Understanding Climate change using Functional data analysis techniques

GLACIER WORKSHOP, Aug 2010

Surajit Ray

Boston University (Dept of Mathematics and Statistics)

Co-investigators Giles Hooker(Cornell Stat)

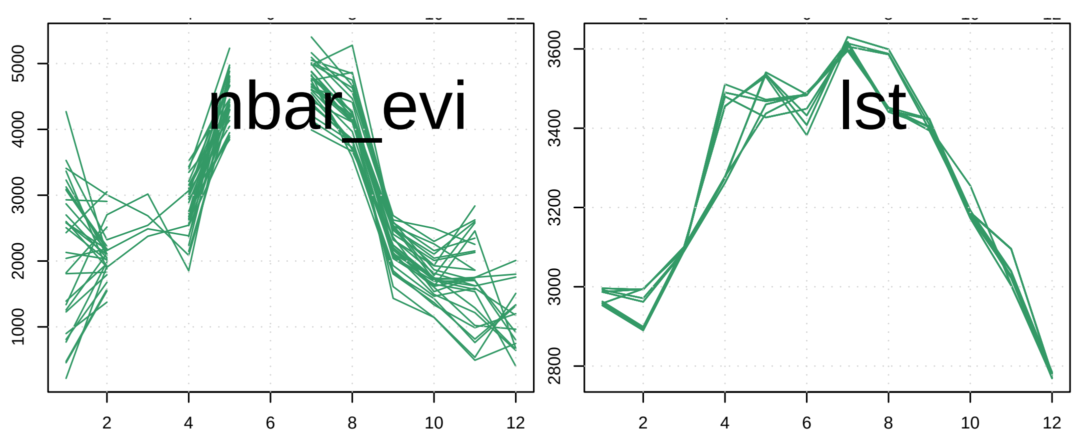
Mark Friedl (BU Geography)

Background

- **Global Change Analysis:** Understand the dynamics and the response of natural vegetation to variation in climate.
  - Both are smooth processes
  - Have very distinct seasonal/annual components.
  - Feedback mechanism.

Background

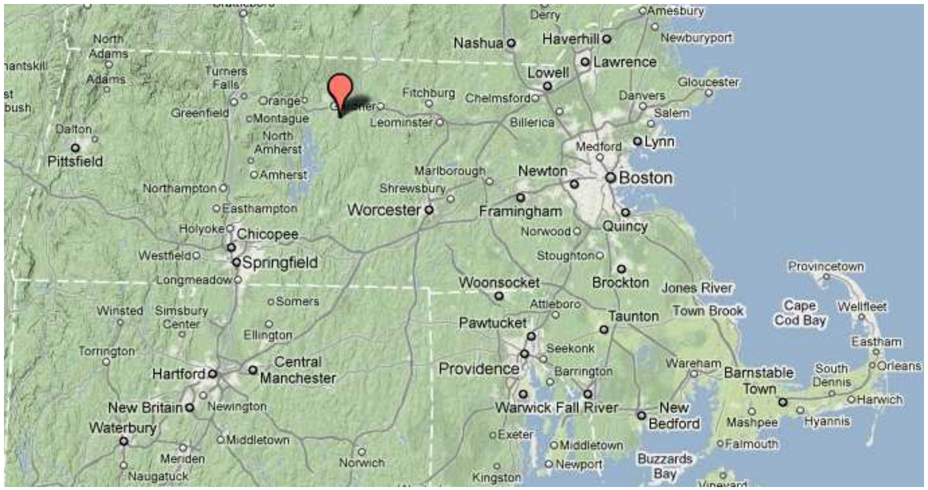
- **Global Change Analysis:** Understand the dynamics and the response of natural vegetation to variation in climate.
  - Both are smooth processes
  - Have very distinct seasonal/annual components.
  - Feedback mechanism.



Types of Data

- **Climate Data:** Rainfall/Precipitation, Land Surface Temperature, Sea Surface Temperature, ...
- **Ecosystem Data:** Vegetation Indices, ecosystem CO_2 exchange, evaporation, and energy flux,...
- **Disturbance events:** Fire, Logging, Insect infestation. (Point Process)

Initial Focus Area: Harvard Forest (far away from Harvard U)



- The google map image of Harvard Forest area in western Massachusetts.
- Some ground data on both climate and phenology available.
- We will mainly use remote sensing data
  - viable means to monitor geographically extensive patterns of coupled climate and vegetation processes

Data available from 2001 to now.

BOSTON UNIVERSITY

Sources of data

- **Ground Data: very limited** Measuring time series data is always challenging. Machine or Humans or both.
- **Remote Sensing of Global Ecosystems:**
- **Vegetation Index:** Obtained as a composite index from several BANDS of remote sensing data.
  - The Enhanced Vegetation Index(EVI) is an 'optimized' index designed to enhance the vegetation signal with improved sensitivity in high biomass regions. EVI is computed as

$$2.5 \times \frac{NIR - RED}{NIR + C_1 \times RED - C_2 \times Blue + L}$$

  - NIR and RED are recorded by MODIS (Moderate Resolution Imaging Spectroradiometer).
  - Other variations are NDVI and EVI2 (Both uses 2 bands)

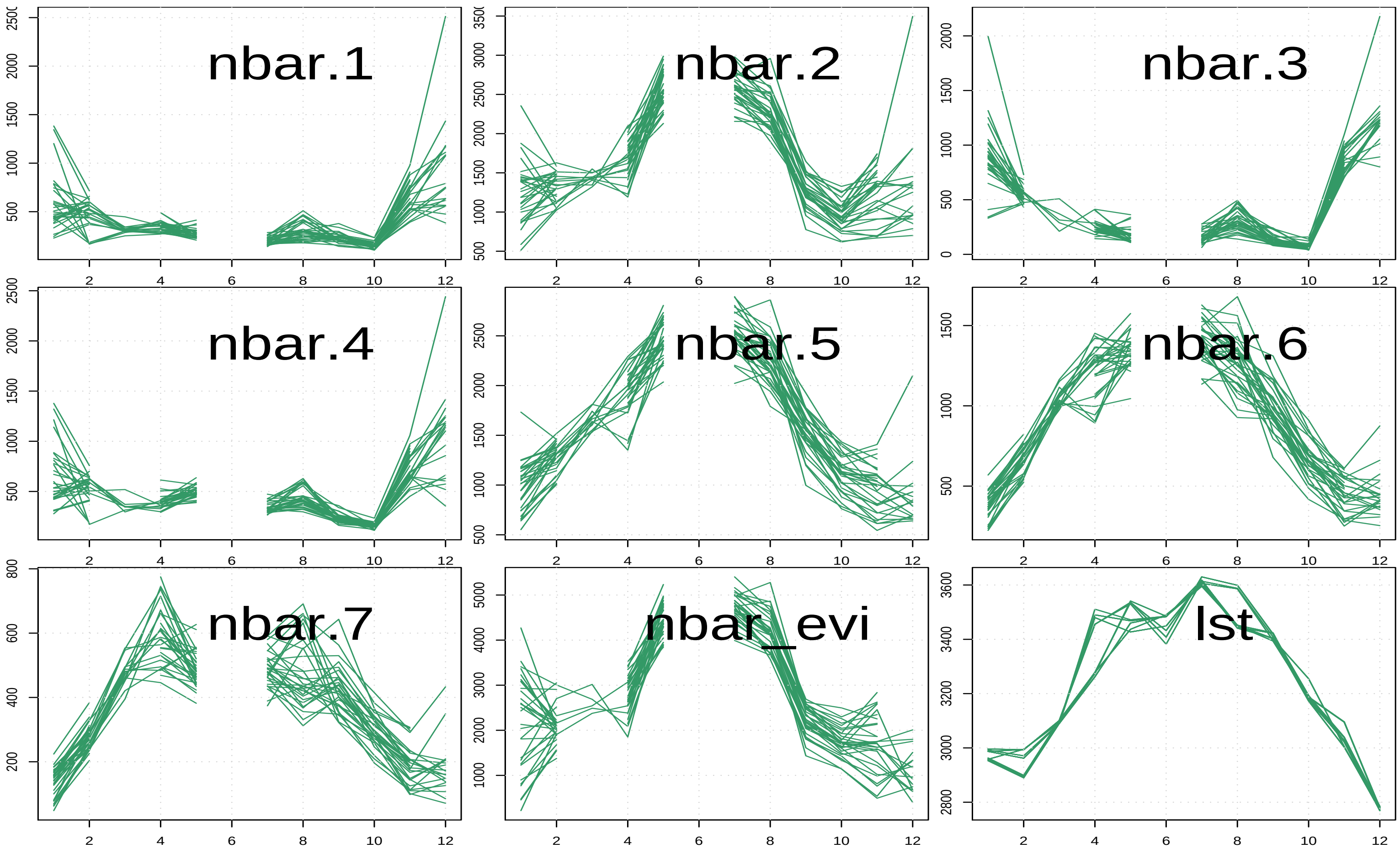
Which functional data to use: No consensus among geoscientist

