Statistical Modeling of Sunspot Cycles

Yaming Yu

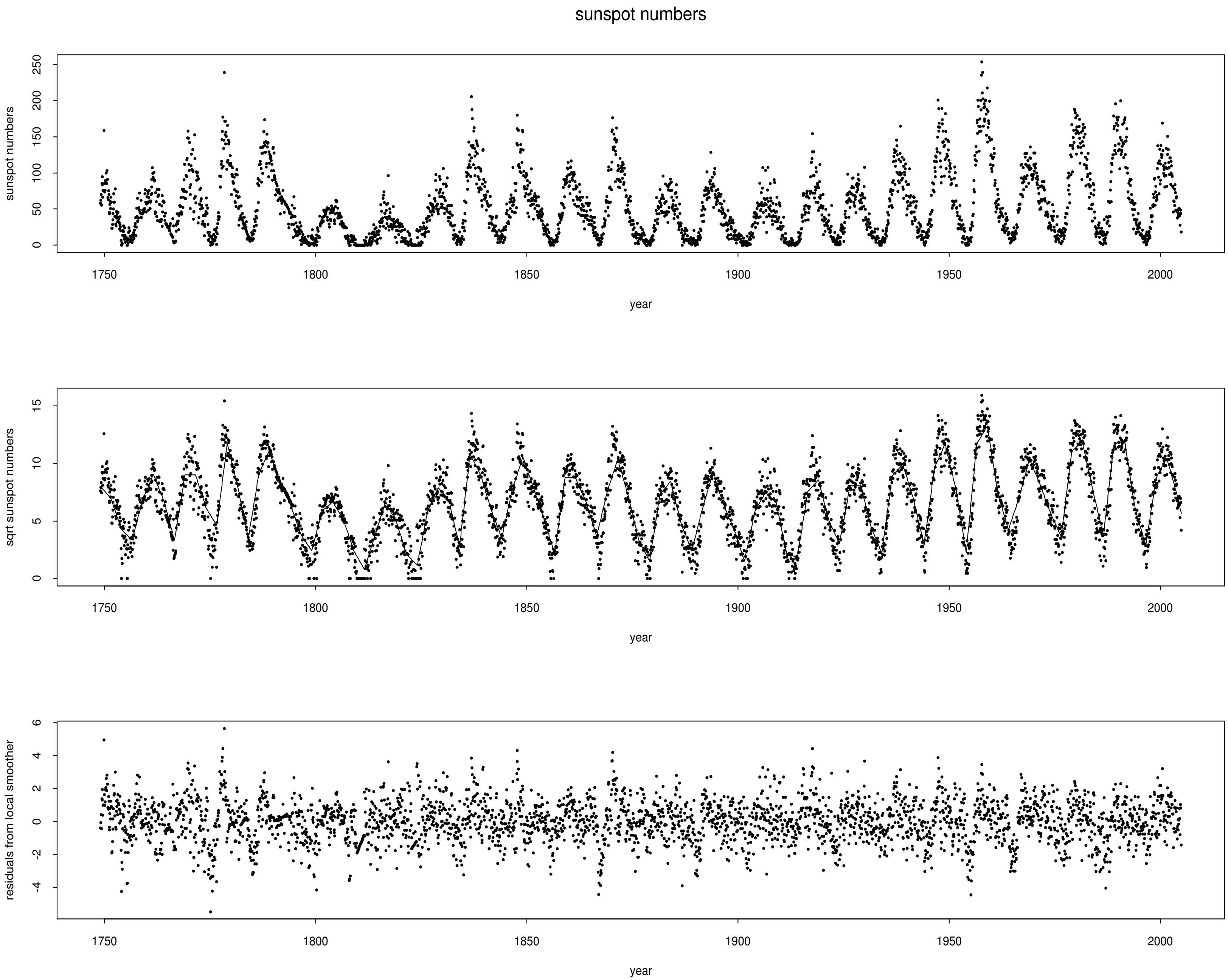
Department of Statistics, Harvard University

Sunspots

- What are they?
  - Sunspots appear as dark spots on the surface of the Sun.
  - Temperature lower than the surrounding photosphere.
    Strong magnetic fields.
  - They typically last several days; some may live for weeks.
- The longest directly observed index of solar activity
  - 1610: Galileo first viewed sunspots with his new telescope.
  - 1749: Daily observations were started at the Zurich Observatory.
  - 1849: Continuous (daily) observations were obtained with the addition of more observatories.

Sunspot Number (SSN) Data

- Sunspots occur in groups.
- Sunspot No. = No. of individual spots + 10 $\times$ No. of groups
- – The International Sunspot Number: compiled by the Sunspot Index Data Center in Belgium.
  – The NOAA sunspot number: compiled by the US National Oceanic and Atmospheric Administration.
- – Top: monthly averages of the International Sunspot Numbers.
  – Middle: local smoother fit to sqrt(SSN).
  – Bottom: residuals.


sunspot numbers
sunspot numbers
0
50
100
150
200
250
1750
1800
1850
1900
1950
2000
year
sqrt sunspot numbers
0
5
10
15
year
residuals from local smoother
-4
-2
0
2
4
6
year

Sunspot Cycles

Features of the sunspot number data

- **A lot of noise.**
- **Quasi-periodicity: average cycle length is 11 years (Wolf 1852).**
- **Asymmetry: rise to maximum is faster than fall to minimum (Waldmeier 1935, 1939).**
- **Waldmeier effect: stronger cycles tend to take less time to rise to maximum amplitude.**
- **Long-term (8–9 cycles) periodicity . . .**

How to quantify the statistical significance?

