

# **A Primer on Using SAS Mixed Models to Analyze Biorhythm Data**

Howard Seltman May, 1997

# Table of Contents

<b>Section</b>	<b>Page</b>
Purpose and Scope	3
Scope of Models Considered	4
Theory Review	5
Simulated Example	7
References	10

# 1 Purpose and Scope

This document explains the use of the Mixed Models procedure (PROC MIXED) available in SAS (SAS Institute Inc. Cary, NC, USA.) as applied to biorhythm data. It supplements both the documentation that comes with PROC MIXED and the book “SAS System for Mixed Models” by Littell, et al. by focusing specifically on those aspects of PROC MIXED that are of greatest interest to those researchers working with biorhythm data.

The type of data assumed in this document is multiple time series. Each time series represents the repeated measurements of some biological phenomenon (e.g. levels of a hormone in the blood or core body temperature) on a single subject. Normally there are dozens to hundreds of measurements per subject. Missing values are allowed. Usually the measurements are equally spaced. All of the subjects may come from one group, or there may be, e.g. control and treatment groups. There may be additional levels of hierarchy of the subjects, e.g. there may be control and treatment groups at each of several centers.

The analyses described here are based on fitting a circadian (or other time period) rhythm with a series of sin and cos curves. This approach may be described as harmonic regression, cosinor analysis, or regression on a Fourier basis (Radomski, et al., 1995, Iranmanesh, et al. 1990, Bergendahl, 1996). Theoretically any curve may be approximated as closely as desired by using a sufficient number of harmonics. Practically, an appropriate fundamental frequency and a few harmonics may provide a good approximation to many, but not all biorhythms.

Two key extensions to the usual cosinor analysis are the inclusion of autoregressive error structures to model the strong serial correlation often seen in this type of data (Greenhouse, et al. 1987), and the use of mixed (hierarchical) models to account for the normal subject-to-subject biological variation (Laird and Ware, 1982, Greenhouse, et al. 1993).

The types of data that cannot be analysed using the methods described here include those with unknown fundamental periods (Greenhouse, et al. 1987) or with periods that vary over the time of the study.

## 2 Scope of Models Considered

Here is an overview of how one might think about a typical biorhythm study. The overall shape of the time series for any individual subject is constructed from a series of sin and cos curves as described below. You must know the fundamental frequency (e.g. 1/24 hour for a circadian rhythm), but the number of harmonics can be decided on as part of the model selection procedure. Technically, sin and cos terms are columns in the design matrix for the fixed (non-random) part of the model. (They may also be in the random effects design matrix.) The coefficients of these terms relate to the phase and amplitude of the various frequency components of the curve estimate.

Additional components are often present in the fixed effects design matrix. These are used to account for differences in the level or shape of the curve due to covariates (e.g. male vs. female) or treatment groups. Effects that are represented simply as additional variables in the fixed effects allow for changes in the level of the curve for different levels of the variable. Effects that are represented as interaction terms between additional variables and the sin/cos terms introduce shape changes.

The next component of the model is the serial correlation of the samples. Since samples are usually closely spaced, treating them as independently distributed is usually inappropriate. SAS PROC MIXED can accommodate AR1 (first order autoregressive) or ARMA (AR1 plus first order moving average) error structures; these are likely to provide acceptable modeling of the serial correlation.

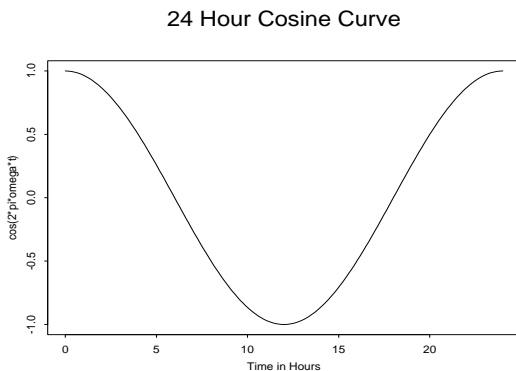
Next we consider the hierarchical portion of the model. This relates to the random effects. One form of subject-to-subject variation is a random intercept, in which each subject has an individual level for his or her biorhythm curve. If the model has fixed effects covariates, then the random intercept reflects random variation of the individual curve level among all subjects having the same covariates. The inclusion of sin and cos terms in the random effects results in random slopes, i.e. random variation of the shapes of the curves (e.g. time of peak) around the average for subjects with a specific set of fixed effect covariates.