Why not use the raw bands (multivariate functional data)

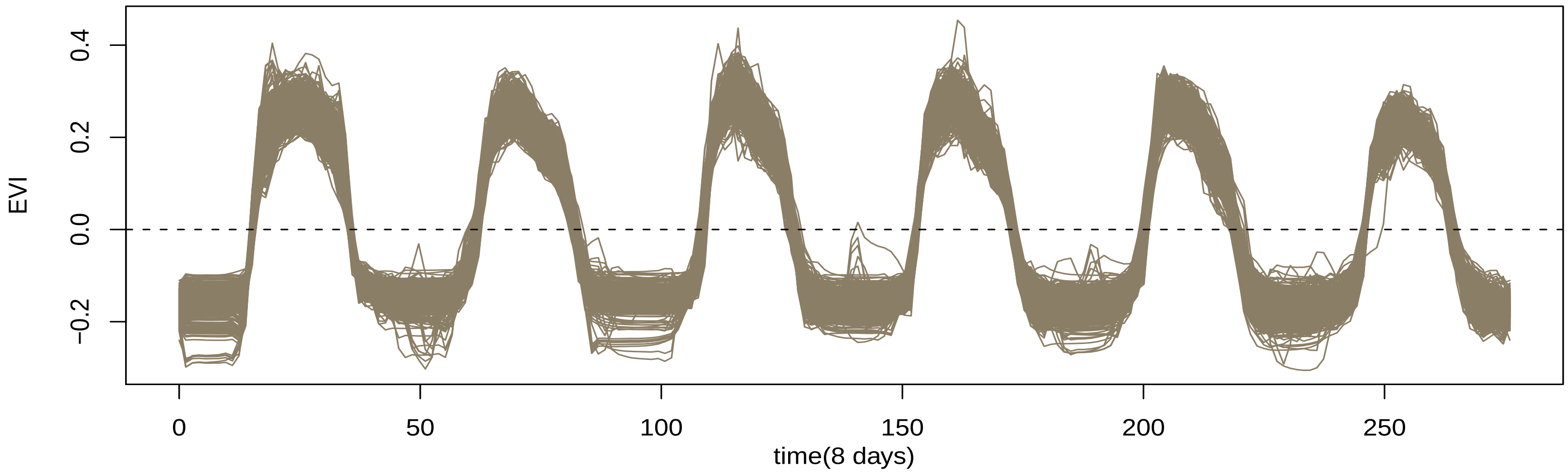
Statistical variable selection method might provide an answer

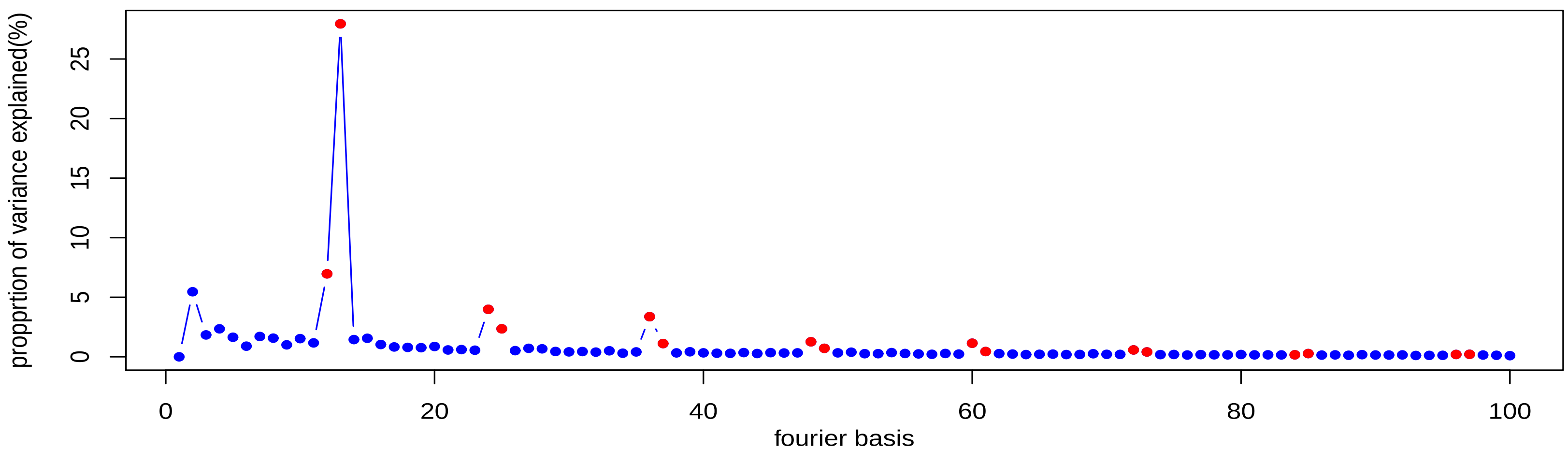
Remote Sensing Data for Harvard Forest

Several bands including derived EVI and LST for tile ID 1312 belonging to Landclass "Evergreen broadleaf forests"



Remote Sensing Data (MODIS) (Contd.)





BOSTON UNIVERSITY

(Geo) Science questions

Single variable Questions

- Can we decompose the variation within these curves and separate annual and extra-annual information for trend analysis.
- Are the timing events e.g. "start of season", "length of season", "start of browning" changing over the years
- Are tree lines moving northward?

Multiple variable Questions

- Are there changes in seasonal variation due to change in climate?
- Are the timing changing due to climate?
- Are tree lines moving northward due to change in climate?

Specific goals and their challenges: Factor Rotation

- One Goal is to decompose the variation within these curves and try to separate annual and extra-annual information for trend analysis.
- Standard functional or multivariate factor analysis is limited.
- One solution: Principal periodic functions.
- Ideally want to bring spatial correlation too.

Extraction of Timing Events: Registration

- Functional Data analysis provides several way of registration (Ramsay and Silverman)
- "zero crossing" of first and second derivatives provide phenologically interesting "markers"
- We wish to automate the process of aligning the timing event.
- Timing event rather than the registered curves are more important.

Challenges

- Automation requires numeric grid search in time (multiple zero crossings).
- Strong spatial dependence.
- No known model for using spatial dependence to uncover timing events or registration.

BOSTON UNIVERSITY

How does Climate change affect changes in Eco System: Functional Regression

Ecology 101

- Rather than the whole phenological curves scientists are often interested in "timing" event.
- Derivative and zero crossing of derivatives natural way of accommodating the above.
- Temperature preceding timing event more important than temperature on specific days
- Geoscientist do not have a single definition of Growing degree days- varies widely
- Relationship of climate and eco-system changes drastically by landclass.
- No standard model for accommodating disturbance events.

