Functional Data Analysis

Lecture – 1

April 23, 2018

Motto

In God we trust, all others bring data

attributed to
William Edwards Deming* (1900-1993)

*American statistician, known, in particular, for promotion of statistical methods in industrial production and management

Outline

1 Organization

2 Assignments

3 Projects

4 Lecture–1

Instructor, webpage, etc.

Instructor, webpage, etc.

- Krzysztof (Krys) Podgórski and Sreekar Vadlamani

Instructor, webpage, etc.

- Krzysztof (Krys) Podgórski and Sreekar Vadlamani
- The easiest way to communicate is through e-mail: Krzysztof.Podgorski@stat.lu.se, Sreekar.Vadlamani@stat.lu.se

Instructor, webpage, etc.

- Krzysztof (Krys) Podgórski and Sreekar Vadlamani
- The easiest way to communicate is through e-mail: Krzysztof.Podgorski@stat.lu.se, Sreekar.Vadlamani@stat.lu.se
- Krys' webpage: https://krys.neocities.org

Instructor, webpage, etc.

- Krzysztof (Krys) Podgórski and Sreekar Vadlamani
- The easiest way to communicate is through e-mail: Krzysztof.Podgorski@stat.lu.se, Sreekar.Vadlamani@stat.lu.se
- Krys' webpage: https://krys.neocities.org
- The main source of the information about the course: https://krys.neocities.org/Teaching/FDA/FDA.html (webpage)
- Information will also be updated at Live@Lund
- Office hours each Wednesday between 11:00 and 12:00 or by appointment.

Course Organization

Course Organization

- Syllabus is available at the webpage (print it on your own, if a hard copy is more convenient for you).

Course Organization

- Syllabus is available at the webpage (print it on your own, if a hard copy is more convenient for you).
- Three parts of a big comprehensive “examination”:

Assignments, **Projects**, **Presentation**

Course Organization

- Syllabus is available at the webpage (print it on your own, if a hard copy is more convenient for you).
- Three parts of a big comprehensive "examination":

  Assignments, **Projects**, **Presentation**

- Assignments – Individually at home plus discussion in the classroom

Course Organization

- Syllabus is available at the webpage (print it on your own, if a hard copy is more convenient for you).
- Three parts of a big comprehensive "examination":

  Assignments, **Projects**, **Presentation**

- Assignments – Individually at home plus discussion in the classroom
- Computer Projects – In groups of two or, in exceptional cases, individually. Mostly done in the computer lab but if not completed then can be finished at home, and then a printed report showing that the tasks have been completed has to be submitted.

Course Organization

- Syllabus is available at the webpage (print it on your own, if a hard copy is more convenient for you).
- Three parts of a big comprehensive "examination":

  Assignments, **Projects**, **Presentation**

- Assignments – Individually at home plus discussion in the classroom
- Computer Projects – In groups of two or, in exceptional cases, individually. Mostly done in the computer lab but if not completed then can be finished at home, and then a printed report showing that the tasks have been completed has to be submitted.
- Presentation – 30min presentation of your own study based on one of the provided data sets (this will be discussed in detail when halfway through the course).

Grade

Grade

- For each of the three parts score will be assigned on the scale from 0-100

Grade

- For each of the three parts score will be assigned on the scale from 0-100
- This score contributes equally to the total score which is computed according to the formula:

$$T = (S_1 + S_2 + S_3)/3,$$

where S_1, S_2, S_3 represent scores for the corresponding parts.

Grade

- For each of the three parts score will be assigned on the scale from 0-100
- This score contributes equally to the total score which is computed according to the formula:

$$T = (S_1 + S_2 + S_3)/3,$$

where S_1, S_2, S_3 represent scores for the corresponding parts.

- The final grade will be assigned according to the table:

Percentage	**Grade**
49 - 0	F
54 - 50	E
64 - 55	D
74 - 65	C
84 - 75	B
100 - 85	A

Outline

1 Organization

2 Assignments

3 Projects

4 Lecture–1

Assignments – cover basics and main topics of the course

Assignments – cover basics and main topics of the course

- The total of **five assignments**, that are based on the covered course material although they do not cover the material completely.