Statistical Modeling of Sunspot Cycles

- **Physical models of the solar dynamo are unfortunately lacking/flawed.**
- **But we can build statistical models.**

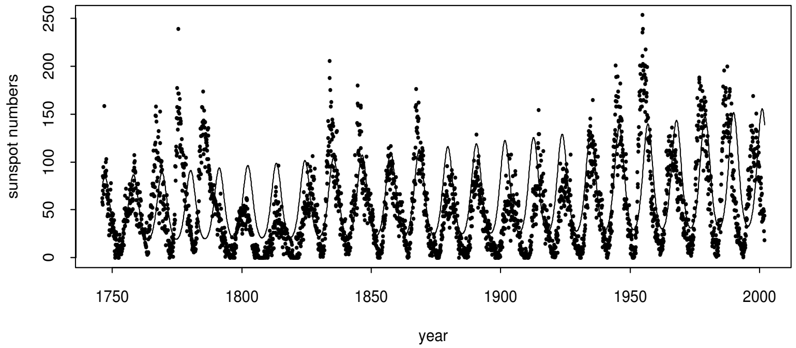
Cycle lengths vary; purely periodic models don't work.

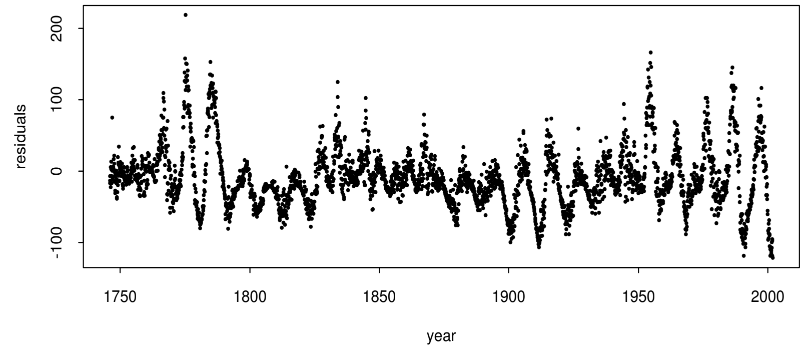
- **A Poisson model with a latent autoregressive process**

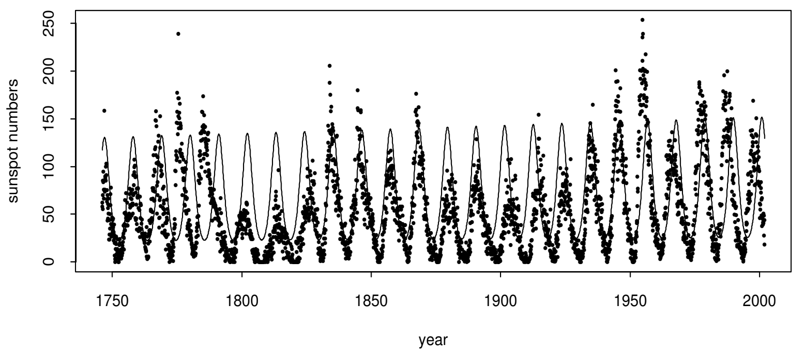
$$Y_t|(\xi_t, \beta) \stackrel{ind}{\sim} Pois\left(e^{\beta_0+\beta_1 t+\beta_2 \cos(2\pi t/T+t_0)+\xi_t}\right);$$

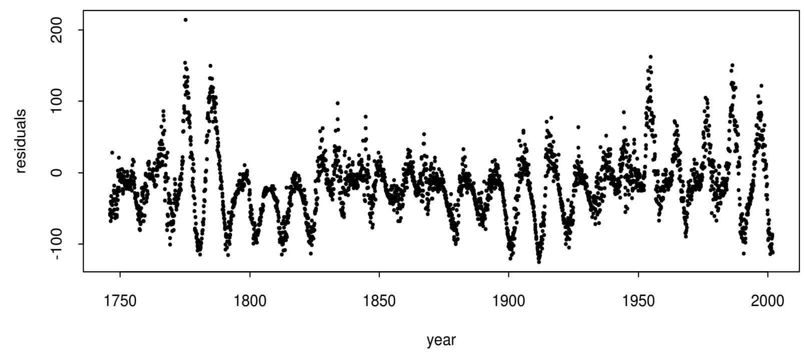
$$\xi_t|(\xi_{<t}, \beta, \rho, \delta) \sim N(\rho\xi_{t-1},\ \delta^2).$$

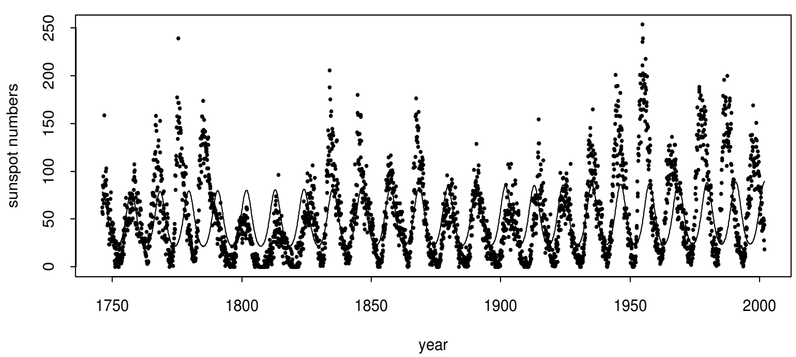
- **Three posterior realizations**
  - **Left: data with fitted curve**
  - **Right: residuals**