Functional Linear Modeling

- $\mathbf{x}(t)$ set of predictors.
- Concurrent Model

$$y(t) = \mu(t) + \beta(t)\mathbf{x}(t) + \epsilon_i(t)$$

- Model allowing a dependence on the recent past:

$$y_i(t) = \mu(t) + \int_{t-\delta}^{t} \beta(s - t + \delta)\mathbf{x}(t)ds + \epsilon_i(t).$$

Challenges for functional regression

- Choose among different models:
  - Concurrent
  - depending on recent past
  - point process (Growing Degree Days)
- Variable selection: Temperature, Rainfall.
- Spatial correlation.
- How to incorporate time specific "disturbance effects"?
- Clustering of functional regression
  - Current Landclass classification not at all perfect

Functional Data Analysis

- **Motivation from the data**

  1. Realizations of remote sensing data are essentially from continuous processes in time.
  2. Treat the vegetation growing process at pixel j, denoted by $f_j(t)$ as the jth realization of an underlying random process $f(t)$.
  3. Consider each $f_j(t)$ as one single 'data point'.

- **Advantages of functional data analysis**

  1. accommodate irregularities
  2. incorporate functional property: periodicity, monotonicity, etc.
  3. functional regression model: functional covariates and responses
  4. modeling derivatives and integrals: easy using the basis function representation

Getting Smoothed Curves

- Due to measurement error, we consider the additive error model for each realization.

$$y_{ji} = f_j(t_i) + \epsilon_j(t_i)$$

where y_{ji} is the ith observation from the jth curve.

- In order to get $f_j(t)$, we solve the following penalized problem

$$\arg\min_{f_j(t)} \sum_{i=1}^{N} (y_{ji} - f_j(t_i))^2 - \lambda \cdot \mathrm{PEN}[f_j(t)]$$

where $\mathrm{PEN}[f_j(t)] = \int f_j''(t)^2 \mathrm{d}t = \int (D^2[f_j(t)])^2 \mathrm{d}t$.

- Using basis expansion, we have

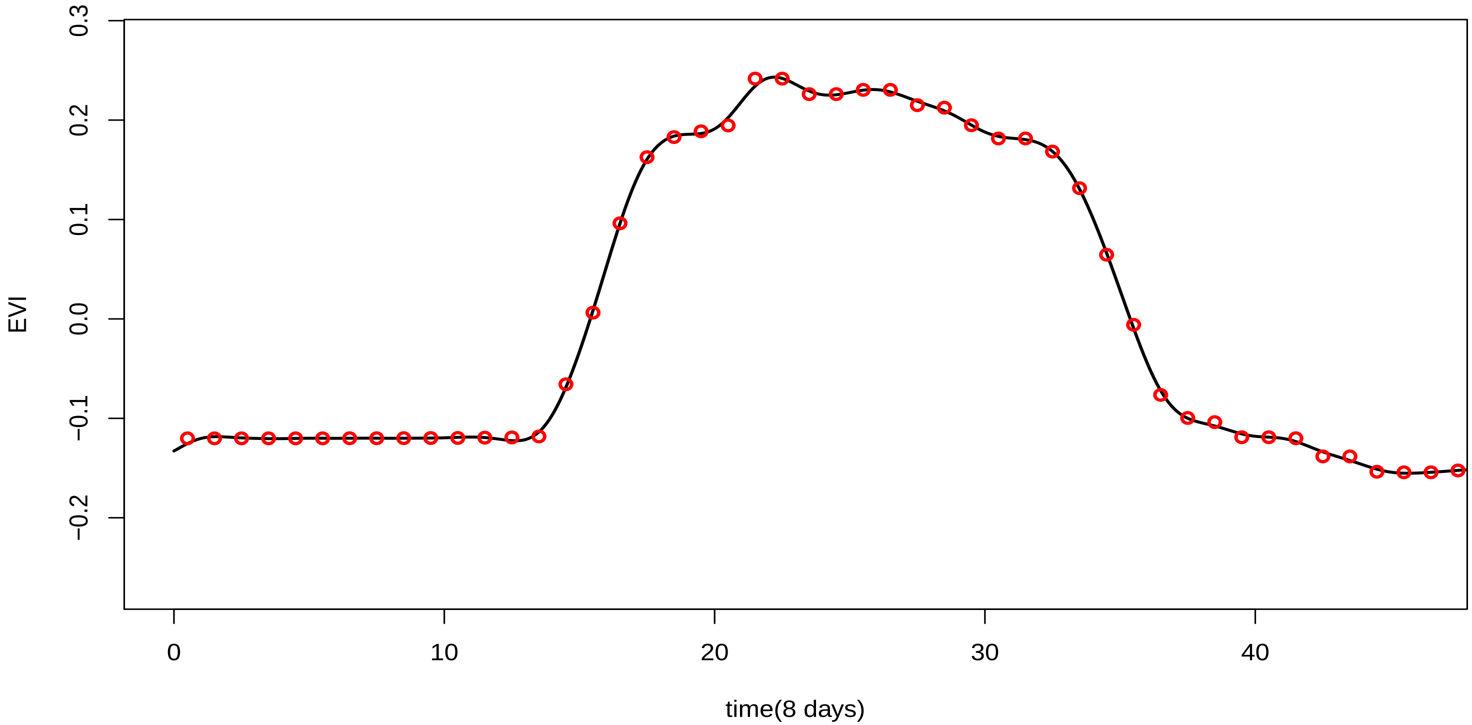
$$f_j(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}' \boldsymbol{\phi}$$

Getting Smoothed Curves (contd.)

In matrix form, we need to solve

$$\arg\min_{\mathbf{c}} (\mathbf{y} - \mathbf{\Phi c})^{\mathbf{T}} (\mathbf{y} - \mathbf{\Phi c}) - \lambda \mathbf{c}' \mathbf{R c}$$

$\mathbf{R}$ is the penalty matrix and $R_{kl} = \int D^2[\phi_k(t)] D^2[\phi_l(t)] \mathrm{d}t$.



Variation Decomposition

- Principal component analysis in the functional context provides a way to decompose the variation within these smoothed curves.
- Define total variation V_T as

$$V_T = \int \mathrm{var}(f(t))\mathrm{d}t$$

- Recall $f(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}'\boldsymbol{\phi}$.
- Under some regularity conditions, we know,

$$\begin{aligned} V_T &= \int \boldsymbol{\phi}^T(t)\left[\mathrm{cov}\left(\mathbf{c}\right)\right]\boldsymbol{\phi}(t)\mathrm{d}t = \mathcal{E}\left[(\mathbf{c}-\bar{\mathbf{c}})'\left(\int \boldsymbol{\phi}(t)\boldsymbol{\phi}^T(t)\mathrm{d}t\right)(\mathbf{c}-\bar{\mathbf{c}})\right] \\ &= \sum_{k=1}^{K} \mathrm{var}(c_k) \end{aligned}$$

- This holds for any orthonormal basis ϕ.