### 3 Theory Review

Good references to the theory of mixed models include Laird and Ware, 1982, Diggle, et al., 1994, and Littell, et al. 1996. This paper uses the SAS notation.

The overall shape fitted to an individual patient's data in harmonic regression is based on sine (sin) and cosine (cos) curves. According to the theory of Fourier analysis, *any* curve can be represented as the sum of many sin and cos curves. In this sense it is theoretically possible to get a good approximation to any curve using the approach given here. Practically, with a reasonable size data set, only a relatively small number of sin and cos curves will be used. If your biorhythms cannot be approximated by a relatively small number of sin and cos curves, harmonic regression is probably not appropriate for your problem.

Harmonic regression uses a pair (one sin plus one cos) of curves for each frequency (usually labelled  $\omega$ ). The frequency is the reciprocal of the period; periods are easier for people to think about, while frequencies are more natural to use in many formulas. For example if the sample times are in hours, and the rhythm is circadian, then the period is 24 and the frequency is  $1/24$  (0.04167). This is the fundamental (lowest frequency), *which is assumed to be known in this document*. Higher frequencies are called harmonics. In the example considered here, the next several frequencies are  $2/24=1/12$ ,  $3/24=1/8$ ,  $4/24=1/6$ , and  $5/24$ ; these correspond to periods of 12, 8, 6, and 4.8 hours respectively. A sin wave can be represented by the equation  $\cos(2\pi\omega t)$  where  $t$  is the time and  $\pi$  is 3.14159. If  $t$  goes from 0 to  $1/\omega$ , the curve traces one complete cycle:



A cos wave of a given frequency can be shifted to the right by an amount called the phase

and labelled  $\phi$  by using the equation  $\cos(2\pi(\omega t - \phi))$ . In this form, the peak of the cos curve is shifted from its standard time 0 position by  $2\pi\phi$  where positive values are a shift to the right, and negative values to the left. *For the rest of this document, the word phase will refer to  $2\pi\phi$  so that the “phase” can be thought of in the same time units as the sampling in the experiment.*

A more complete representation of a time series, “Y”, indexed by time “t”, when offset by mean,  $\mu$ , multiplied by amplitude, R, and perturbed by errors,  $\epsilon_t$ , is given here:

$$Y_t = \mu + \sum_{k=1}^K R_k \cos(2\pi(k\omega t - \phi_k)) + \epsilon_t$$

Often a form linear in the unknown parameters is used:

$$Y_t = \mu + \sum_{k=1}^K [A_k \cos(2\pi k\omega t) + B_k \sin(2\pi k\omega t)] + \epsilon_t$$

In this form, a each phase shifted cos wave is represented as the sum of two unshifted cos and sin wave with appropriate separate amplitudes,  $A_k$  and  $B_k$ . The equations that can be used to convert between these forms are:

$$R = \sqrt{A^2 + B^2} \quad 2\pi\omega\phi = \arctan(B/A)$$

$$A = \frac{R}{\sqrt{1 + \tan(2\pi\omega\phi)^2}} \quad B = \sqrt{R^2 - A^2}$$

## 4 Simulated Example

A detailed example of the analysis of simulated biorhythm data is given here. The simulated data, as well as the Splus (StatSci, Inc, WA) program used to generate in are available on the world wide web at <http://www.stat.cmu.edu/hselman/SASMixed/primer.html>. The data consists of measurements of some unspecified attribute that follows a circadian rhythm. Measurements of the attribute in 10 control and 10 treated subjects are observed every 15 minutes for 24 hours.

The data are simulated as the sum of a 24 period fundamental cosine curve plus an 8 hour period harmonic. All subjects have a harmonic amplitude of 7, a harmonic phase shift of -1 hours, a fundamental amplitude of 5 with a random subject-to-subject variation with standard deviation (sd) equal to 1. The controls have mean 45 with a random variation of sd=2, and a fundamental phase shift of 4 with a random variation of sd=2. The treated subjects have mean of 50 with a random variation of sd=5(sd), and a fundamental phase shift of -2 with a random variation of sd=2.

The analysis steps given here are neither universal nor unique, but some variation of these steps should comprise a reasonable analysis in many situations.

