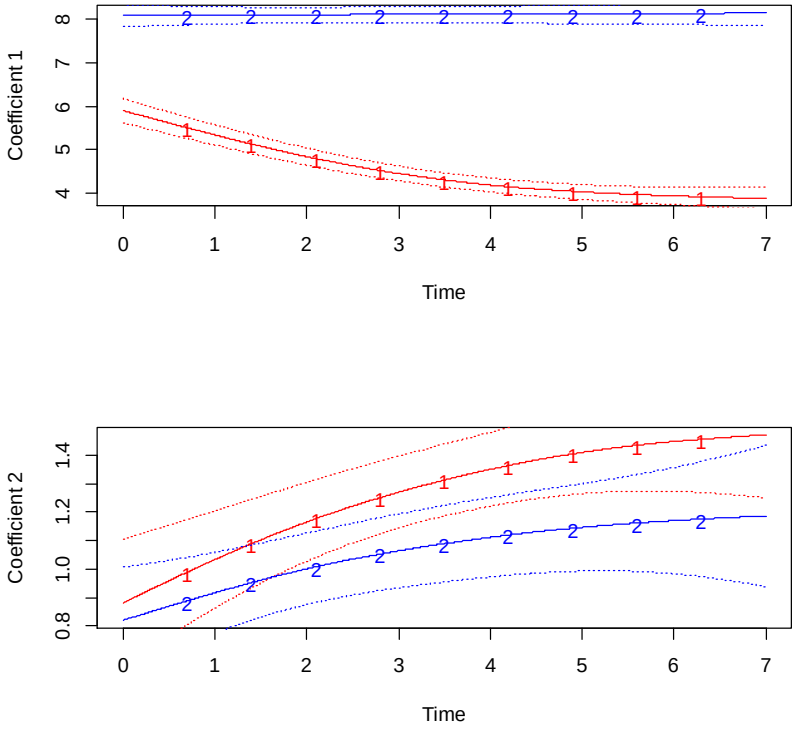
the structure of the within-subject covariance function. The third standard error is used to calculate the test statistic *z*, which in turn is compared to a normal distribution to test the null hypothesis that the corresponding coefficient is zero. In this simulated data, the covariate S3 (minutes to first cigarette) is very, very highly statistically significant. Someone who can wait longer until their first cigarette in a day is perhaps less addicted, and therefore more likely to be in the declining-urge class rather than the high-urge class. Specifically, 1 more minute of waiting is associated with an odds ratio of an estimated exp(0.38635389), or about 1.47, for being in class 2.

We could also explore whether the classes have different rates of relapse by including future relapse, measured at 1 month after quit date, as a covariate. We can use the following code:

```
model4 <- TVEMMixNormal(dep=obsData$Urge,
                        id=obsData$ID,
                        numInteriorKnots=6,
                        numClasses=2,
                        numStarts=10,
                        tcov=cbind(obsData$Intercept,
obsData$CenteredNegAffect),
                        scov=obsData$RelapseAtOneMonth,
                        time=obsData$Time);
```

We get these results:

```
MixTVEM R Function
Number of subjects:                          200
Total number of observations:               4963
Effect of time between knots treated as cubic
Roughness penalty weight:                 193.49
Weighted RSS statistic:                13091.01
Weighted GCV statistic:                    2.66
Log-likelihood:                        -9251.30
AIC:                                   18536.86
BIC:                                   18593.35
Proportion of starting values giving approximately the best obtained ...
 ... log-likelihood:                       1.00
 ... weighted sum squared error:              1.00
Estimated autocorrelation parameter rho:    0.438
Estimated proportion nugget:                0.680
Count number of parameters:                46.00
Smoothed number of parameters:             17.13
Class proportions:
0.4894 0.5106
Total standard deviations:
1.6194 1.6281
Logistic regression for class membership:
  Column Class    Estimate SEUncorrected SECorrected SESandwich          z
1     S1     2 -0.03033283     0.1516068   0.1536516  0.1530854 -0.1981432
2     S2     2  0.51062657     0.4091018   0.2741013  0.2019767  2.5281458
           p
1 0.84293300
2 0.01146667
```



The plots look a little different from before, because the class labels have switched. This does not have any substantive importance, because the numerical labels 1 and 2 on the classes are arbitrary and assigned randomly during the EM algorithm. Now class 2 is the one with the higher intercept function for urge.

The most important results here are in the logistic regression output. There appears to be a significant relationship between relapse (i.e., S2) and membership in the class 2 (high urge intercept) class (0.01146667). Thus it appears that even after adjusting for negative affect, participants with higher urge to smoke were more likely to relapse. Of course, this is not surprising. However, we get an estimated odds ratio here: relapsers are exp(0.51062657) or about 1.67 times as likely to have been in the high class.

References

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*, submitted.

Shiffman, S., Hickcox, M., Paty, J. A., Gnys, M., Kassel, J. D., & Richards, T. (1996). Progression from a smoking lapse to relapse: prediction from abstinence violation effects and nicotine dependence. *Journal of Consulting and Clinical Psychology, 64*, 993-1002.

Shiffman, S., Engberg, J., Paty, J. A., Perz, W., Gnys, M., & Kassel, J. D., & Hickcox, M. (1997). A day at a time: Predicting smoking lapse from daily urge. *Journal of Abnormal Psychology, 106*, 104-116.