Functional Principal Component Analysis

We want to identify a weighting function $\xi_1(t)$ which maximizes

$$\mathrm{var}(\int \xi_1(t) f(t) \mathrm{d}t)$$

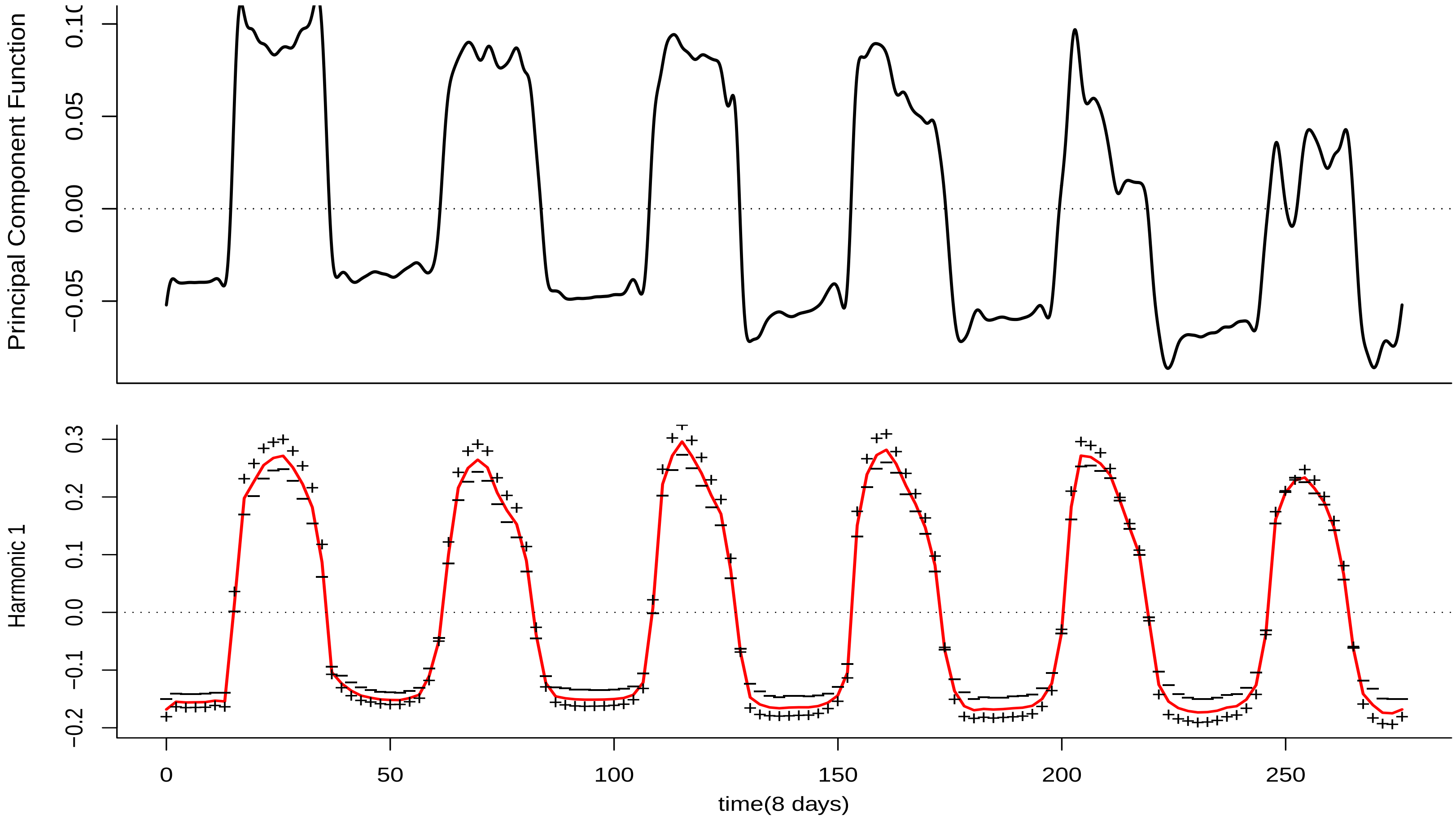
subject to $\int \xi_1(t)^2 \mathrm{d}t = 1$. Note $f(t)$ is random and $\xi_1(t)$ is fixed. Then, we proceed to find weighting function ξ_k such that

$$\xi_k(t) = \underset{\substack{\xi_k \perp \xi_1, \cdots, \xi_{k-1} \\ \|\xi_k\| = 1}}{\arg\max} \mathrm{var}(\int \xi_k(t) f(t) \mathrm{d}t)$$

some facts,

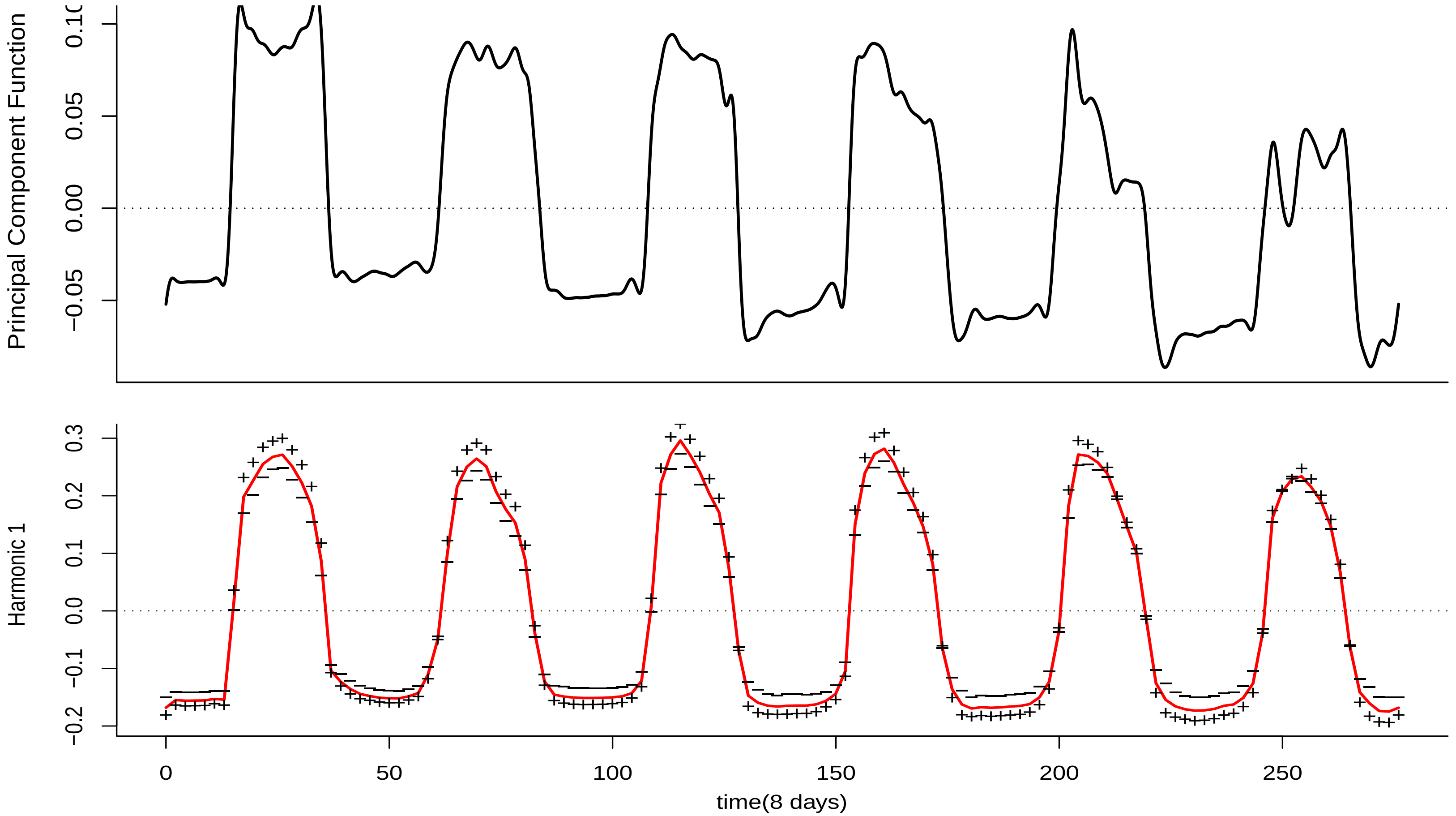
- λ_k is the kth eigenvalue of $\mathrm{cov}(\mathbf{c})$ and $\lambda_k = \mathrm{var}(\int \xi_k(t) f(t) \mathrm{d}t)$
- $\mathrm{V}(s,t) = \mathrm{cov}(f(s), f(t)) = \sum_{k=1}^{K} \lambda_k \xi_k(s) \xi_k(t)$
- $\sum_{k=1}^{K} \lambda_k = V_T$