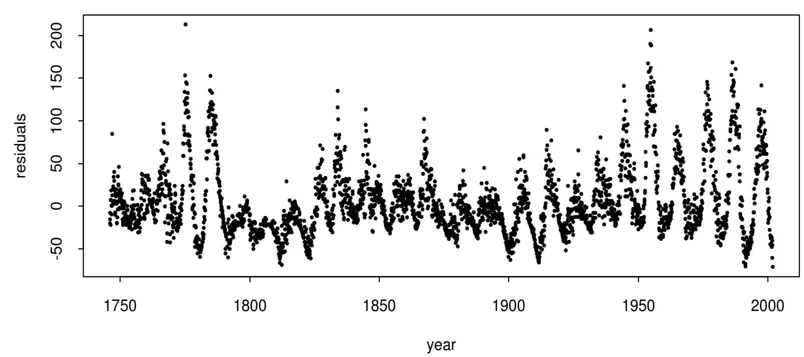

sunspot numbers
0
50
100
150
200
250
1750
1800
1850
1900
1950
2000
year


residuals
-100
0
100
200
1750
1800
1850
1900
1950
2000
year


sunspot numbers
0
50
100
150
200
250
1750
1800
1850
1900
1950
2000
year


residuals
-100
0
100
200
1750
1800
1850
1900
1950
2000
year

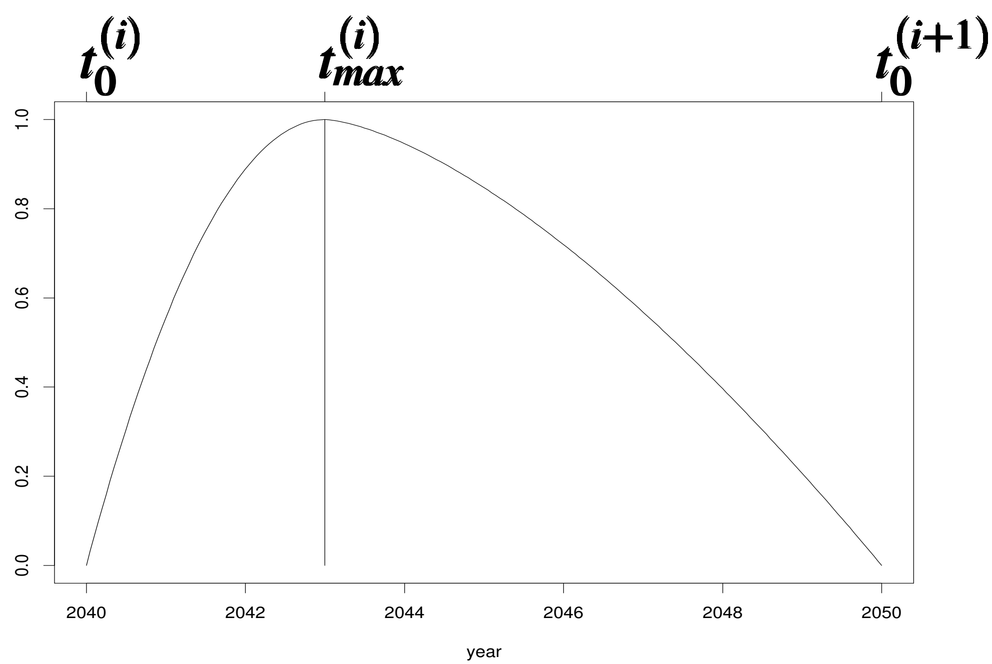

sunspot numbers
0
50
100
150
200
250
1750
1800
1850
1900
1950
2000
year


residuals
-50
0
50
100
150
200
1750
1800
1850
1900
1950
2000
year

Modeling Each Cycle by Simple Functions

Notation for cycle i

- $t_0^{(i)}$: start of cycle i
- $t_{max}^{(i)}$: time at cycle maximum
- $t_0^{(i+1)}$: end of cycle i



- R_t: "average solar activity level" at time t
  - For the rising phase $t < t_{max}^{(i)}$

$$R_t = c_i \left(1 - \left(\frac{t_{max}^{(i)} - t}{t_{max}^{(i)} - t_0^{(i)}}\right)^{\alpha_1}\right);$$

  - For the declining phase $t > t_{max}^{(i)}$

$$R_t = c_i \left(1 - \left(\frac{t - t_{max}^{(i)}}{t_0^{(i+1)} - t_{max}^{(i)}}\right)^{\alpha_2}\right).$$

- cycle length = $t_0^{(i+1)} - t_0^{(i)}$;
  time to rise to maximum = $t_{max}^{(i)} - t_0^{(i)}$;
  amplitude = c_i.
- $\alpha_1, \alpha_2 > 1$: the same shape parameters for all cycles.

A Nonlinear Regression Model

- **Model sqrt of sunspot numbers to stablize the variance:**

$$\sqrt{Y_t} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 t + R_t,\ \sigma^2)$$

- **Cycle-specific parameters**
  - $T_0 = (t_0^{(i)},\ i = 0, 1, \ldots, k)$;
  - $T_{max} = (t_{max}^{(i)},\ i = 0, \ldots, k-1)$;
  - $C = (c_i,\ i = 0, \ldots, k-1)$.

  Total number of available cycles $k = 24$.

