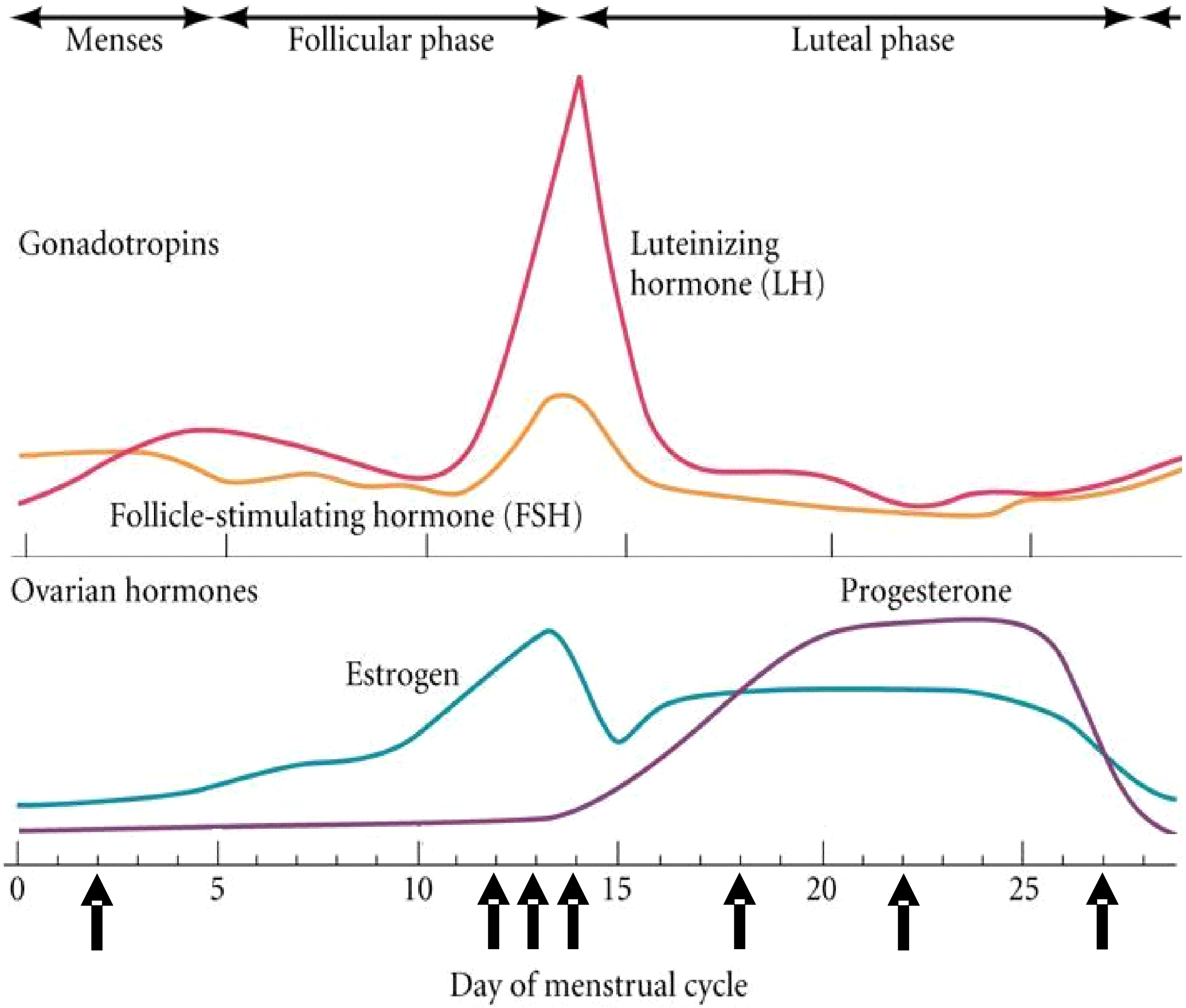
# Approaches to Modeling Menstrual Cycle Function

Paul S. Albert (albertp@mail.nih.gov)
Biostatistics & Bioinformatics Branch
Division of Epidemiology, Statistics, and Prevention Research
NICHD

SPER Student Workshop
June 22, 2010

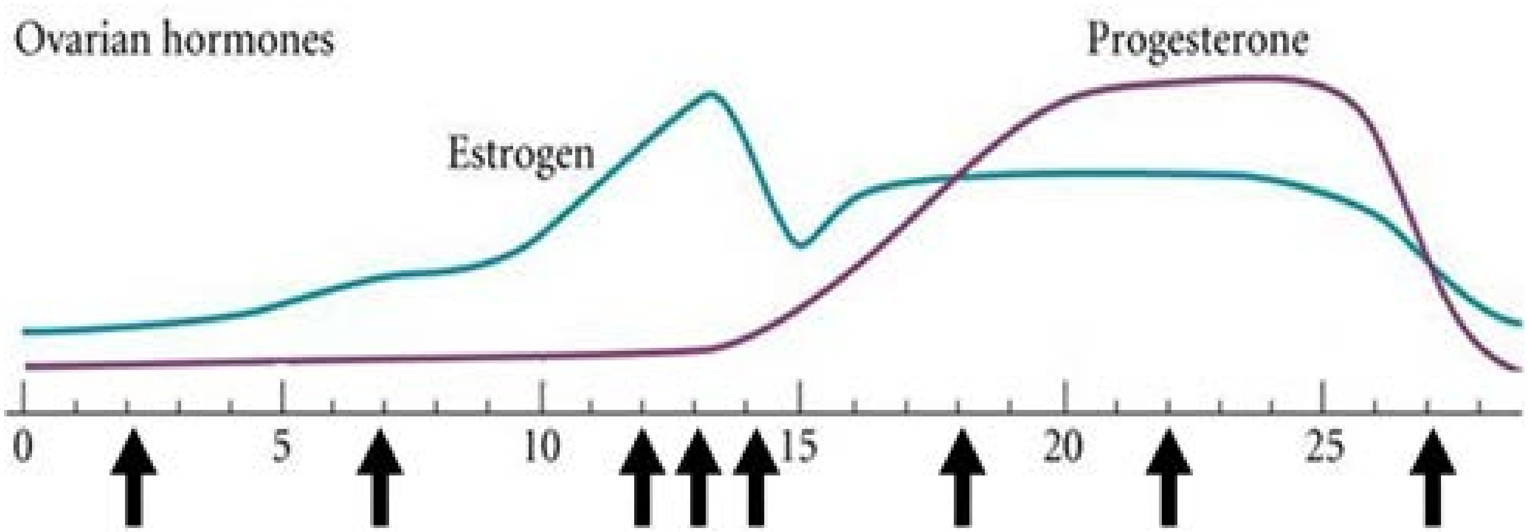
# Menstrual Cycle Function



# Scientific Interest

- To characterize the “typical” menstrual cycle pattern in a population of women.
  - What is the effect of a subject-specific covariate  on typical menstrual cycle?
- To characterize the variation in menstrual cycle function across women and across consecutive cycles on the same woman.
  - What is the inter-relationship between multiple hormones across the menstrual cycle?
- To characterize the relationship between menstrual cycle function and menstrual cycle length.

# Longitudinal Menstrual Cycle Pattern



Menstrual

Early Follicular

Ovulation

Early Luteal

Mid Luteal

Late Luteal

Urine measurements taken daily
-or-
Serum measurements taken at regularly scheduled time points.

# Time Scale for Menstrual Cycle Function

1. Assign visit days to phases in the cycle
2. Actual visit days
3. Actual visit days standardized by cycle length
4. Registered cycles:
   - Standardize by day of ovulation for time period before ovulation.
   - Standardize by time between ovulation and menstruation for time period between ovulation and menstruation.

# Linear Mixed Model

## Pinheiro and Bates-Springer, 2000

- $Y_{ij} = X_{ij}\beta + Z_{ij}b_i + \varepsilon_{ij}$

  Fixed Effects Random Effects

- For follow-up data at fixed scheduled time points:
  - $Y_{ij} = \beta_0 + \beta_1 X_i + \sum_{d=2}^{8} I_{(j=d)}\beta_d + b_i + \varepsilon_{ij}$
  - $Y_{ij} = \beta_0 + \beta_1 X_i + \sum_{d=2}^{8} \beta_d I_{(j=d)} + \sum_{d=2}^{8} \eta_d I_{(j=d)} X_i + b_i + \varepsilon_{ij}$

# Linear Mixed Model

## Pinheiro and Bates-Springer, 2000

$$Y_{ij} = \beta_0 + \beta_1 X_i + \sum_{d=2}^{8} \beta_d I_{(j=d)} + \sum_{d=2}^{8} \eta_d I_{(j=d)} X_i + b_i + \varepsilon_{ij}$$

- Test of whether a covariate influences the menstrual cycle pattern of a particular hormone can be formulated as a likelihood ratio test of $\eta_2 = \eta_3 = \ldots = \eta_8$.