Assignments – cover basics and main topics of the course

- The total of **five assignments**, that are based on the covered course material although they do not cover the material completely.
- They will comprise of a set of simple questions that will help to clarify introduced topics.

Assignments – cover basics and main topics of the course

- The total of **five assignments**, that are based on the covered course material although they do not cover the material completely.
- They will comprise of a set of simple questions that will help to clarify introduced topics.
- Some questions will also help in preparing to the lab sessions.

Assignments – cover basics and main topics of the course

- The total of **five assignments**, that are based on the covered course material although they do not cover the material completely.
- They will comprise of a set of simple questions that will help to clarify introduced topics.
- Some questions will also help in preparing to the lab sessions.
- Assignments will be worked out at home and **due the week following the date they are posted in our schedule**.

Assignments – cover basics and main topics of the course

- The total of **five assignments**, that are based on the covered course material although they do not cover the material completely.
- They will comprise of a set of simple questions that will help to clarify introduced topics.
- Some questions will also help in preparing to the lab sessions.
- Assignments will be worked out at home and **due the week following the date they are posted in our schedule**.
- They can be submitted in written or electronic form (can be send by e-mail).

Assignments – cover basics and main topics of the course

- The total of **five assignments**, that are based on the covered course material although they do not cover the material completely.
- They will comprise of a set of simple questions that will help to clarify introduced topics.
- Some questions will also help in preparing to the lab sessions.
- Assignments will be worked out at home and **due the week following the date they are posted in our schedule**.
- They can be submitted in written or electronic form (can be send by e-mail).
- They will be available at the webpage but since I can change them, please, download a copy only on the date at which they are listed in the schedule.

Course contents and the textbook

Functional Data Analysis with R and MATLAB by J. O. Ramsay, G. Hooker, and S. Graves; Springer; 2009.

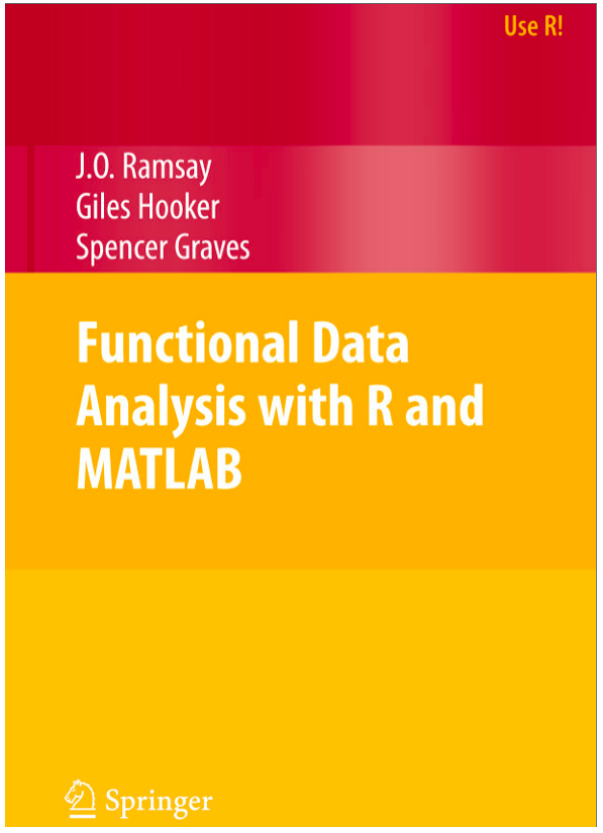
Course contents and the textbook

Functional Data Analysis with R and MATLAB by J. O. Ramsay, G. Hooker, and S. Graves; Springer; 2009.

- Provides a complete overview of the subject together with a healthy mix of examples.

Course contents and the textbook

Functional Data Analysis with R and MATLAB by J. O. Ramsay, G. Hooker, and S. Graves; Springer; 2009.



- Provides a complete overview of the subject together with a healthy mix of examples.
- Also provides help with `R` whenever needed.

Reference for `R` package on FDA:

`https://cran.r-project.org/web/packages/fda/index.html`

Access to the textbook

Access to the textbook

- Downloadable from the Springer Link.

Access to the textbook

- Downloadable from the Springer Link.
- Hard copy is also accessible at Springer.