Priors

- flat on $t_0^{(i)}$, $i = 1, \ldots, k-1$ and T_{max} subject to

$$t_0^{(i)} < t_{max}^{(i)} < t_0^{(i+1)};$$

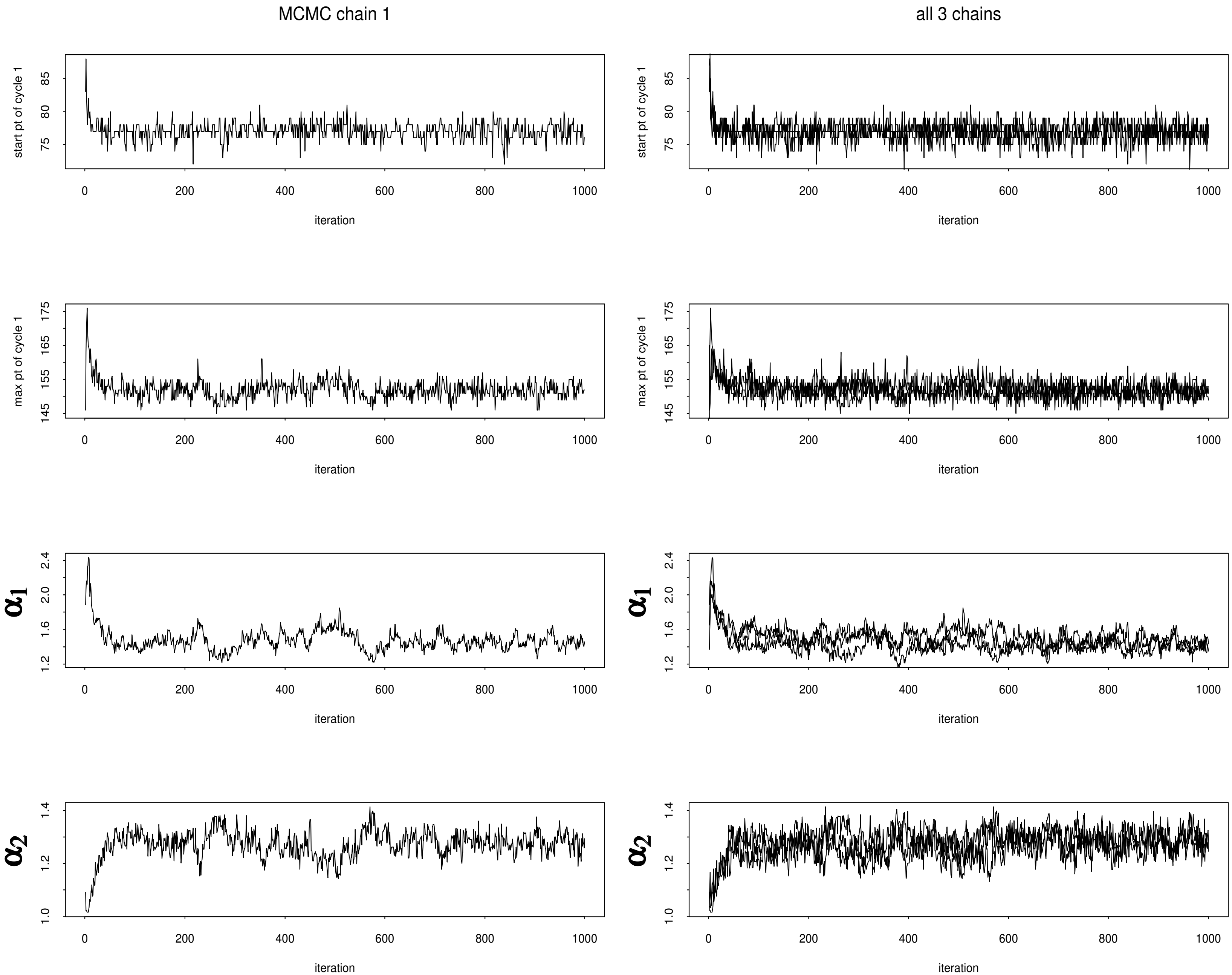
- flat but with extra constraint on $t_0^{(0)}$, $t_0^{(k)}$ and $\alpha = (\alpha_1,\ \alpha_2)$;
- standard prior on C, $\beta = (\beta_0,\ \beta_1)$, and σ^2.

Model-fitting Procedure

- **Gibbs sampler with M–H steps.**
- **Lots of local modes in simulations.**

Note: Given T_0, T_{max} and α, posterior of (C, β, σ^2) follows standard normal-inverse χ^2. So

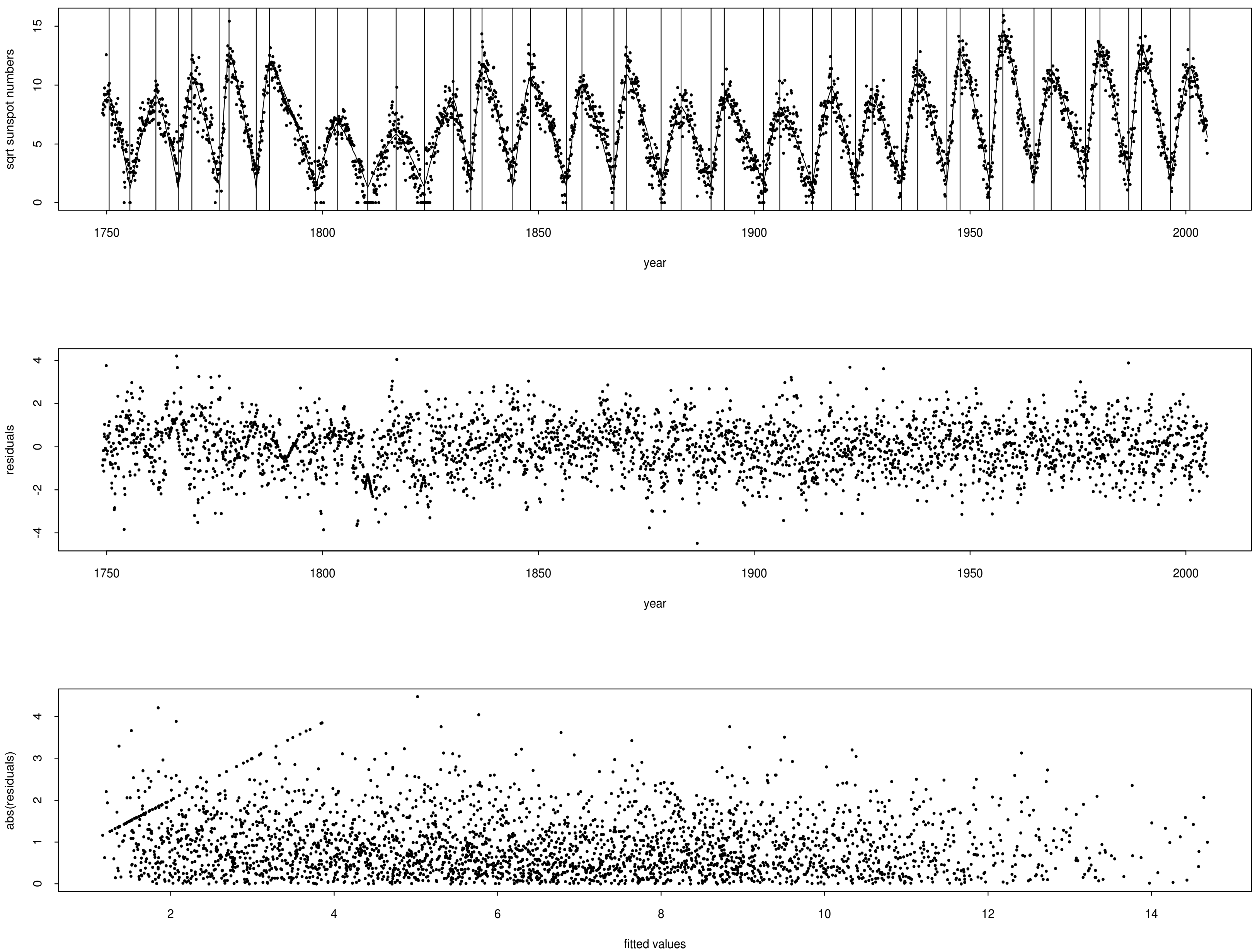
- **update (T_0, T_{max}, α) one coordinate at a time according to its conditional density, but with (C, β, σ^2) integrated out;**
- **draw (C, β, σ^2) given (T_0, T_{max}, α) using OLS routines.**