# Linear Mixed Model

## Pinheiro and Bates-Springer, 2000

- Advantages:
  - Simple to implement in standard software
  - Easily allows for flexible correlation structure (example: AR-1, ARIMA, exponential model, spherical model)
  - Model diagnostics “built in” in standard software such as lme.

# Linear Mixed Model

## Pinheiro and Bates-Springer, 2000

- Advantages (Continued):
  - Can be easily extended to multiple cycles per women

$$Y_{ij} = \beta_0 + \beta_1 X_i + \sum_{d=2}^{8} \beta_d I_{(j=d)} + \sum_{d=2}^{8} \eta_d I_{(j=d)} X_i + b_i + b_{ik} + \varepsilon_{ij}$$

- Disadvantages:
  - Requires uniform measurement times on each subject.
  - Global tests are not very powerful.
  - Difficult to study effect of covariates on hormonal pattern.

# Semi-parametric Stochastic Mixed Model

## Zhang, Lin, Raz, and Sowers - JASA 1998

$$Y_{ij} = X'_{ij}\beta + f(t_{ij}) + Z'_{ij}b_i + U_i(t_{ij}) + \varepsilon_{ij}$$

Fixed Effects

Smoothed function of time

Random Effects

Random process

- Random process follows a nonhomogenous Ornstein-Uhlenbeck (NOU) process

# Semi-parametric Stochastic Mixed Model

## Zhang, Lin, Raz, and Sowers - JASA 1998

- NOU process:

$$Var(U_i(t)) = \zeta(t)$$

with $\log\{\zeta(t)\} = \zeta_0 + \zeta_1 t$

$$Corr(U_i(t), U_i(s)) = \rho^{|t-s|}$$

# Semi-parametric Stochastic Mixed Model

## Zhang, Lin, Raz, and Sowers - JASA 1998

- Advantages:
  - Allows for irregularly spaced measurements.
  - Accounts for serial correlation and variances of measurements that change over the menstrual cycle.
  - Allows for very flexible hormonal patterns.
- Disadvantages:
  - Covariates affect cycle only by a constant shift in mean.
  - Needs specialized software to implement.
  - Does not easily extend to multiple cycles per women.

# Semi-parametric Stochastic Mixed Model

## Zhang, Lin, and Sowers - Biometrics 2000

- Comparing the non-parametric curves across two groups

$$Y_{kij} = X'_{kij}\beta + f(t_{kij}) + Z'_{ij} b_{ki} + U_{ki}(t_{kij}) + \varepsilon_{kij}$$

Groups K=1 or 2

Are the two curves different across groups?

$$\Delta\{f_1(.), f_2(.)\} = \int_0^T [f_1(t) - f_2(t)]^2 dt$$

- Chi-Squared test to compare the two curves

# Semi-parametric Stochastic Mixed Model

## Zhang, Lin, and Sowers - Biometrics 2000

- Advantages:
  - Very flexible and general test
- Disadvantages:
  - Only compares two discrete groups
  - Difficult to implement with standard software

# Characterizing Hormonal Profiles Using Functional Data Analysis

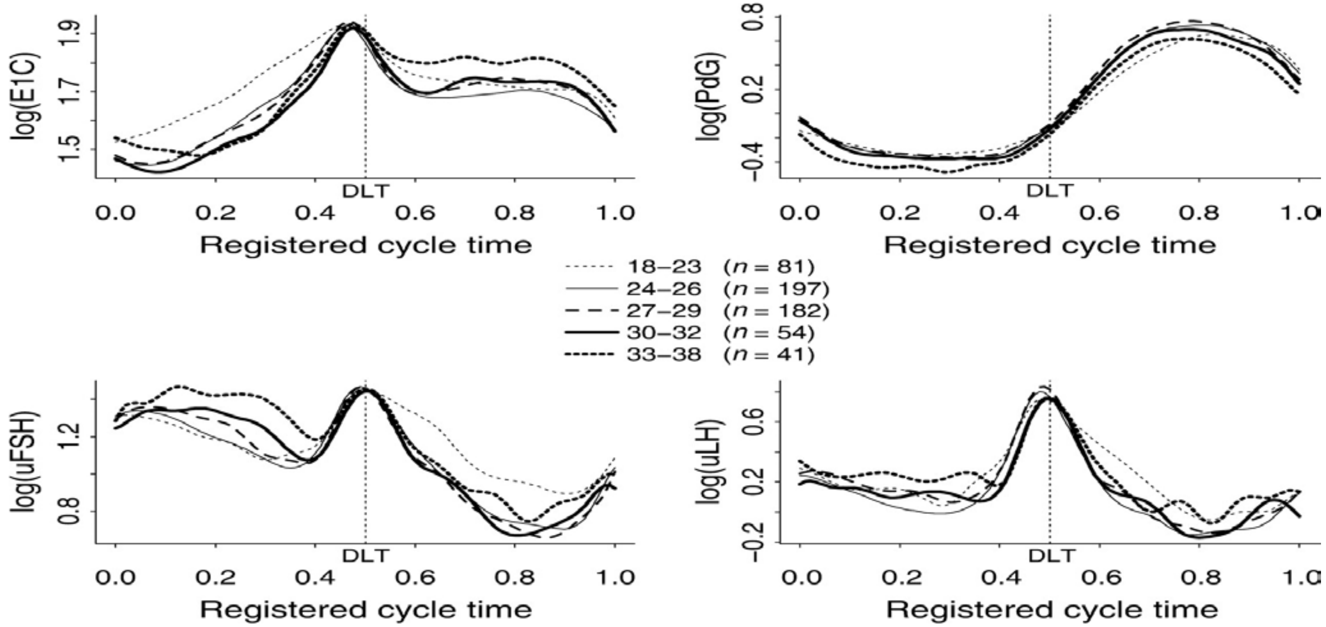
## Meyer, Zeger, Harlow, et al. - AJE 2007

- Fit separate curves to each subject’s menstrual cycle data.
- Cubic B-splines with 20 basis functions were used with associated weights estimated from the data.
- Mean curves for each group were estimated by averaging estimated weights and multiplying average weights by each basis function.

# Characterizing Hormonal Profiles Using Functional Data Analysis

## Meyer, Zeger, Harlow, et al. - AJE 2007

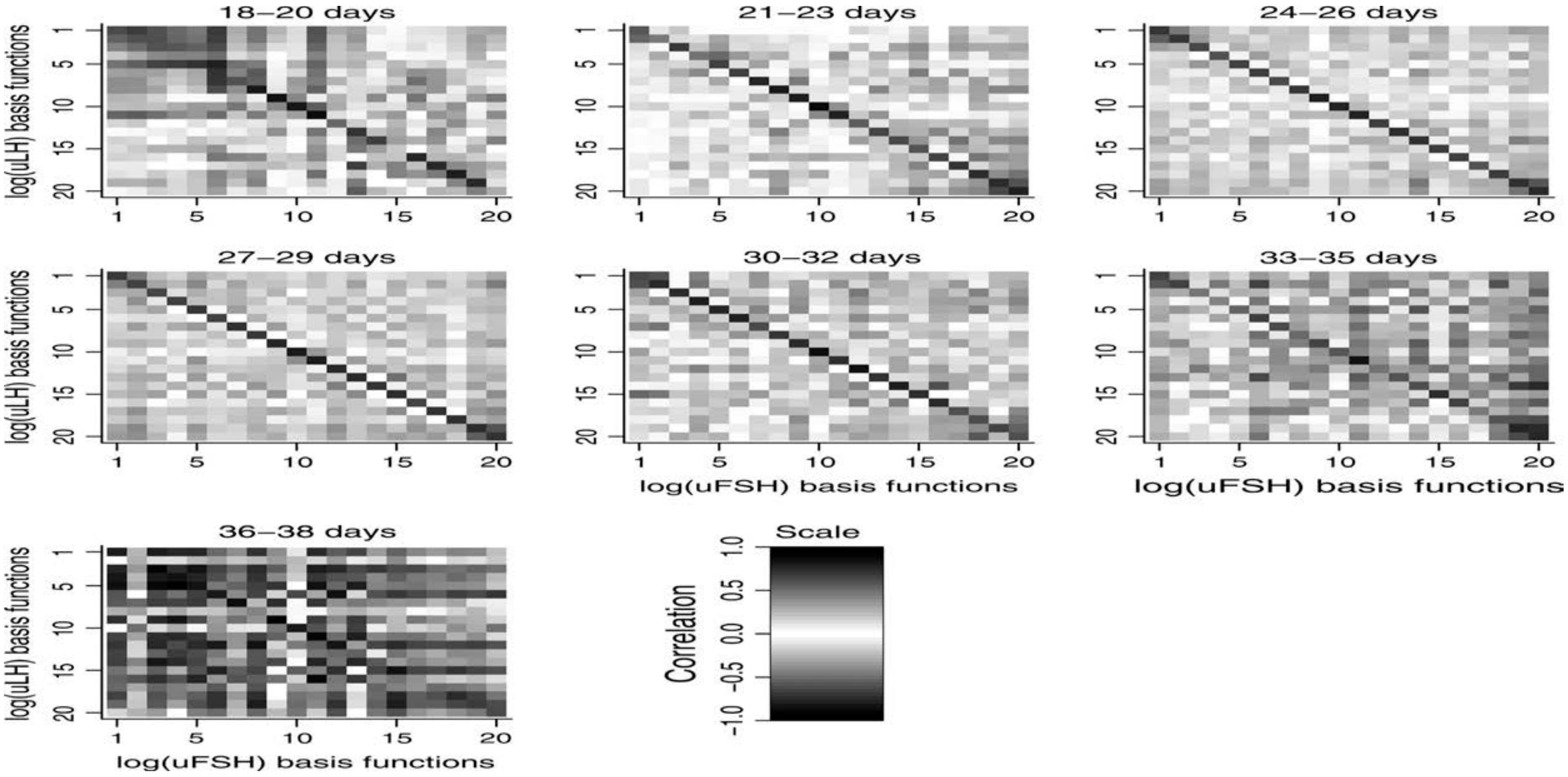
- Interest is on comparing menstrual cycle function by cycle length