Functional PCA Result



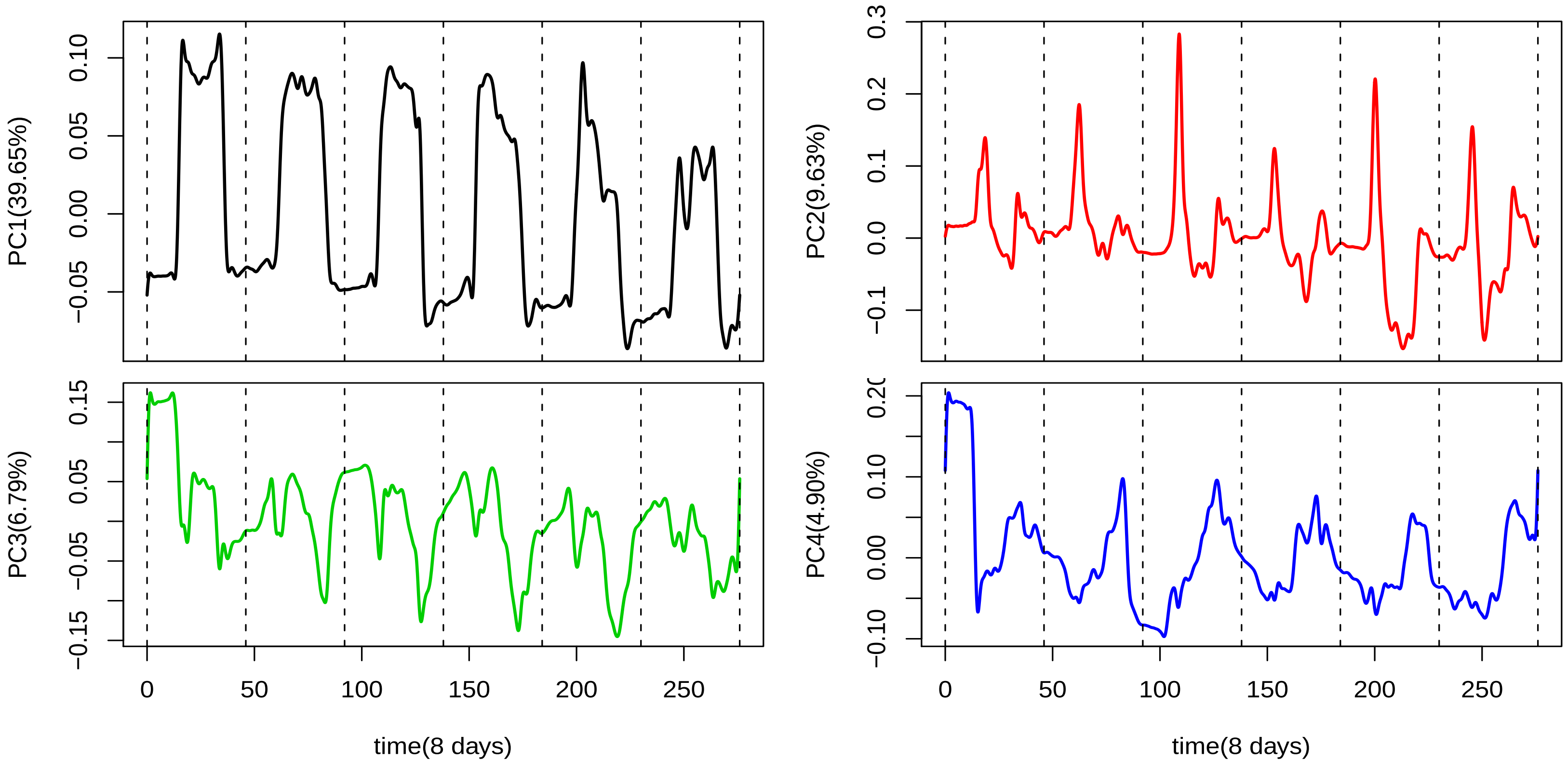
- The first PC function shows some level of periodicity.
- The first PC captures the variation during the summer and winter time.
- Higher summer index and lower winter index will receive higher PC score.

Functional PCA Result



- The first PC function shows some level of periodicity.
- The first PC captures the variation during the summer and winter time.
- Higher summer index and lower winter index will receive higher PC score.

Functional PCA result contd.



- All four PC functions suggest some level of periodicity.
- The second PC seems to capture the variation of spring feature.
- The third PC seems to capture the variation of fall feature.