MCMC chain 1
all 3 chains
start pt of cycle 1
max pt of cycle 1
α1
α2
iteration

Posterior Inference

Fitted model and residuals

- – **Top: sqrt(SSN) with fitted values.**
  Vertical lines represent one posterior draw of (T_0, T_{max}).
  – **Middle: residuals vs. time (year).**
  – **Bottom: residuals vs. fitted values.**

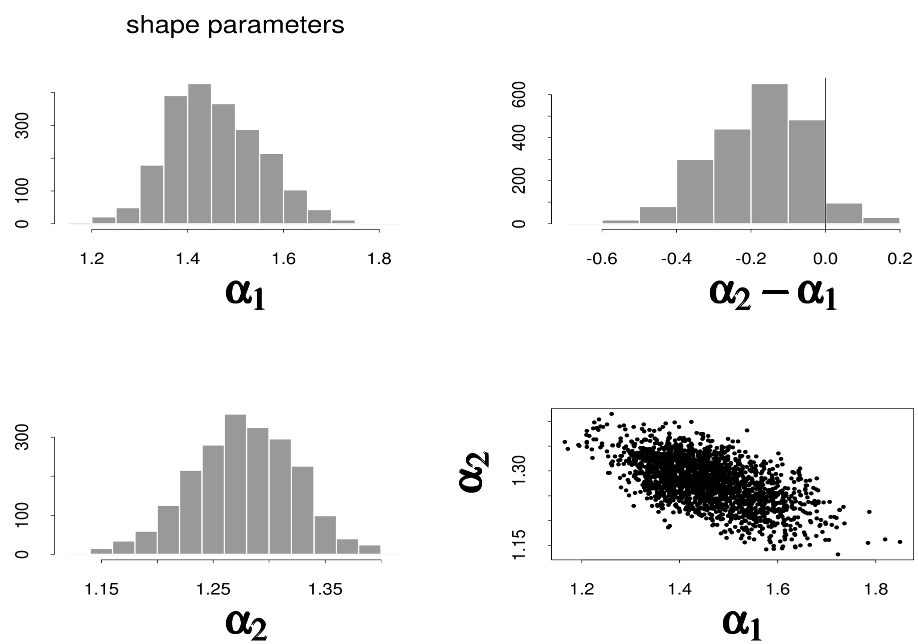

sqrt sunspot numbers
year
residuals
year
abs(residuals)
fitted values

- **It's a fairly good fit. Much better than the local smoother.**
- **The fit is better for recent data ($year > 1850$) than for the less reliable data in the past.**
- **The 45 degree streak is an artifact caused by zero SSN observations.**

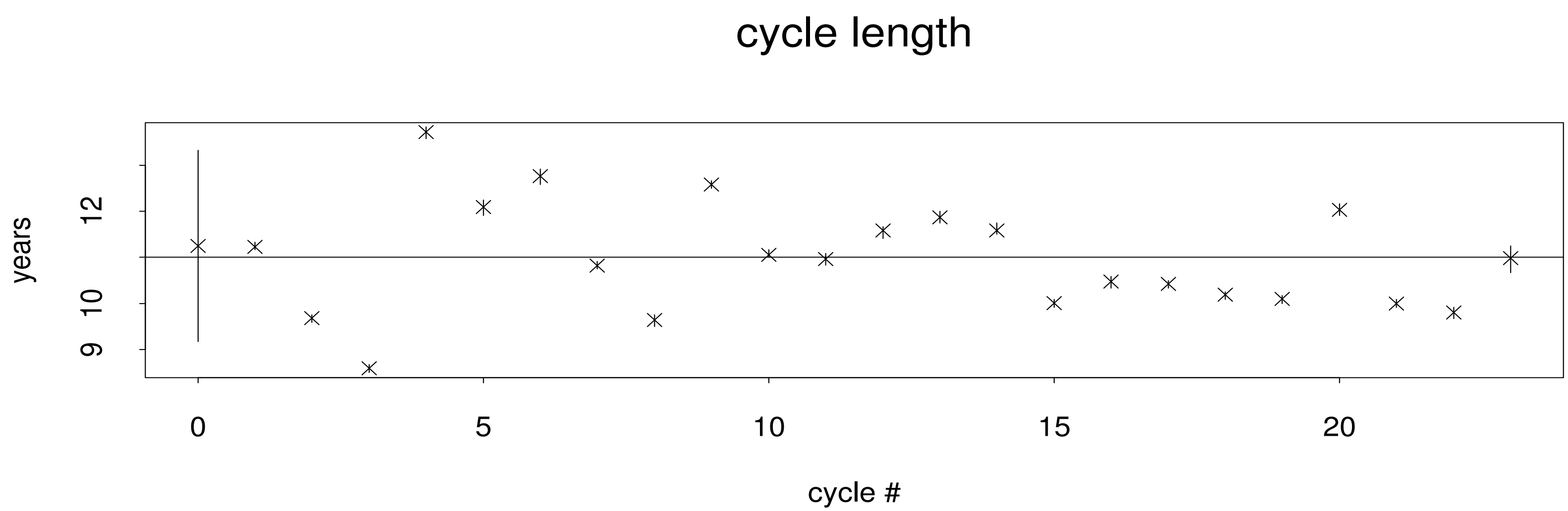
Posterior Inference: Shape Parameters α_1, α_2