# Characterizing Hormonal Profiles Using Functional Data Analysis

## Meyer, Zeger, Harlow, et al. - AJE 2007

- Evaluating synchrony between FSH and LH:



# Characterizing Hormonal Profiles Using Functional Data Analysis

## Meyer, Zeger, Harlow, et al. – AJE 2007

- Advantages:
  - Summarizes menstrual cycle function in a sensible way.
  - Easily allows for examining how different hormones relate to each other.
- Disadvantages:
  - Needs many measurements on each cycle. Good for urine. May not work for serum measurements.
  - Does not easily adapt to a regression framework.
  - Focus is not on statistical testing, but is more descriptive.

# Nonlinear Mixed Models: Shape Invariant Model

Wang, Ye, and Brown - Biometrics 2003

Albert and Hunsberger - Biometrics 2005

- $y_{ij} = \phi_{1i} + \exp(\phi_{2i}) f\{t_{ij} - a\log it(\phi_{3i})\} + \varepsilon_{ij}$

Mean level

Amplitude: nadir to peak distance

Phase shift

$$a\log it(x) = \exp(x)/\{1+\exp(x)\}$$

$$\phi_{1i} = X_{1i}\beta + Z_{1i}b_{1i}$$

$$\phi_{2i} = X_{2i}\beta + Z_{2i}b_{2i}$$

$$\phi_{3i} = X_{3i}\beta + Z_{3i}b_{3i}$$

- f(t) is a periodic function with t ranging from 0 to 1

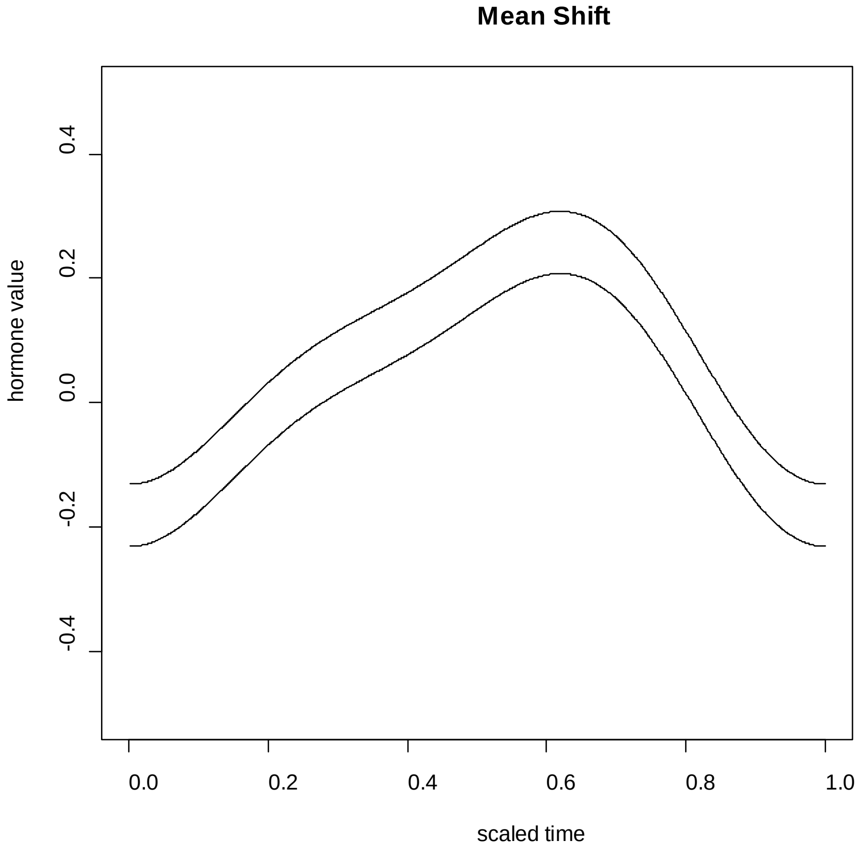
# Nonlinear Mixed Models: Shape Invariant Model

## Wang, Ye, and Brown- Biometrics 2003

## Albert and Hunsberger - Biometrics 2005

- $y_{ij} = \phi_{1i} + \exp(\phi_{2i})\, f\{t_{ij} - a \log it(\phi_{3i})\} + \varepsilon_{ij}$

Mean level



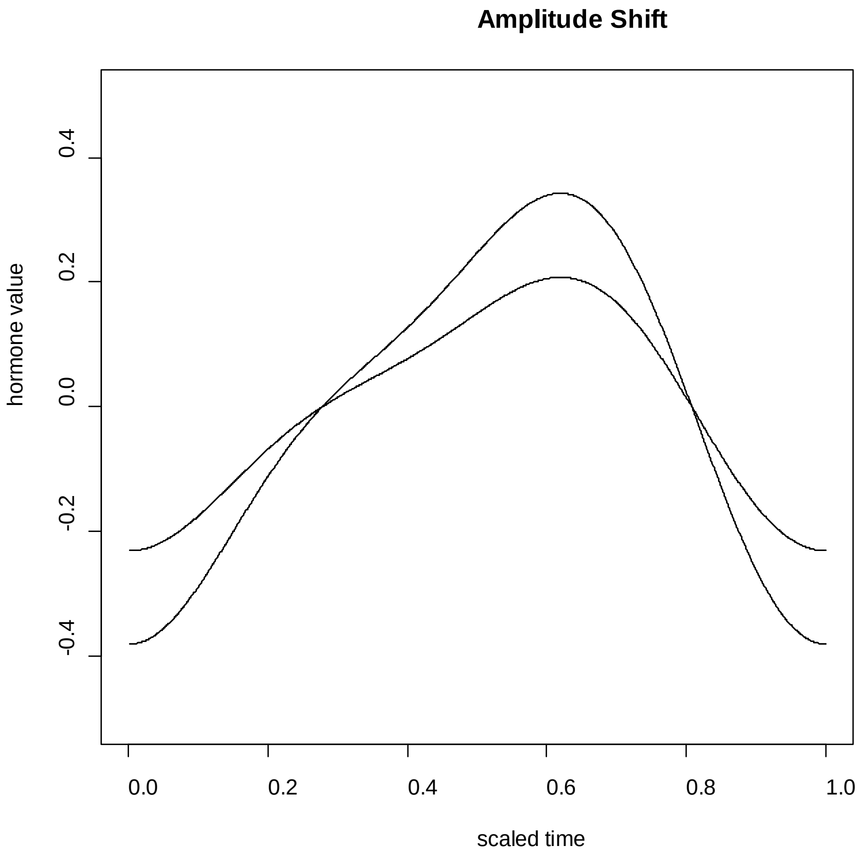
# Nonlinear Mixed Models: Shape Invariant Model

Wang, Ye, and Brown- Biometrics 2003

Albert and Hunsberger - Biometrics 2005

- $y_{ij} = \phi_{1i} + \exp(\phi_{2i}) f\{t_{ij} - a \log it(\phi_{3i})\} + \varepsilon_{ij}$