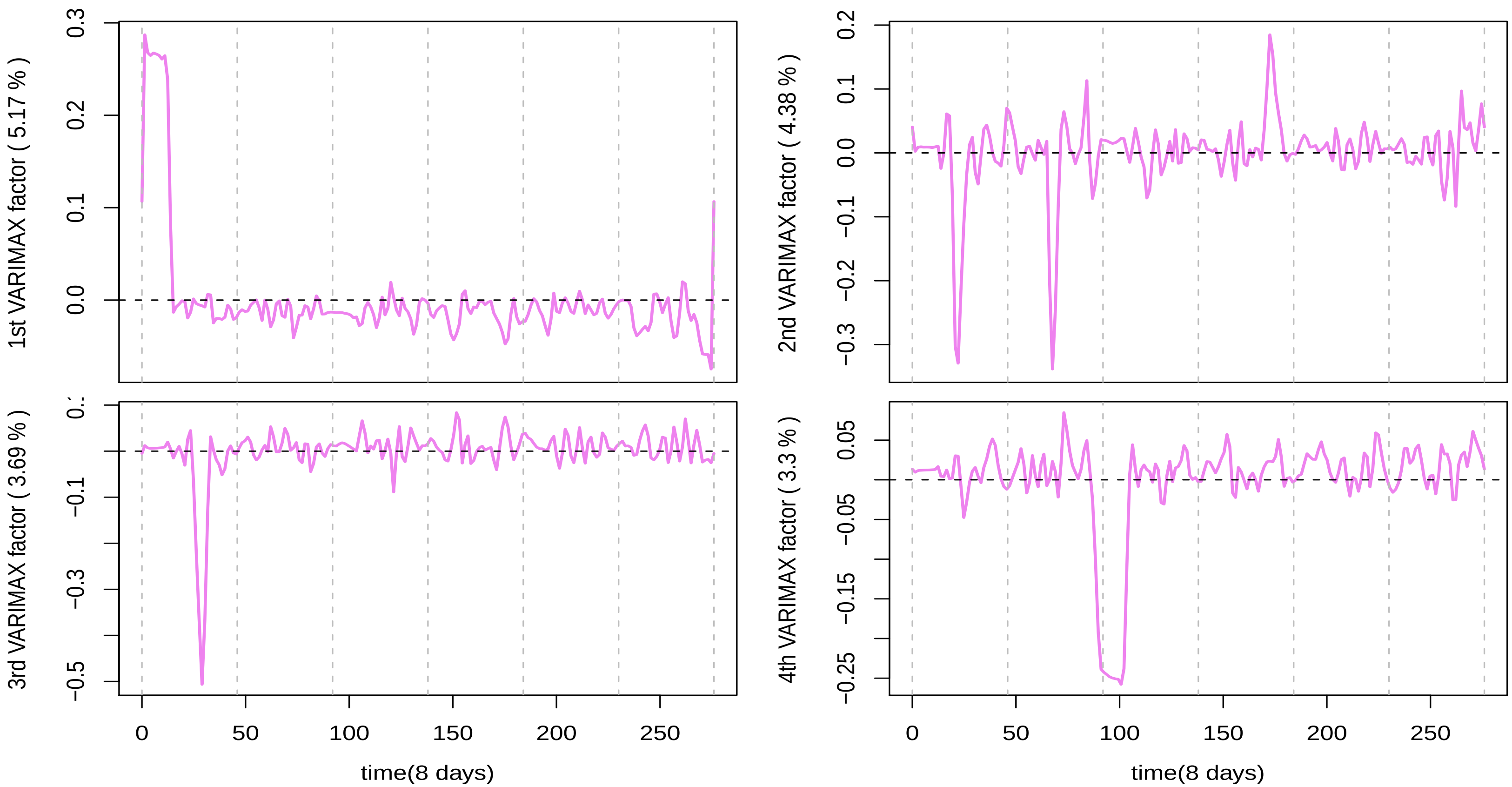
Principal Periodic Component Motivation

Our goal is to extract annual information from main signals contained in these curves.

- Original space contains noise.
- Major PC functions exclude noises and extract signals.
- The fourier bases with annual period are not fully embedded in the subspace S_m spanned by retained PC functions $\xi_1, \cdots, \xi_m$ for $m < 277$.
- Decompose S_m with a new basis $\eta_1, \cdots, \eta_m$ which can be ranked by their levels of periodicity.
- The VARIMAX rotation is not designed to capture periodicity feature.

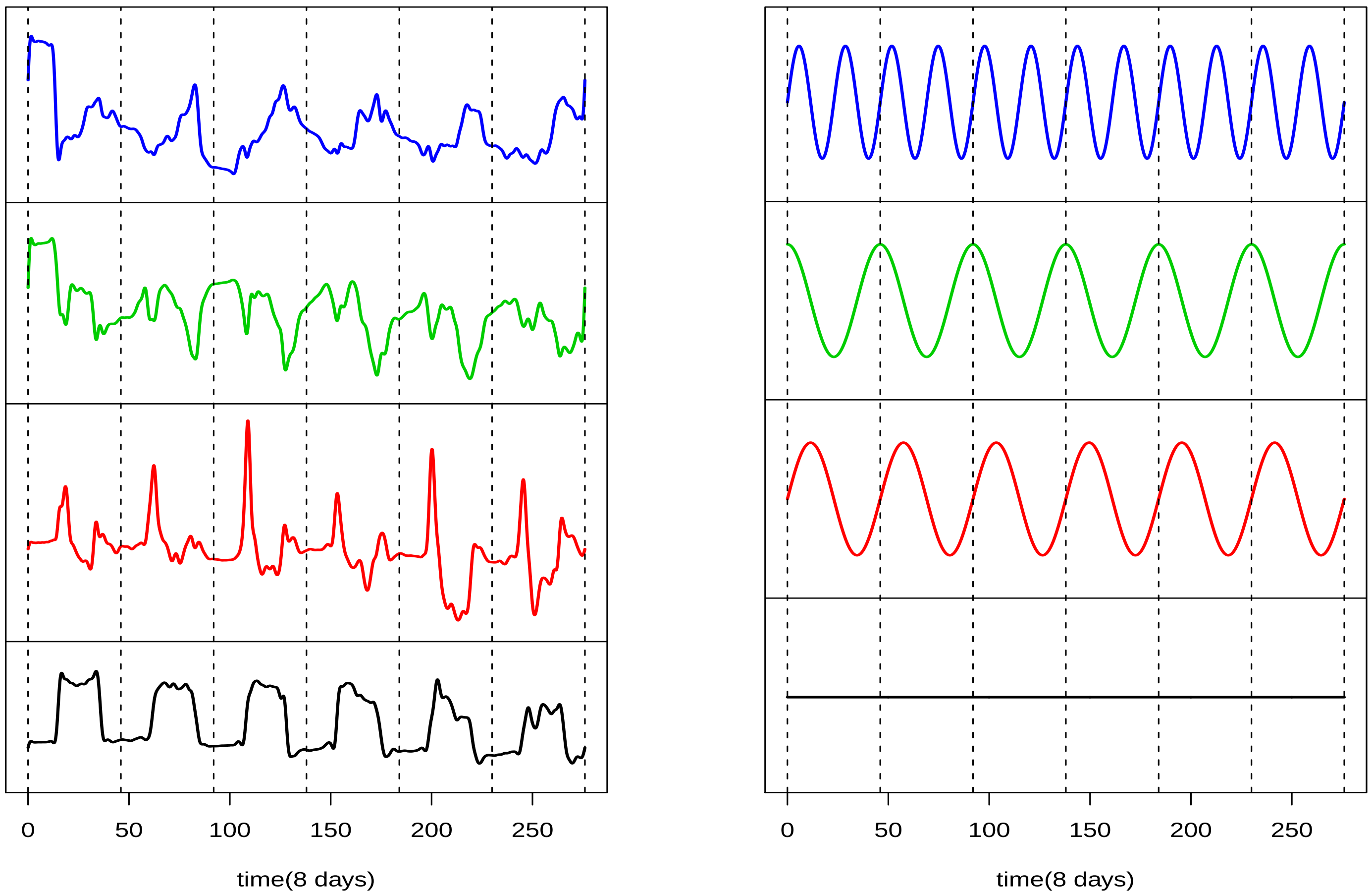
We propose to measure periodicity by standardized inner product between η and a potential benchmark function θ.

VARIMAX rotation result



- VARIMAX tends to make component functions spiky at various places

PPC Motivation (Contd.)



- We take linear combinations of functions from each sets of functions and measure their standardized inner products

PPC Formulation

- Suppose we have two vectors $\mathbf{a}_k$ and $\mathbf{b}_k$ of real value and let ξ_i be the ith retained PC function and ϕ_j be the jth annual fourier basis function.
- Define $\theta_k = \mathbf{b}_k^T \boldsymbol{\phi}$ and $\eta_k = \mathbf{a}_k^T \boldsymbol{\xi}$.
- Then, we want to solve the following problem,