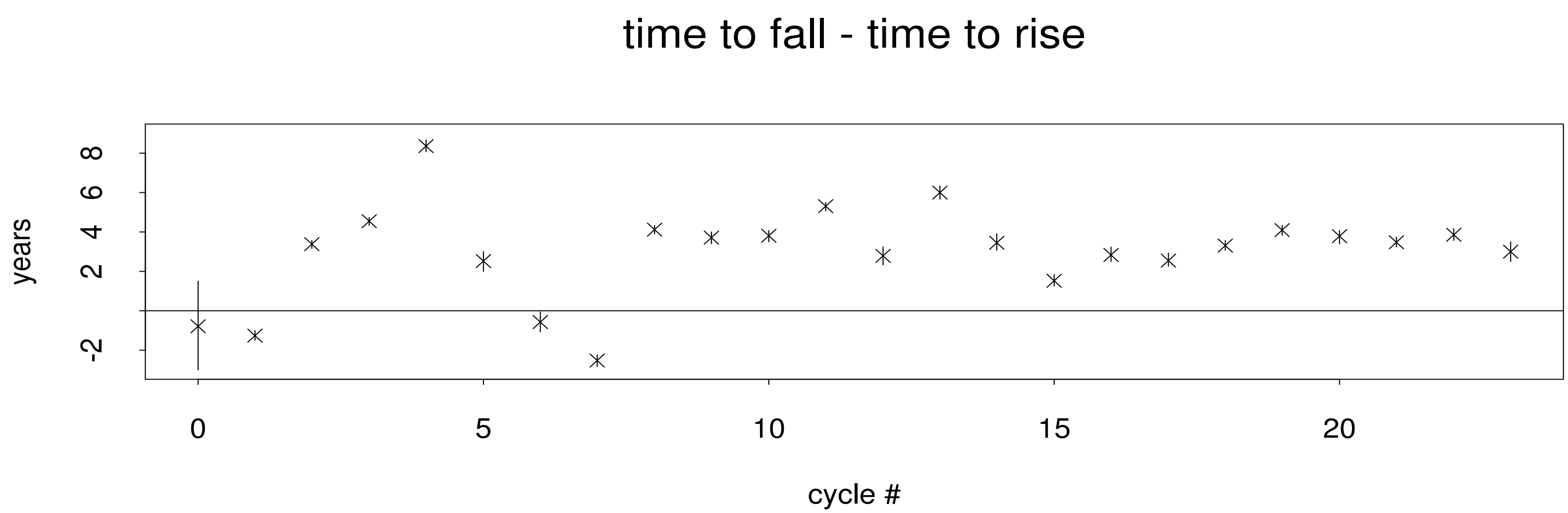
	mean	s.e.	2.5%	97.5%
α_1	1.46	0.10	1.29	1.66
α_2	1.28	0.05	1.18	1.36

$$\Pr(\alpha_2 - \alpha_1 < 0 | Y) = 0.94$$



Cycle Length Patterns





- **Average cycle length is around 11 years**
  ($\times$'s mark posterior means)

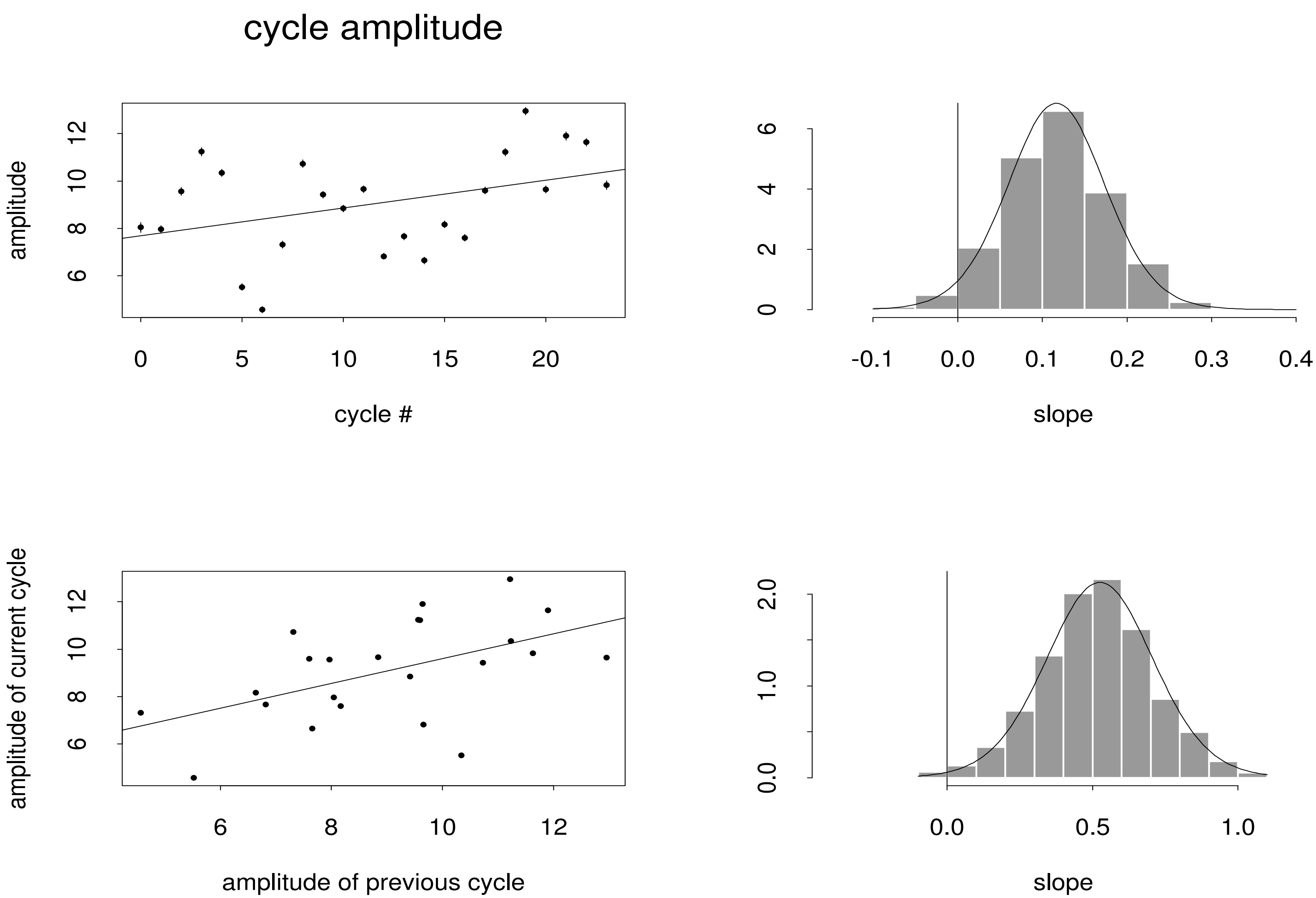
- **Error bars are small**
  (Vertical bars represent the 50% marginal credible intervals)