Amplitude: nadir to peak distance



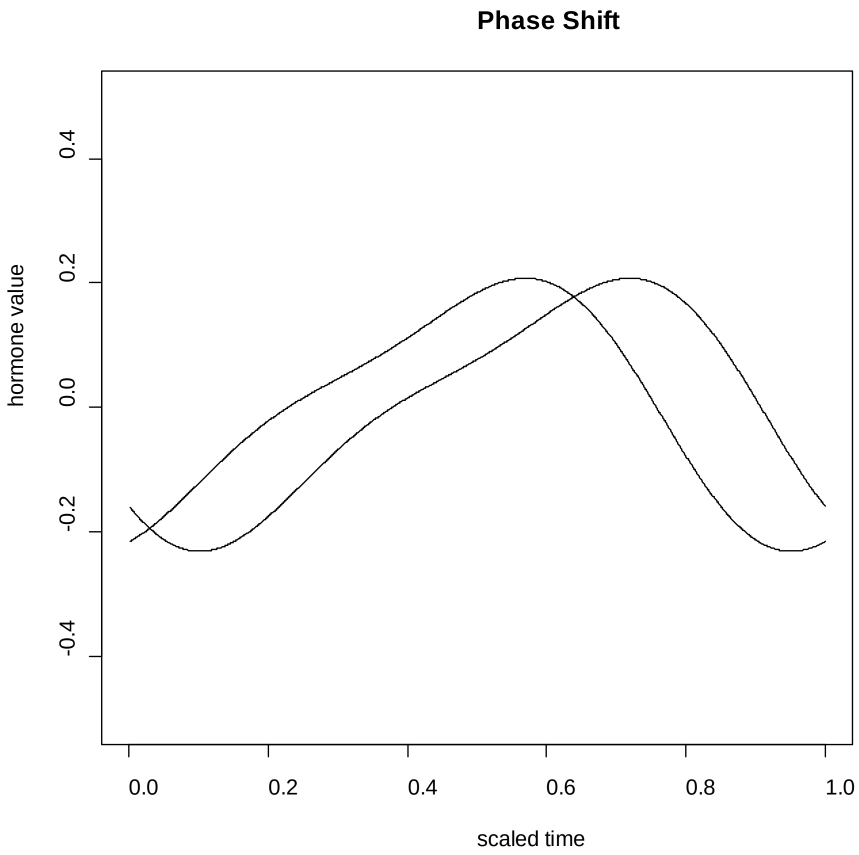
# Nonlinear Mixed Models: Shape Invariant Model

Wang, Ye, and Brown- Biometrics 2003

Albert and Hunsberger - Biometrics 2005

- $y_{ij} = \phi_{1i} + \exp(\phi_{2i}) f\{t_{ij} - a \log it(\phi_{3i})\} + \varepsilon_{ij}$

Phase shift


Phase Shift

# Nonlinear Mixed Models: Shape Invariant Model

Wang, Ye, and Brown- Biometrics 2003

Albert and Hunsberger - Biometrics 2005

$$y_{ij} = \phi_{1i} + \exp(\phi_{2i}) f\{t_{ij} - a \log it(\phi_{3i})\} + \varepsilon_{ij}$$

- f is a periodic function which can be represented in a number of ways:
  - – Periodic splines (Wang, Ke, and Brown 2003)
  - – Harmonic models (Albert and Hunsberger 2005)

# Nonlinear Mixed Models: Shape Invariant Model

## Characterizing f

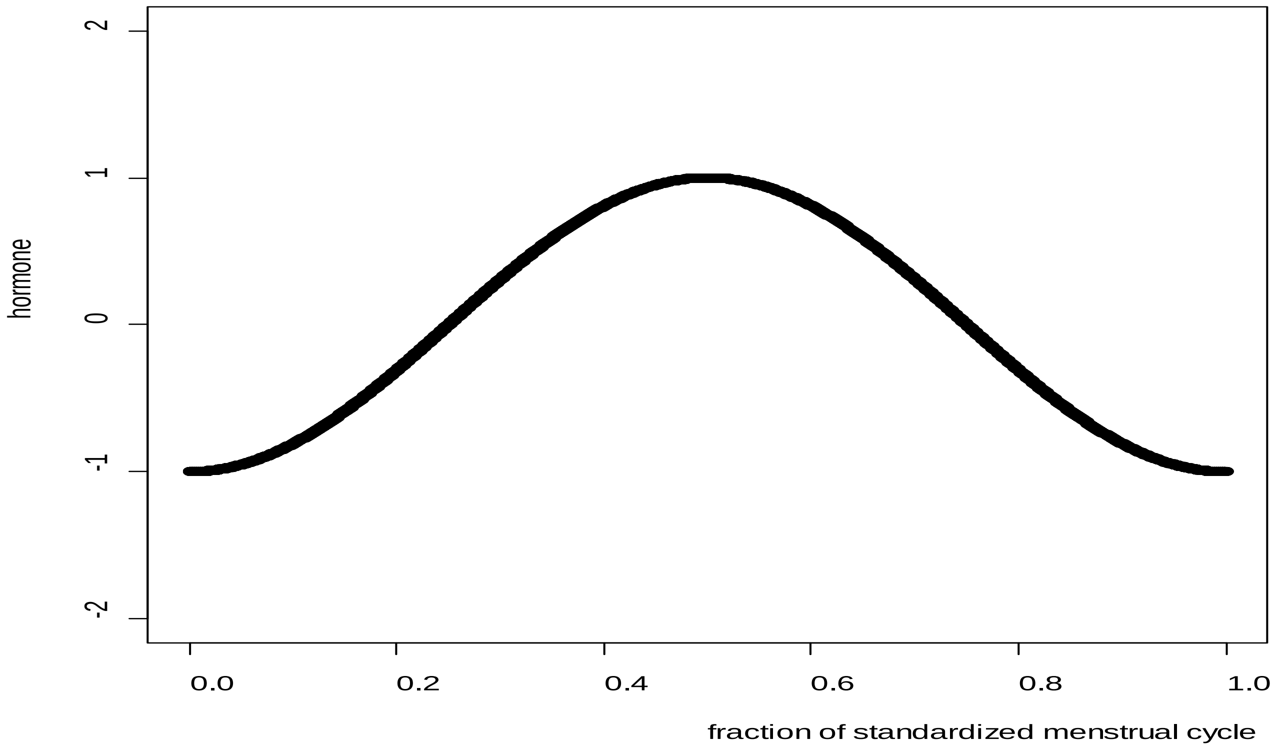
- Using periodic splines:
  - f and f' are absolutely continuous
  - f(0)=f(1), f'(0)=f'(1)
  - $\int_0^1 f(t)dt = 0$
- Using harmonic models:

$$f(t) = \sum_{k=1}^{K} \beta_k \cos(2\pi kt + \theta_k)$$

# Nonlinear Mixed Models: Shape Invariant Model

## Characterizing f with harmonic models

- K=1 (one harmonic)