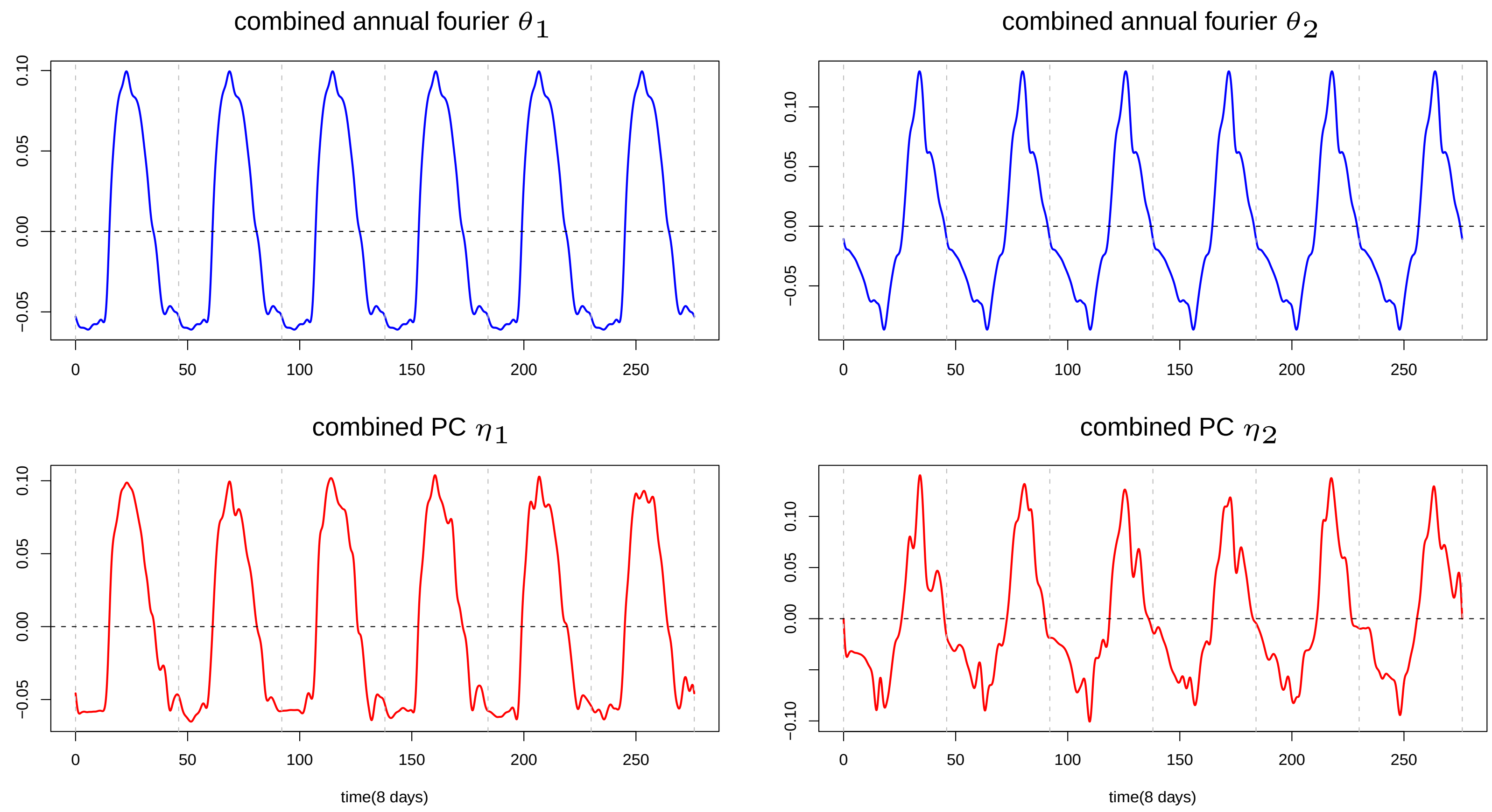
$$
\begin{aligned}
& \arg\max_{\mathbf{a_k},\mathbf{b_k}} \frac{< \mathbf{a}_k^T \boldsymbol{\xi}, \mathbf{b}_k^T \boldsymbol{\phi} >}{\|\mathbf{a}_k^T \boldsymbol{\xi}\| \cdot \|\mathbf{b}_k^T \boldsymbol{\phi}\|} \\
= \quad & \arg\max_{\mathbf{a_k},\mathbf{b_k}} \frac{< a_{1k}\xi_1 + \cdots + a_{mk}\xi_m, b_{1k}\phi_1 + \cdots + b_{nk}\phi_n >}{\|\mathbf{a}_k^T \boldsymbol{\xi}\| \cdot \|\mathbf{b}_k^T \boldsymbol{\phi}\|} \\
= \quad & \arg\max_{\mathbf{a_k},\mathbf{b_k}} \frac{\mathbf{a}_k^T \mathbf{V} \mathbf{b}_k}{\|\mathbf{a}_k^T \boldsymbol{\xi}\| \cdot \|\mathbf{b}_k^T \boldsymbol{\phi}\|}
\end{aligned}
$$

subject to

$$
\|\eta_k\| = 1, \; < \eta_k, \eta_l >= 0, \; \|\theta_k\| = 1, \; < \theta_k, \theta_l >= 0, \; < \eta_k, \theta_l >= 0
$$
$$
\forall k, l = 1, 2, \cdots, \min(m, n)
$$

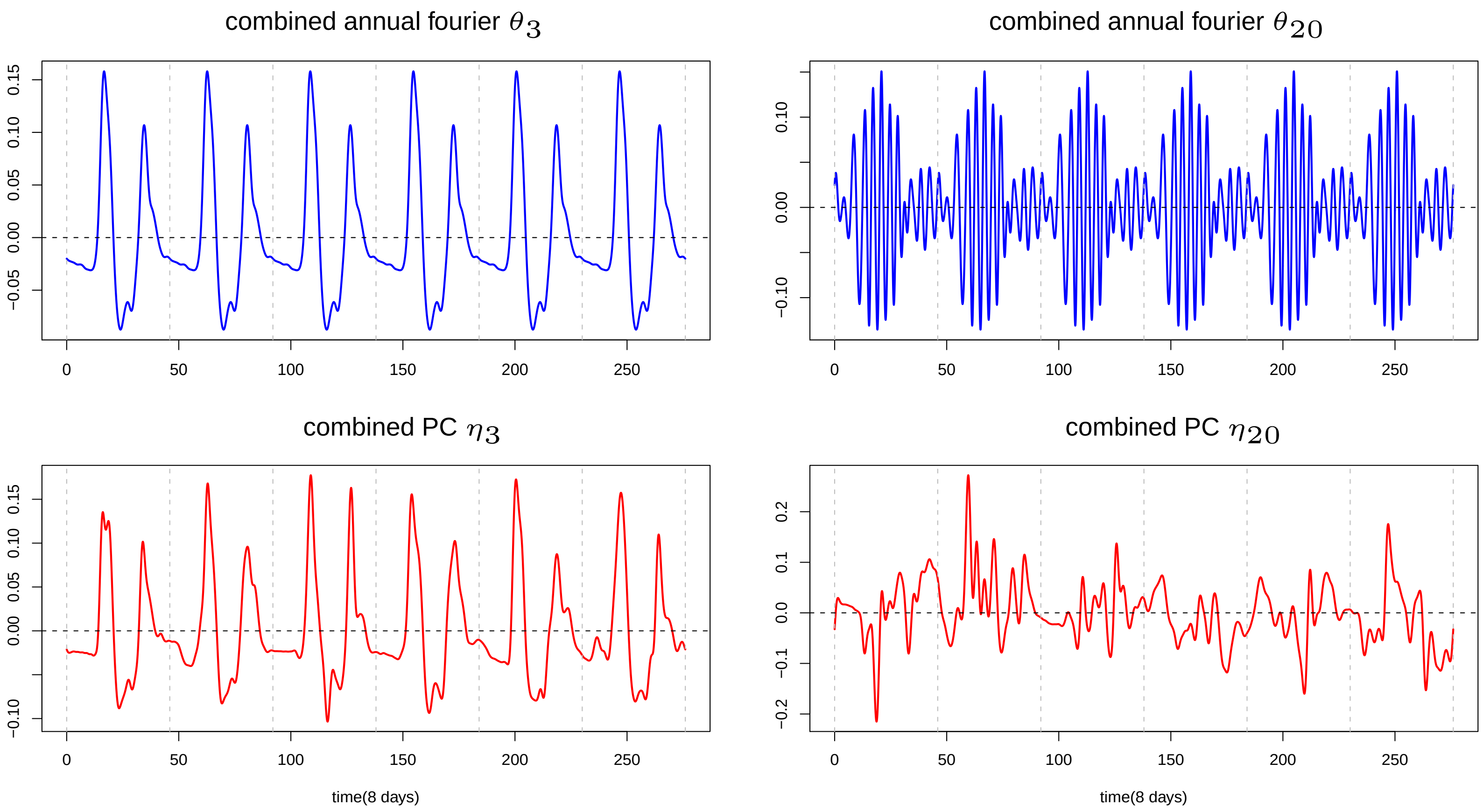
- same as the multivariate canonical correlation analysis objective