### 4.1 Reading in the data

The following header and data sections are common to all SAS programs in this example. The data is read in from a file, and new variables are created for the autoregression times (needed to determine the order and spacing of the samples in the AR analysis), and the sin and cos variables:

```
options linesize=80;
data Sim1;
infile 'Sim.dat';
input id rx time attrib;
      cos24=cos(2*3.14159/24*time);
```

```

cos12=cos(2*3.14159/12*time);
cos8=cos(2*3.14159/8*time);
cos6=cos(2*3.14159/6*time);
cos4p8=cos(2*3.14159/4.8*time);
sin24=sin(2*3.14159/24*time);
sin12=sin(2*3.14159/12*time);
sin8=sin(2*3.14159/8*time);
sin6=sin(2*3.14159/6*time);
sin4p8=sin(2*3.14159/4.8*time);
artime=int(time*100); /* remove decimals to get correct sort order */
run;

```

The input statement assigns names to the four columns in the data file. The `cosxx` and `sinxx` terms define the cosine and sine curves for each frequency used. The `artime` converts times from decimal hours to integer hundredths of an hour; the creates numbers that will be correctly sorted by SAS when interpreted as a “class” variable, as is needed for the “repeated” statement.

## 4.2 Test of AR parameter

The first step is to use a set of fixed effects that should be conservatively sufficient to model the curves in any given subject. By including more harmonic frequencies that one would expect to need, testing of the serial correlation (AR parameter) and the other random effects will be little effected my misspecification of the fixed effects.

The SAS program used is:

```

title 'Sim Step 1: Fundamental + 4 Harmonics + Treatment +/-AR(1)';

title2 'Without AR(1)';
proc mixed data=Sim1 info;
  class rx;

```



```

model attrib = rx cos24 sin24 cos12 sin12 cos8 sin8 cos6 sin6 cos4p8 sin4p8
  rx*cos24 rx*sin24 rx*cos12 rx*sin12 rx*cos8 rx*sin8
  rx*cos6 rx*sin6 rx*cos4p8 rx*sin4p8;
random int /type=VC sub=id;
run;

```

```

title2 'With AR(1)';
proc mixed data=Sim1 info;
  class rx artime;
  model attrib = rx cos24 sin24 cos12 sin12 cos8 sin8 cos6 sin6 cos4p8 sin4p8
    rx*cos24 rx*sin24 rx*cos12 rx*sin12 rx*cos8 rx*sin8
    rx*cos6 rx*sin6 rx*cos4p8 rx*sin4p8;
  random int /type=VC sub=id;
  repeated artime / type=ar(1) sub=id;
run;

```

## 5 References

Bergendahl, M, Vance, ML, Iranmanesh, A, Thorner, MO and Veldhuis, JD, Fasting as a metabolic stress paradigm selectively amplifies cortisol secretory burst mass and delays the time of maximal nyctohemeral cortisol concentrations in healthy men. *Journal of Clinical Endocrinology and Metabolism*, 81: 692-9, 1996.

Chi, EM and Reinsel, GC, Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, 84: 452-459.

Diggle, PJ, Liang, K-Y, and Zeger, SL, *Analysis of longitudinal data*. Oxford University Press, Oxford, UK, 1994.

Greenhouse, JB, Kass, RE, and Tsay, RS, Fitting nonlinear models with ARMA errors to biological rhythm data. *Statistics in Medicine*, 6: 167-183, 1987.

Greenhouse, JB, Kass, RE, Lam, T and Tsay, RS, A hierarchical model for serially correlated data: Analysis of biological rhythm data. Technical Report, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 1993. (Available at <http://lib.stat.cmu.edu>.)

Iranmanesh, A, Lizarralde, G, Johnson, ML, and Veldhuis, JD, Dynamics of 24-hour endogenous cortisol secretion and clearance in primary hypothyroidism assessed before and after partial thyroid hormone replacement. *Journal of Clinical Endocrinology and Metabolism* 70: 155-61, 1990.

Laird, NM, and Ware, JH, Random effects models for longitudinal data. *Biometrics*, 38, 963-974, 1982.

Littell, RC, Milliken, GA, Stroup, W. and Wolfinger, RD. *SAS System for Mixed Models*, SAS Institute, Inc., Cary NC, 1996.

Pauler, D, The Schwarz criterion for mixed effects models. Dissertation for Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 1996.

Radomski, MW, Buguet, A., Montmayeur, A., Bogui, P., Bourdon, L., Doua, F., Lonsdorfer, A., Tapie, P. and Dumas, M., Twenty-four-hour plasma cortisol and prolactin in human African trypanosomiasis patients and healthy African controls. *American Journal of Tropical Medicine and Hygiene*, 52: 281-6, 1995.

Schwarz, G, Estimating the dimension of a model. *Annals of Statistics*, 6:461-464, 1978.