# Nonlinear Mixed Models: Shape Invariant Model

## Characterizing f with harmonic models

- More than one harmonic

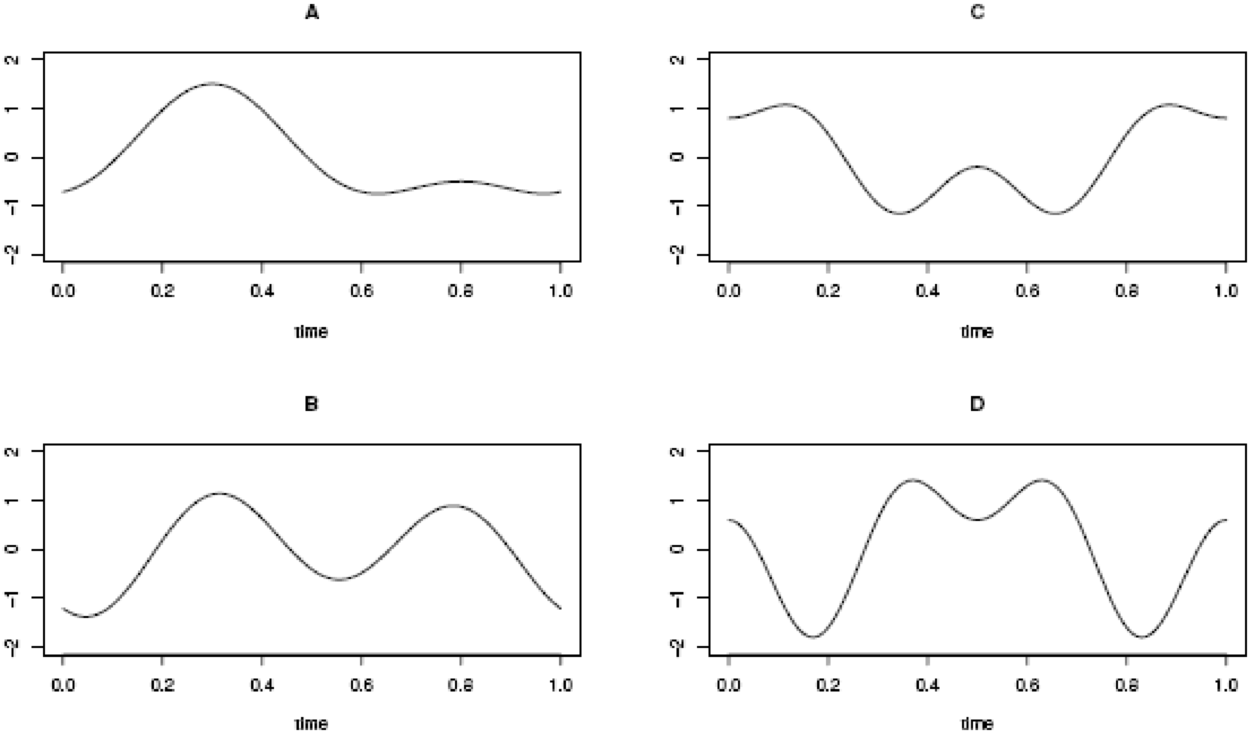


**Figure 1.** Circadian rhythms represented by $f$ with harmonic models. (A) and (B) show two-harmonic models, while (C) and (D) show three-harmonic models. Patterns are generated as $f(t) = A \times \cos(2\pi(t + B)) + C \times \cos(4\pi(t + D)) + E \times \cos(6\pi(t + F))$, where for (A): $A = 1$, $B = 0.7$, $C = 0.5$, $D = 0.7$, $E = 0$, $F = 0$, for (B): $A = 0.4$, $B = 0.5$, $C = 1$, $D = 0.2$, $E = 0$, $F = 0$, for (C): $A = 1$, $B = 0$, $C = 0.3$, $D = -0.5$, $E = 0.5$, $F = -0.5$, and for (D): $A = 1$, $B = -0.5$, $C = 0.6$, $D = 0$, $E = 1$, $F = 0$.

# Nonlinear Mixed Models: Shape Invariant Model

## Wang, Ye, and Brown- Biometrics 2003

## Albert and Hunsberger - Biometrics 2005

- $$y_{ij} = \phi_{1i} + \exp(\phi_{2i}) f\{t_{ij} - a \log it(\phi_{3i})\} + \varepsilon_{ij}$$

Mean level

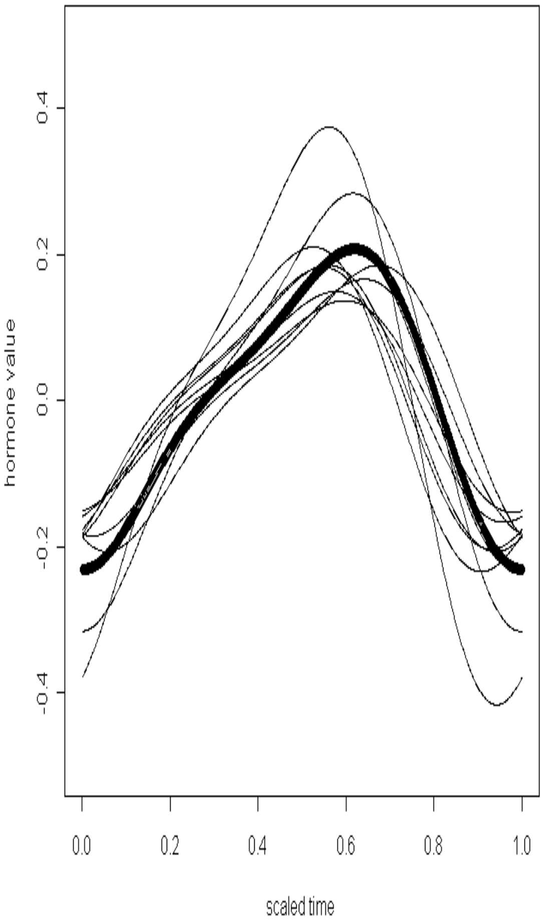
Amplitude: nadir to peak distance

Phase shift

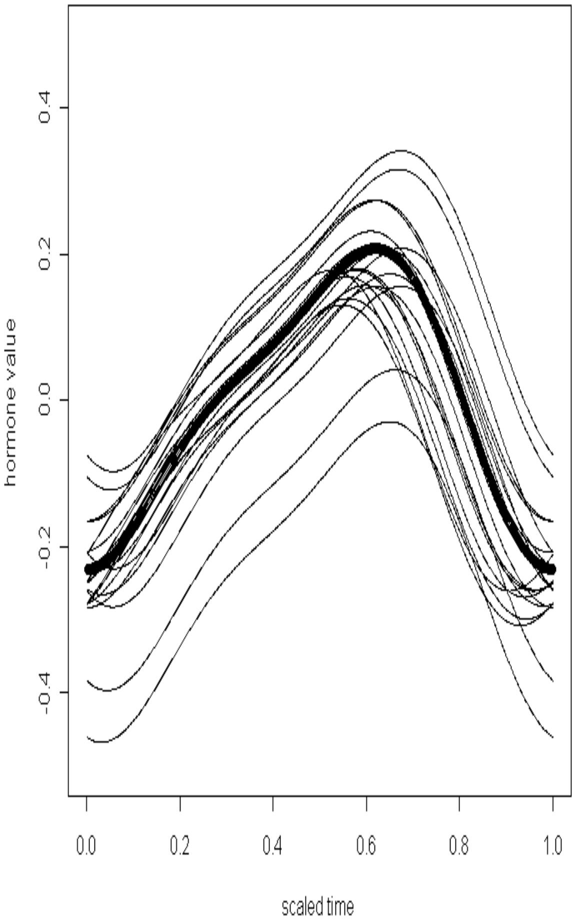
- The model allows for complex between subject variation in mean, amplitude, and phase shift

# Complex Between-Subject Variation

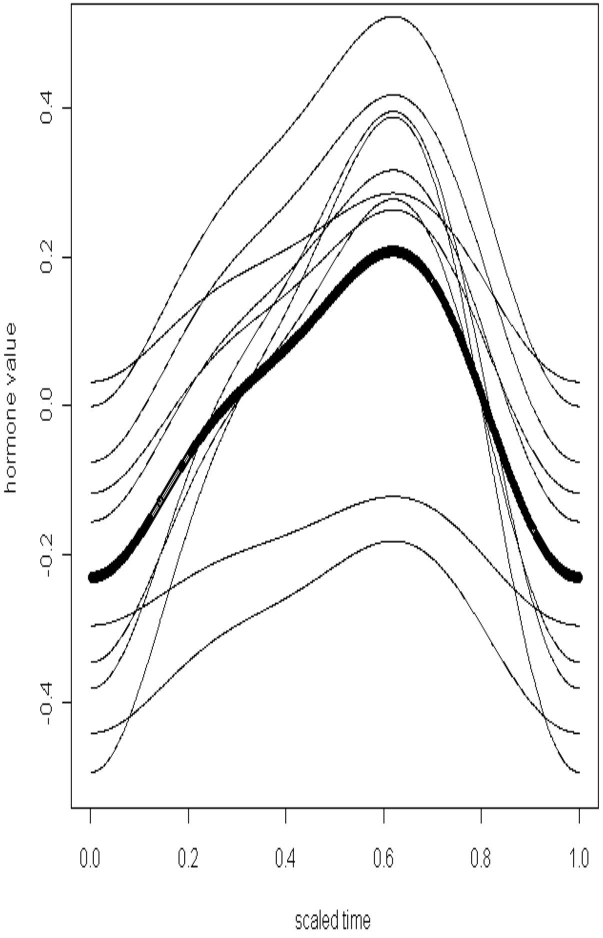
Variation in
Amplitude & Phase
**Not** Mean



Variation in
Mean & Phase
**Not** Amplitude



Variation in
Mean & Amplitude
**Not** Phase



# How to Choose the Number of Harmonics (K)?

- Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for choosing K.

$$AIC(p) = -2\log L + 2p$$

$$BIC(p) = -2\log L + p\log(\sum_{i=1}^{n} n_i)$$

- Since f is very flexible with 4 harmonics, we recommend choosing 1,2, 3, or 4 harmonics.
- Inference on covariates not very sensitive to choice of number of harmonics.