Access to the textbook

- Downloadable from the Springer Link.
- Hard copy is also accessible at Springer.
- All the datasets included in the textbook come along with the `fda` package.

Outline

1 Organization

2 Assignments

3 Projects

4 Lecture–1

Lab projects – working with data

Lab projects – working with data

- Analysis of data using the methods discussed in the lecture using the statistical package `R`.

Lab projects – working with data

- Analysis of data using the methods discussed in the lecture using the statistical package `R`.
- At the moment the most popular mathematical/statistical softwares are **R** and **Python**, although commercial packages such as Matlab are also often utilized.

Lab projects – working with data

- Analysis of data using the methods discussed in the lecture using the statistical package `R`.
- At the moment the most popular mathematical/statistical softwares are **R** and **Python**, although commercial packages such as Matlab are also often utilized.
- `R`-package – free and very popular statistical package, very good for statistical computing although less compelling in handling large matrices and multivariate visualization

Lab projects – working with data

- Analysis of data using the methods discussed in the lecture using the statistical package `R`.
- At the moment the most popular mathematical/statistical softwares are **R** and **Python**, although commercial packages such as Matlab are also often utilized.
- `R`-package – free and very popular statistical package, very good for statistical computing although less compelling in handling large matrices and multivariate visualization
- We have opted for `R`-package to present analysis of the data and the methods of functional data analysis.

Lab projects – working with data

- Analysis of data using the methods discussed in the lecture using the statistical package R.
- At the moment the most popular mathematical/statistical softwares are **R** and **Python**, although commercial packages such as Matlab are also often utilized.
- R-package – free and very popular statistical package, very good for statistical computing although less compelling in handling large matrices and multivariate visualization
- We have opted for R-package to present analysis of the data and the methods of functional data analysis.
- The choice is largely dictated by the decision to stick to the textbook.

Downloading R-package and first steps

Downloading R-package and first steps

- Statistical R-package available for free download here

Downloading R-package and first steps

- Statistical R-package available for free download here
- Available on any PC platform (Mac, Windows, Linux).

Downloading R-package and first steps

- Statistical R-package available for free download here
- Available on any PC platform (Mac, Windows, Linux).
- Worry free and fast downloading procedure (a couple of minutes).

Downloading R-package and first steps

- Statistical R-package available for free download here
- Available on any PC platform (Mac, Windows, Linux).
- Worry free and fast downloading procedure (a couple of minutes).
- We will be working in the command line window of R (most direct way of accessing R-package).

Downloading R-package and first steps

- Statistical R-package available for free download here
- Available on any PC platform (Mac, Windows, Linux).
- Worry free and fast downloading procedure (a couple of minutes).
- We will be working in the command line window of R (most direct way of accessing R-package).
- No experience is required – all of the code that will be needed will be provided on our webpage!

Downloading R-package and first steps

- Statistical R-package available for free download here
- Available on any PC platform (Mac, Windows, Linux).
- Worry free and fast downloading procedure (a couple of minutes).
- We will be working in the command line window of R (most direct way of accessing R-package).
- No experience is required – all of the code that will be needed will be provided on our webpage!
- There some so-called R front-ends (such *R Commander* or *R-Studio* or *Jupyter*) that ease writing more complex programming in R – while you can use and utilize them, I assume **only a very basic R installation** with the primitive *copy-and-paste-to-the-command-line* approach as a method of running the programs.