PPC Result



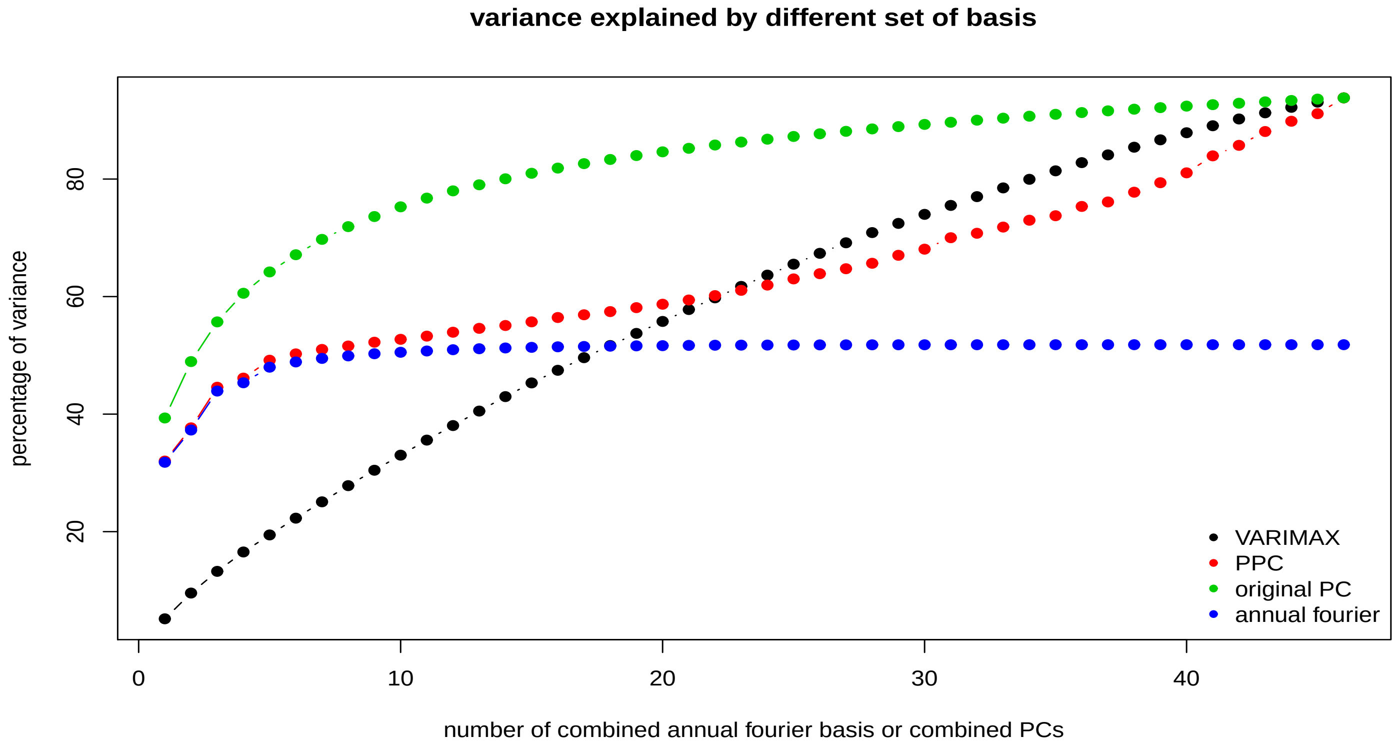
- The periodicity of the PPC functions is high

PPC Result Contd.



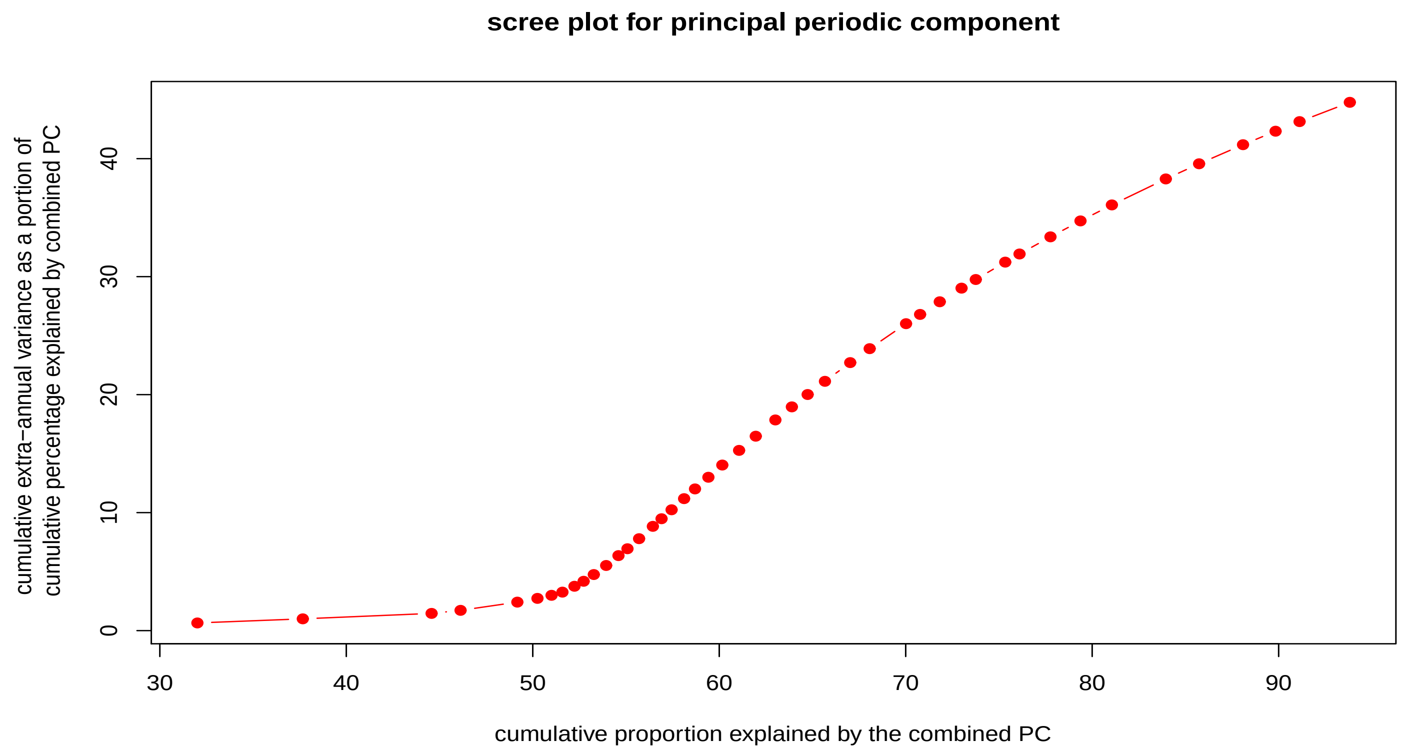
- The periodicity of the PPC function decreases

Variance Decomposition



- The rotated components $\{\eta_k\}_{k=1}^{m}$ drag the green curve down to the blue curve as much as possible.

Determine the Number of PPC



- The obvious kink in the plot suggests a place to cut off and separate annual from extra-annual signals.

Landclass Classification: Functional Clustering

- Functional Clustering

Acknowledgement:

- **NSF Award #0947950**
  NSF GK-12 Graduate STEM Fellows in K-12 Education GLACIER-Global Change Initiative-Education & Research
- **NSF Award #0934739–CMG:**
  Functional Data Modeling of Climate-Ecosystem Dynamics
- **NASA Carbon Cycle & Ecosystem Grant:**
  MODIS Algorithm Refinement and Earth Science Data Record Development for Global Land Cover and Land Cover Dynamics. NNX08AE61A