# Testing the Shape Invariant Assumption

- Simple model for two groups (G=0 or 1):

$$y_{ijG} = \tau_G + b_{1i} + \exp(b_{2i} + d_1 G) \\ \times f\{(t_{ij} - a\log it(b_{3i} + d_2 G))\} + \varepsilon_{ijk}$$

- Likelihood ratio test for shape invariance:

$$f(t) = \sum_{k=1}^{K} \beta_{1k} \cos(2k\pi(t_{ij} + \theta_{1k} - a\log it(\phi_{3i})) \\ + \beta_{2k} G \cos(2k(t_{ij} + \theta_{2k} G - a\log it(\phi_{3i})))$$

Test $\beta_{2k} = \theta_{2k} = 0$ for $k = 2,3,\ldots,K$

# Nonlinear Mixed Models: Shape Invariant Model

## Non-parametric versus harmonic modeling of f

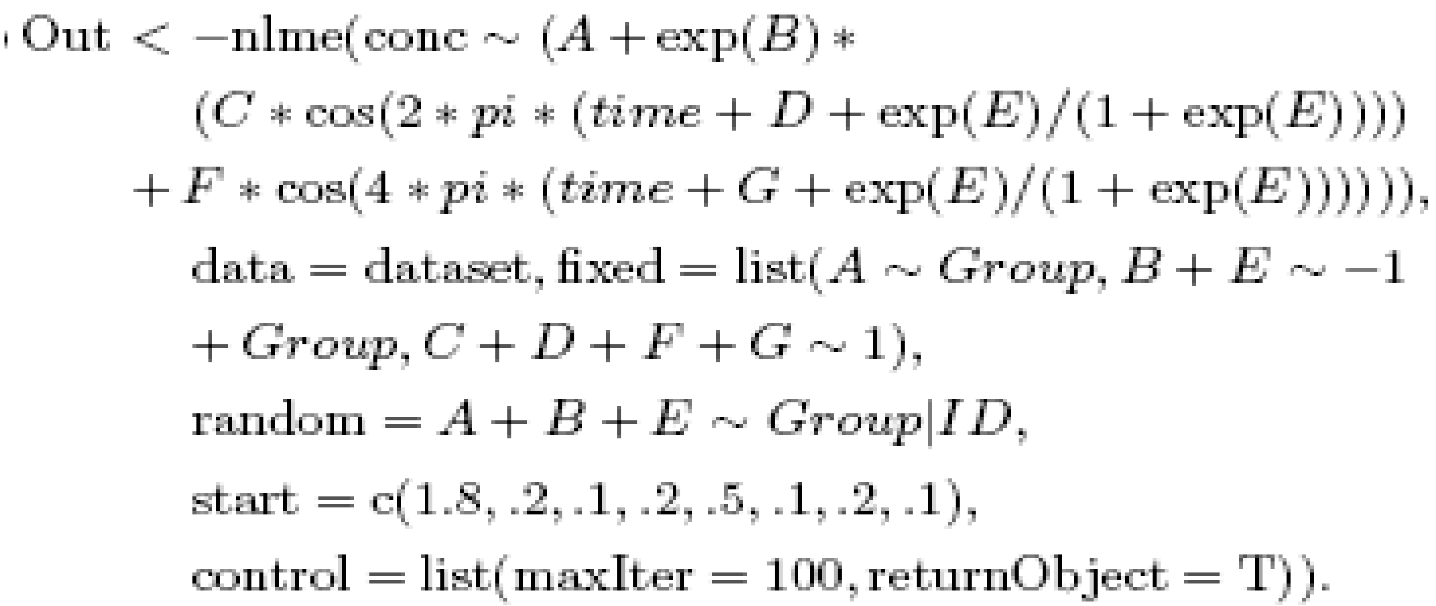
- Non-parametric models must choose the smoothing parameter while harmonic models must choose the number of harmonics.

- Non-parametric models require specialized software. Wang et al. developed a package called ASSIST for SPLUS.

- Harmonic models are easy to fit in R or SAS with standard code.

# Estimation in R using nlme

- Uses an estimation algorithm described by Lindstrom and Bates (1990)
- Sample R code:



```
Out < −nlme(conc ∼ (A + exp(B) ∗
      (C ∗ cos(2 ∗ pi ∗ (time + D + exp(E)/(1 + exp(E)))
    + F ∗ cos(4 ∗ pi ∗ (time + G + exp(E)/(1 + exp(E))))),
      data = dataset, fixed = list(A ∼ Group, B + E ∼ −1
      + Group, C + D + F + G ∼ 1),
      random = A + B + E ∼ Group|ID,
      start = c(1.8, .2, .1, .2, .5, .1, .2, .1),
      control = list(maxIter = 100, returnObject = T)).
```

# Estimation in R using nlme:
## Choosing starting values

- Choose starting values for fixed effects by fitting a non-linear model without random effects.
- Sensitivity to starting values is more pronounced when the number of harmonics is large…

# Statistical Properties of the Procedure

- Even in small samples, estimation is unbiased and tests of coefficients have the correct type-I error (with and without estimating the number of harmonics).

# Using the Shape Invariant Model for Menstrual Cycle Longitudinal Data

$$y_{ijk} = \phi_{1ik} + \exp(\phi_{2ik})\, f\{t_{ijk} / T_{ik} - a\log it(\phi_{3ik})\} + \varepsilon_{ijk}$$

k<sup></sup>th cycle

jth time point

ith subject

Standardized cycle length

$$a\log it(x) = \exp(x) / \{1 + \exp(x)\}$$

# Using the Shape Invariant Model for Menstrual Cycle Longitudinal Data

- Random Effects:

$$\phi_{1ik} = X_{1i}\beta + b_{1i} + b_{1ik}$$

$$\phi_{2ik} = X_{2i}\beta + b_{2i} + b_{2ik}$$

$$\phi_{3ik} = X_{3i}\beta + b_{3i} + b_{3ik}$$

- $(b_{1i}, b_{2i}, b_{3i})$ is $N(0, \Sigma_b)$

Between-subject variation

- $(b_{1ik}, b_{2ik}, b_{3ik})$ is $N(0, \Sigma_c)$

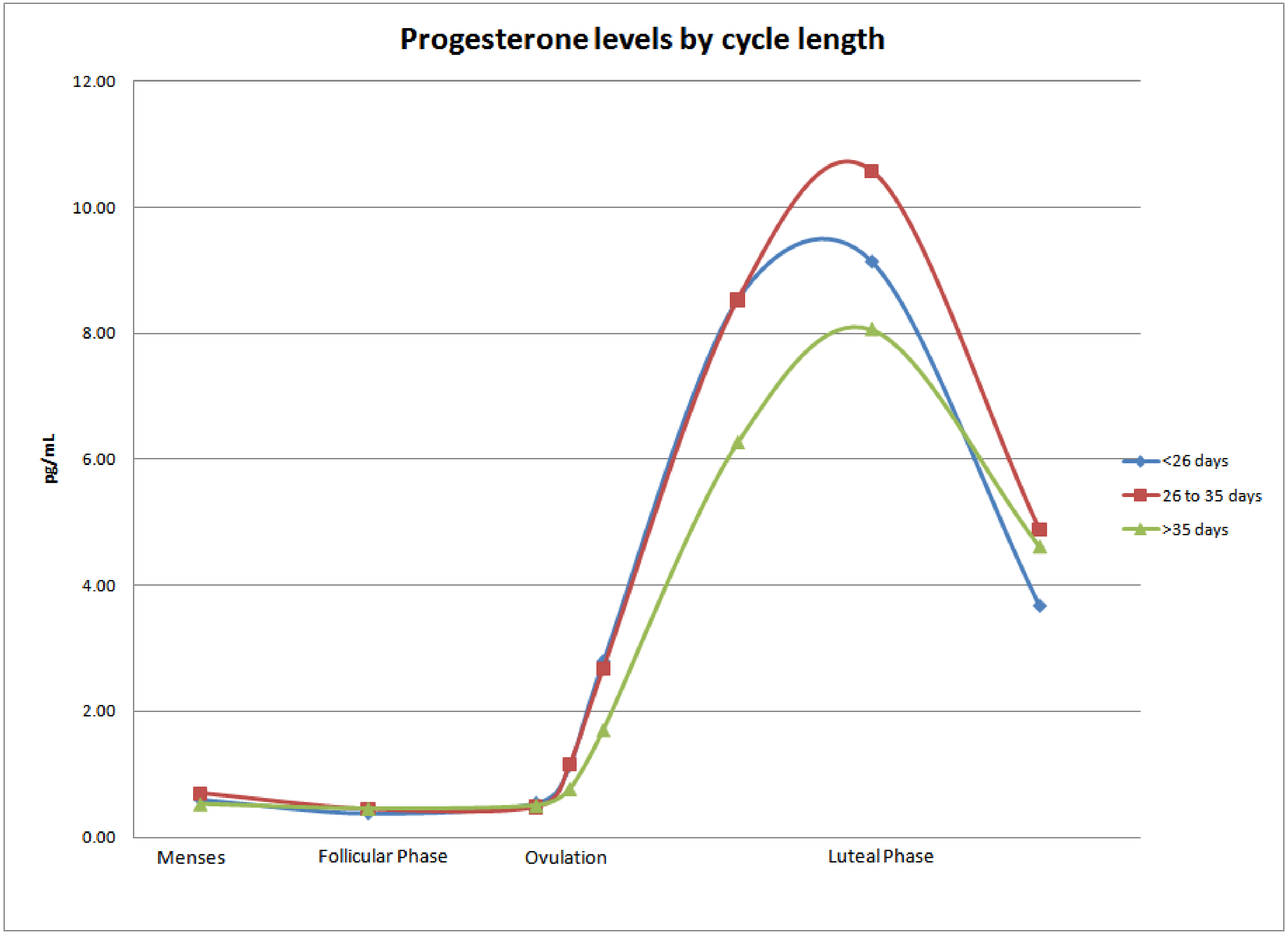
Within-subject variation

# A Comparison of Methods

| Method | Advantages | Disadvantages |
|---|---|---|
| Linear Mixed Model | - Simple to implement<br>- Built-in diagnostics<br>- Easy to incorporate multiple cycles | - Uniform measurements<br>- Studying effect of covariates on hormonal process is limited |
| Semi-parametric Stochastic Mixed Model | - Irregularly timed measurements<br>- Flexible hormonal pattern | - Studying effect of covariates on hormonal process is limited<br>- Needs specialized software |
| Non-linear Mixed Model: Shape Invariant Model | - Irregularly timed measurements<br>- Flexible way to examine covariates on hormonal process | - Shape invariance assumption |