Outline

1 Organization

2 Assignments

3 Projects

4 Lecture–1

Lecture – 1

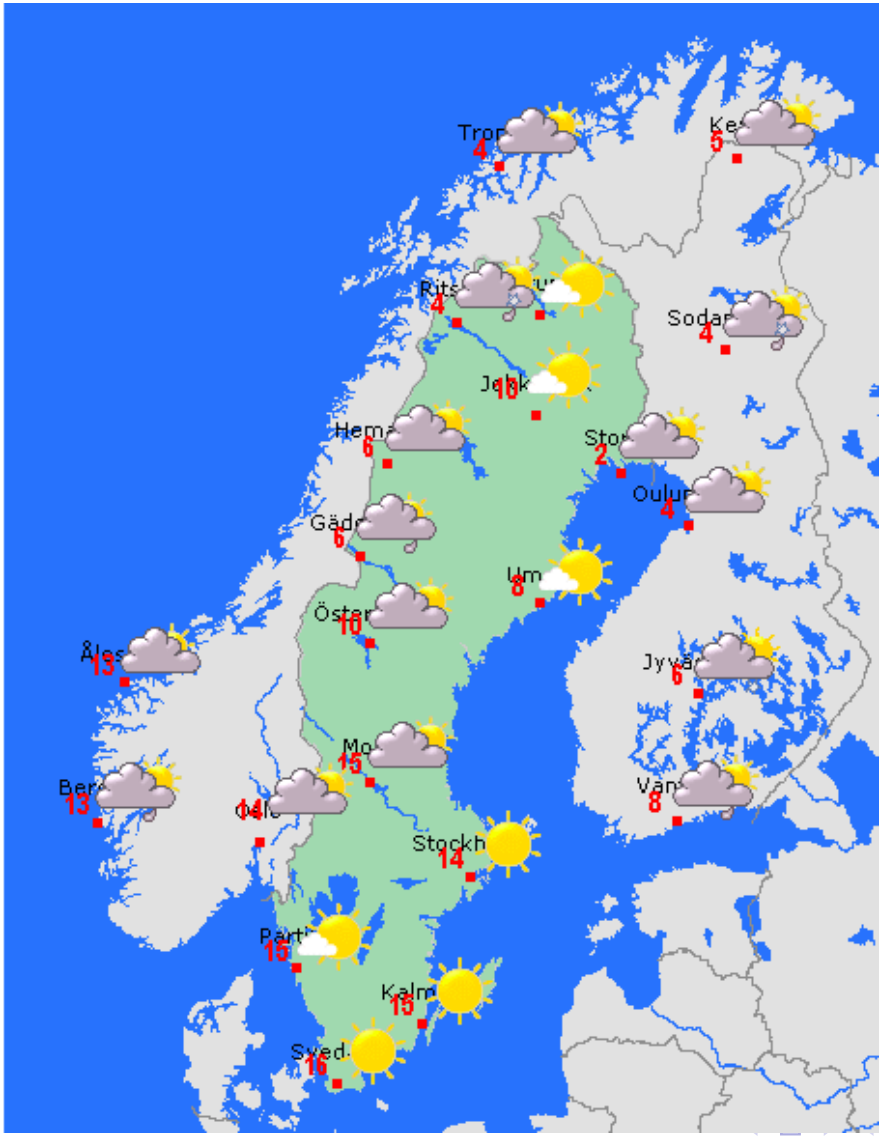
What are functional data?

"Data providing information about curves, surfaces or anything else varying over a continuum. In its most general form, under an FDA framework each sample element is considered to be a function. The physical continuum over which these functions are defined is often time, but may also be spatial location, wavelength, probability, etc."