- **The cycle length has no apparent upward or downward trend.**

- **With few exceptions, cycles take more time to decline than to rise.**

- **Only about half of Cycle # 0 is observed, hence the large error bars.**

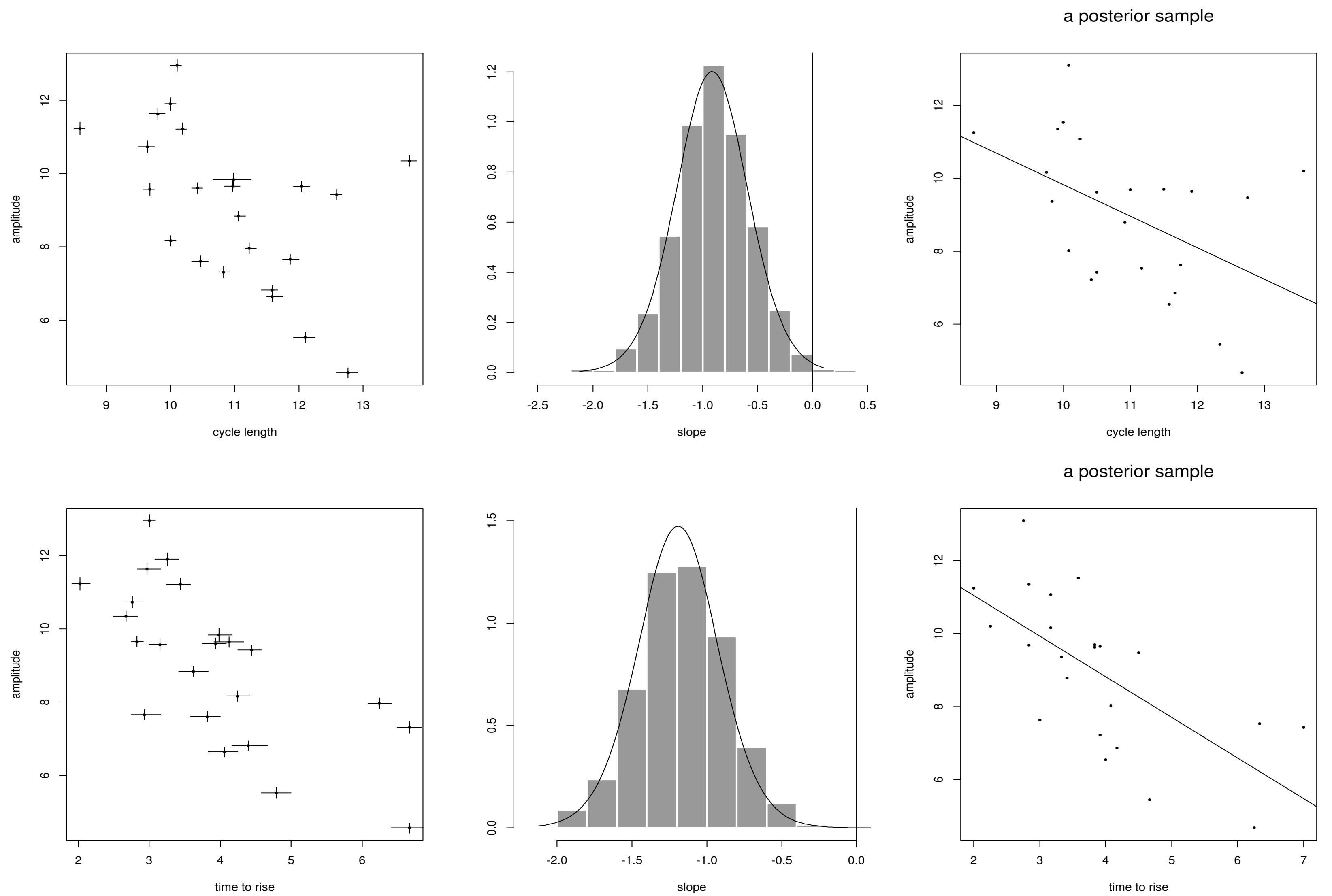
Cycle Amplitude Patterns



Evaluating Statistical Significance

- **Wrong procedure: simple linear regression using the posterior mean as the true amplitudes.**
- **Ideally we should fit a hierarchical model.**
- **A two-stage simulation procedure:**
  - **Draw posterior samples of the cycle amplitudes (done).**
  - **For each sample, fit the regression model of amplitude vs. cycle #, and then draw from the posterior of the regression coefficient.**
- **Because error bars are small, results (histogram) are nearly identical to those of simple linear regression (solid curve).**