# Advanced Topics and New Methods

# Menstrual Cycle Pattern May Vary By Cycle Length



# Joint modeling of longitudinal menstrual cycle and menstrual cycle length data

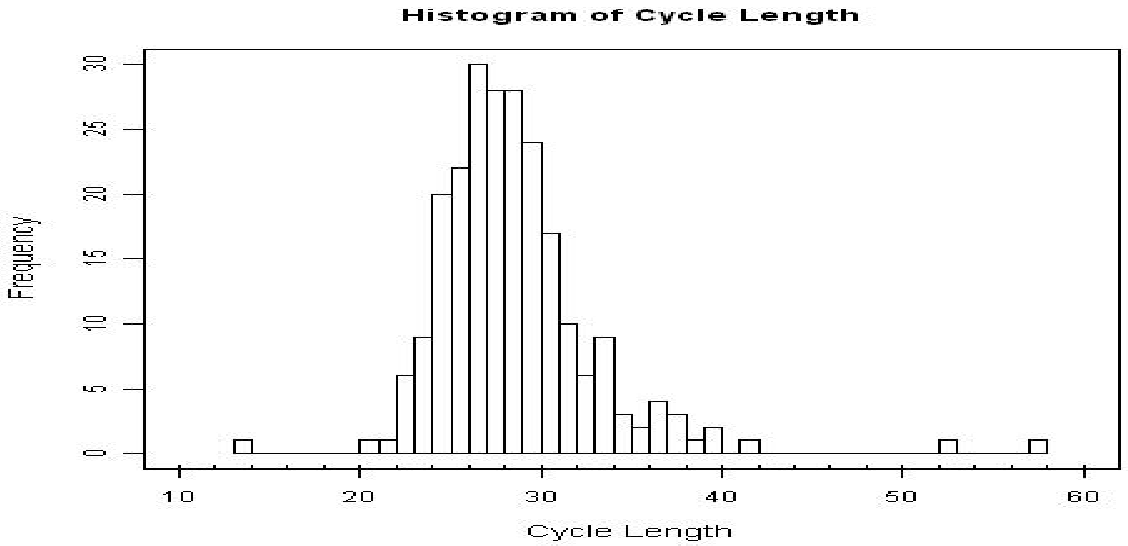
- Measure hormones at multiple “fixed” time points over a 28 day period.
- Actual cycle length has a much wider range: 12-56 days in BioCycle.
- Actual pattern may vary by menstrual cycle length

$$y_{ij} = \phi_{1i} + \exp(\phi_{2i}) f\{t_{ij} / T_i - a \log it(\phi_{3i})\} + \varepsilon_{ij}$$

$T_i$ is $N(\mu_0 + \mu_1 \phi_{1i} + \mu_2 \phi_{2i} + \mu_3 \phi_{3i}, \sigma^2)$

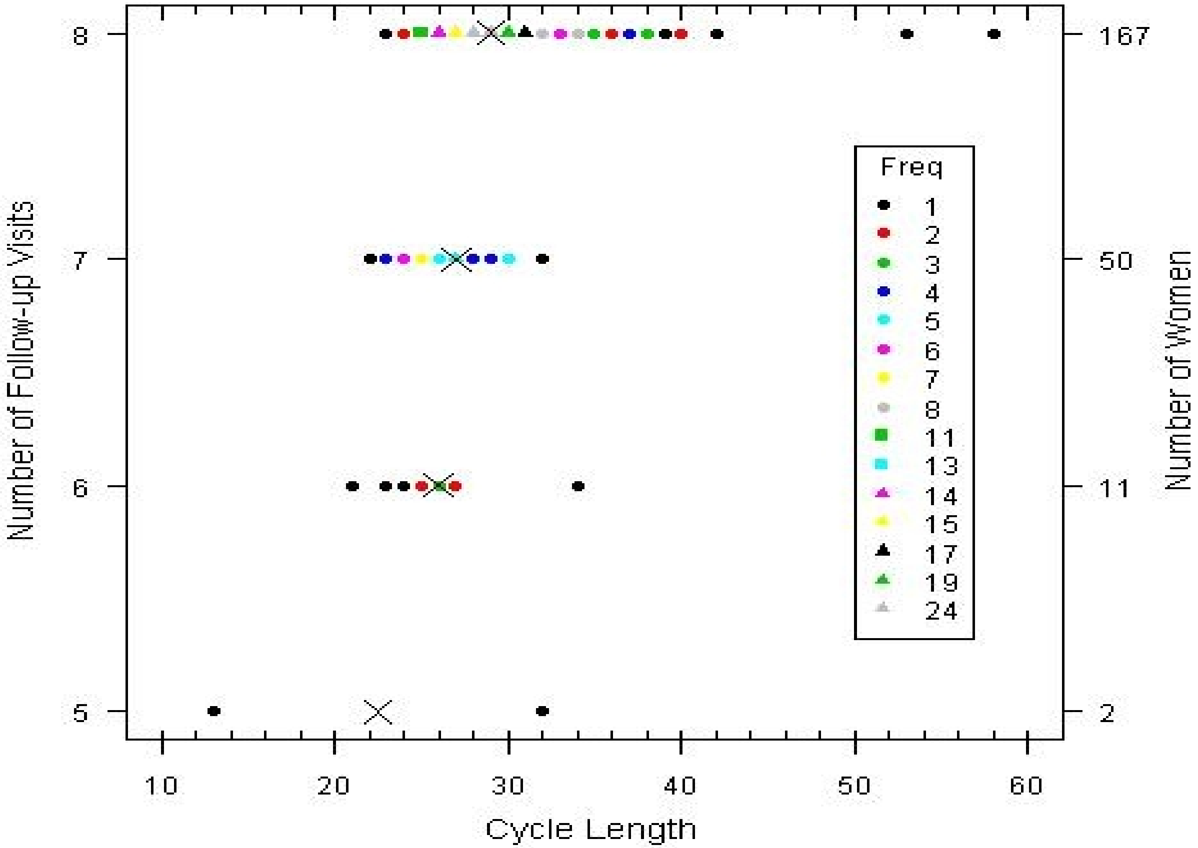
# Joint modeling of longitudinal menstrual cycle and menstrual cycle length data

- In BioCycle, women enrolled with “regular” menstrual cycle length



- Modeling menstrual cycle length using a mixture distribution (Guo et al. Biostatistics 2002)

# Relationship Between Menstrual Cycle Length and Number of Measurements



- Informative missing data problem…

# Estimation of Joint Model

- $$y_{ij} = \phi_{1i} + \exp(\phi_{2i}) f\{t_{ij} / T_i - a \log it(\phi_{3i})\} + \varepsilon_{ij}$$

  $T_i$ is $N(\mu_0 + \mu_1\phi_{1i} + \mu_2\phi_{2i} + \mu_3\phi_{3i}, \sigma^2)$

- $$L = \prod_{i=1}^{I} \int_b \prod_{j=1}^{J_i} f(y_{ij} \mid T_i, b) f(T_i \mid b) g(b) db$$