Relationship Between Cycle Length and Amplitude



- **Row 1: amplitude vs. cycle length**
  - **Left: Scatterplot of the posterior means.**
    Vertical (horizontal) bars are 50% credible intervals for cycle amplitude (length).
  - **Middle: Statistical significance of the regression slope.**
    Little difference between simple linear regression and two-stage simulation.
  - **Right: A posterior sample and its regression line.**
- **Row 2: amplitude vs. time to rise to cycle maximum**
  - **Middle: the error bars are large enough to make a (very small) difference.**

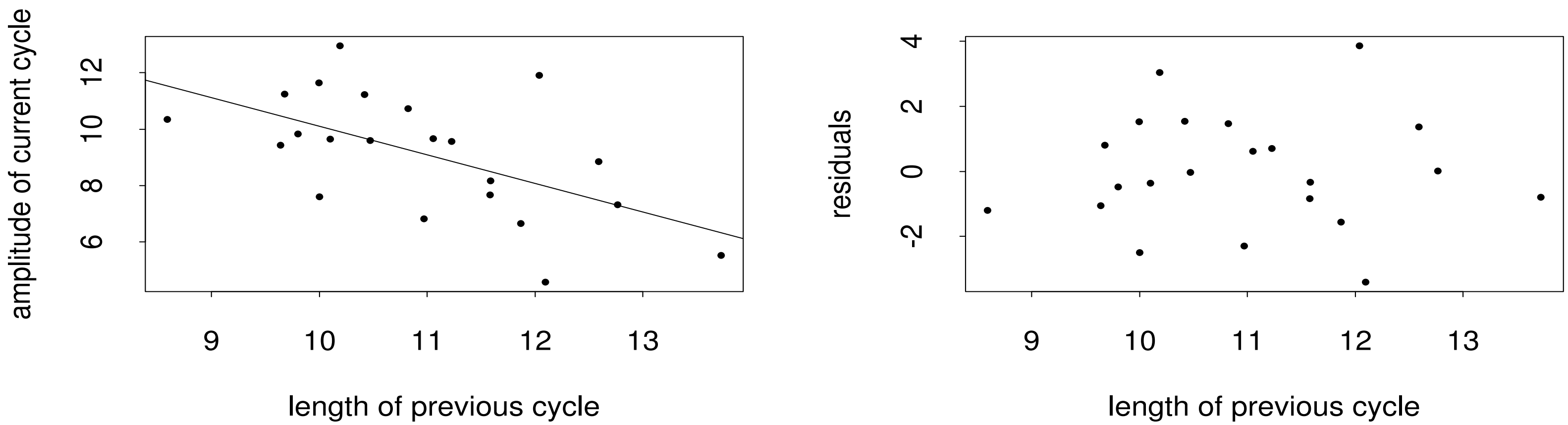
Forecasting Problems

- **Predict the rest of a partially observed cycle**
- **Predict the length and amplitude of an unobserved future cycle**

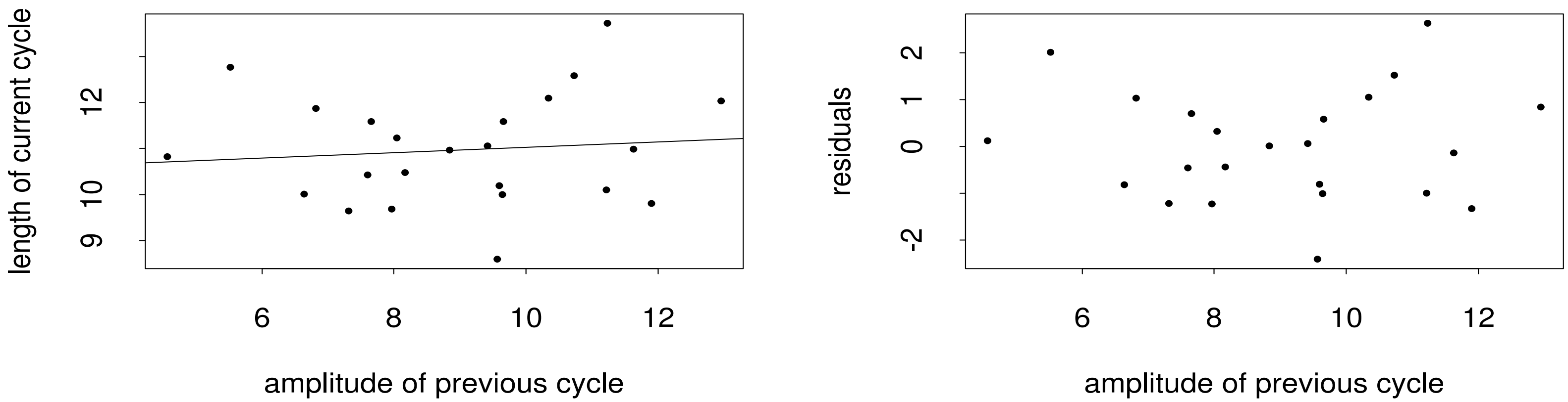
The amplitude-length (amplitude-period) relations:

- **Length of the previous cycle is a fairly good predictor of the amplitude of the current cycle.**
- **Amplitude of the previous cycle has little correlation with the length of the current cycle.**

prediction



prediction



Work in Progress

- **Data quality problems.**
- **Incorporating additional information, e.g., spatial location of sunspots, magnetic polarity information; joint modeling with 10.7cm flux, etc.**
- **A more elaborate model to link cycle length, time to rise, and amplitude through hyperparameters.**
- **Allowing the start of cycle $i+1$ to be slightly different from the end of cycle i.**
- **Comparison with similar models in the literature.**
- **Better algorithms. More efficient computer code.**
- $\cdots$