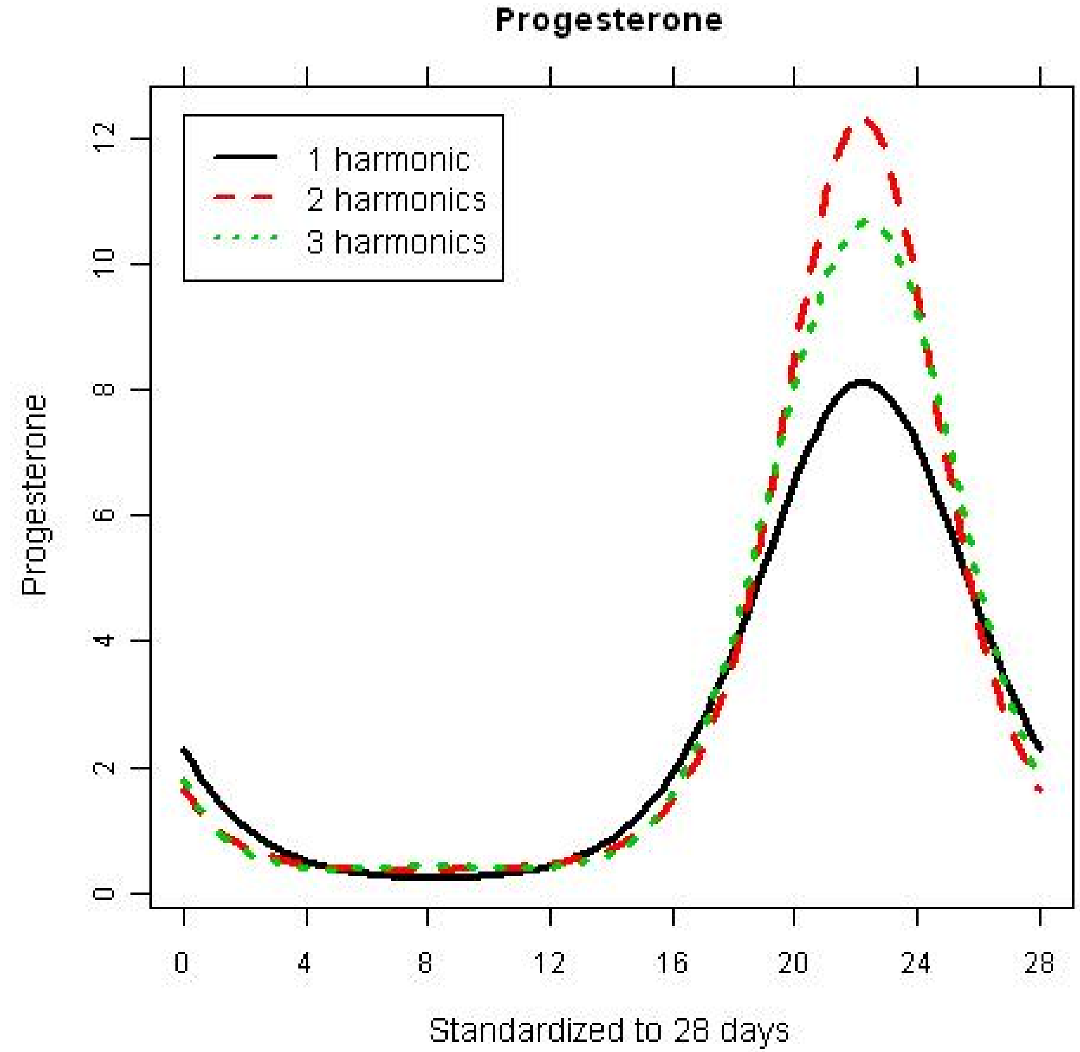
- Use trapezoidal rule for numerical integration
- Obtain maximum-likelihood estimates

# Results of Analyses for Progesterone

- 3 harmonics model: Joint model and naïve model not considering menstrual cycle length

| | Joint model | | Naïve model | |
|---|---|---|---|---|
| **Parameters** | **Estimate** | **SE** | **Estimate** | **SE** |
| $\mu_0$ | 29.2 | 0.30 | 29.0 | 0.30 |
| $\mu_1$ | -5.06 | 1.47 | | |
| $\mu_2$ | -6.31 | 2.81 | | |
| $\mu_3$ | -0.19 | 2.66 | | |
| $\beta_1$ | 0.39 | 0.03 | 0.39 | 0.03 |
| $\beta_2$ | 0.55 | 0.02 | 0.56 | 0.02 |
| $\beta_3$ | 1.33 | 0.02 | 1.33 | 0.02 |
| $\sigma_{b1}$ | 0.38 | 0.02 | 0.38 | 0.02 |
| $\sigma_{b2}$ | 0.20 | 0.02 | 0.19 | 0.02 |
| $\sigma_{b3}$ | 0.28 | 0.02 | 0.28 | 0.02 |

# Results for Progesterone



# Simulation Studies

- Under the correct joint model - 3 harmonics, 50 individuals, 500 simulated datasets, and 5 time points: 2, 7, 13, 18, and 27.

| Parameters | Truth | Estimate | MC SE | Aympt SE |
|---|---|---|---|---|
| $\mu_0$ | 29 | 29.0 | 0.25 | 0.24 |
| $\mu_1$ | 1 | 1.02 | 0.27 | 0.27 |
| $\mu_2$ | 1 | 1.01 | 0.26 | 0.24 |
| $\mu_3$ | 1 | 1.00 | 0.24 | 0.22 |
| $\beta_1$ | 1 | 1.01 | 0.14 | 0.14 |
| $\beta_2$ | 1 | 0.99 | 0.13 | 0.11 |
| $\beta_3$ | 0 | 0.00 | 0.04 | 0.02 |

# Simulation Studies

- Naïve model - 3 harmonics, 50 individuals, 500 simulated datasets, and 5 time points: 2, 7, 13, 18, and 27.

| Parameters | Truth | Estimate | MC SE | Aympt SE |
|---|---|---|---|---|
| $\beta_1$ | 1 | 1.00 | 0.16 | 0.14 |
| $\beta_2$ | 1 | 1.00 | 0.16 | 0.11 |
| $\beta_3$ | 0 | 0.00 | 0.18 | 0.02 |

# Simulation Studies

- Naïve model -3 harmonics, 50 individuals, 1000 simulated datasets, and 5 time points: 2, 7, 13, 18, and 27 :

$$\mu_1 = \mu_2 = 10$$

| Parameters | Truth | Estimate | MC SE | Aympt SE |
|---|---|---|---|---|
| $\beta_1$ | 1 | 1.21 | 0.19 | 0.15 |
| $\beta_2$ | 1 | 1.26 | 0.38 | 0.10 |
| $\beta_3$ | 0 | 0.00 | 0.01 | 0.01 |

# Extensions to Modeling Multivariate Hormonal Profiles

mth hormone

$$y_{ijmk} = \phi_{1imk} + \exp(\phi_{2imk}) f\{t_{ijk} / T_{ik} - a\log it(\phi_{3imk})\} + \varepsilon_{ijmk}$$

$$a\log it(x) = \exp(x) / \{1 + \exp(x)\}$$

$$\phi_{1imk} = X_{1i}\beta_m + b_{1im} + b_{1ikm}$$

$$\phi_{2imk} = X_{2im}\beta_m + b_{2im} + b_{2ikm}$$

$$\phi_{3imk} = X_{3im}\beta_m + b_{3im} + b_{3ikm}$$

# Extensions to Modeling Multivariate Hormonal Profiles

- Let $b_{im} = (b_{1im}, b_{2im}, b_{3im})'$ and $b_{ikm} = (b_{1ikm}, b_{2ikm}, b_{3ikm})'$

- Assume that

$$Cov((b_{i1}, b_{i2}, \ldots, b_{iM})') = \Sigma_b$$

Between-Subject Variation

$$Cov((b_{ik1}, b_{ik2}, \ldots, b_{ikM})') = \Sigma_W$$

Within-Subject Variation

# Acknowledgements

- This work was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health & Human Development, National Institutes of Health.
- BioCycle working group: Sunni Mumford, Anna Pollack, Enrique Schisterman, Aijun Ye, Jean Wactawski-Wende, Edwina Yeung, Cuilin Zhang

# References

1. Albert, P.S. and Hunsberger, S. (2005). On analyzing circadian rhythms data using nonlinear mixed models with harmonic terms. *Biometrics* **61**, 1115-1120.
2. Guo, Y., Manatunga, A.K., Chen, S., and Marcus, M. (2006). Modeling menstrual cycle length using a mixture model. *Biostatistics* **7**, 100-114.
3. Meyer, P.M., Zeger, S.L., et al. (2007). Characterizing daily urinary hormone profiles for women at midlife using functional data analysis. *American Journal of Epidemiology* **165**, 936-945.
4. Pinheiro, J.C. and Bates, D.M. *Mixed-Effects Models in S and S-PLUS*. Springer, New-York. 2000. ISBN 0-387-98957-9
5. 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.
6. Zhang, D., Raz, J., and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93**, 710-719
7. Zhang, D. Lin, X., and Sowers, M. (2000). Semiparametric regression for periodic longitudinal hormone data with multiple menstrual cycles. *Biometrics* **56**, 31-39.