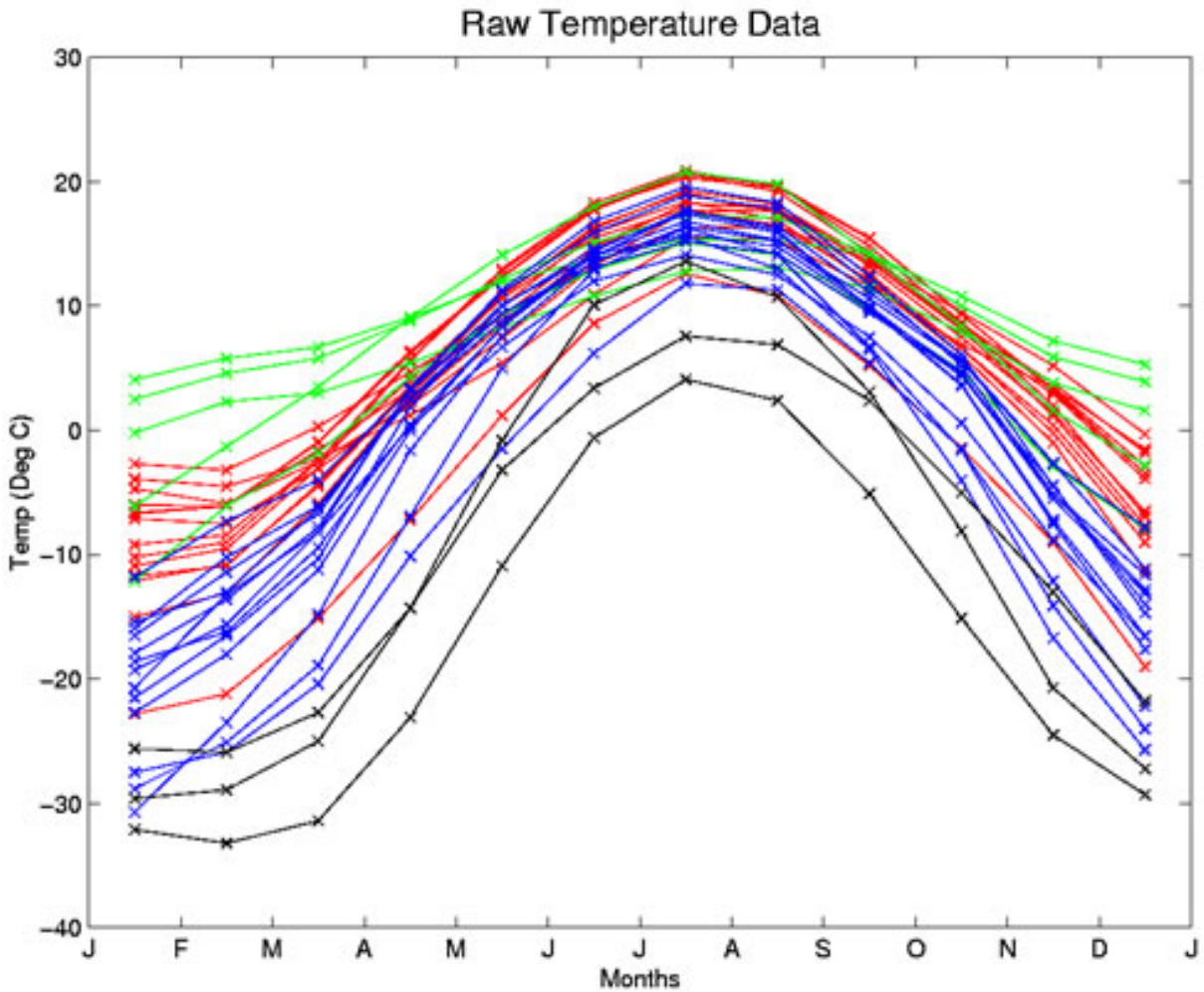
What are functional data?

"Data providing information about curves, surfaces or anything else varying over a continuum. In its most general form, under an FDA framework each sample element is considered to be a function. The physical continuum over which these functions are defined is often time, but may also be spatial location, wavelength, probability, etc." (Wikipedia)

Examples: Temperature over a Sweden– spatial



Examples: Temperature at a location over a given period of time



What questions can we ask of the data? Climatologists may have specific scientific hypotheses about temperature data such as these. For example:

What questions can we ask of the data? Climatologists may have specific scientific hypotheses about temperature data such as these. For example:

- What's the typical weather pattern for a city?

What questions can we ask of the data? Climatologists may have specific scientific hypotheses about temperature data such as these. For example:

- What's the typical weather pattern for a city?
- Which of the months have the most variable weather across the given period?

What questions can we ask of the data? Climatologists may have specific scientific hypotheses about temperature data such as these. For example:

- What's the typical weather pattern for a city?
- Which of the months have the most variable weather across the given period?
- Are the summer temperatures just a mirror image of the winter temperatures? That is, are the weather patterns balanced in the length of seasons and in their intensity? Or do some cities tend to have extreme temperatures that last for a long time?

What questions can we ask of the data? Climatologists may have specific scientific hypotheses about temperature data such as these. For example:

- What's the typical weather pattern for a city?
- Which of the months have the most variable weather across the given period?
- Are the summer temperatures just a mirror image of the winter temperatures? That is, are the weather patterns balanced in the length of seasons and in their intensity? Or do some cities tend to have extreme temperatures that last for a long time?
- How do the shapes of the weather patterns differ among the Pacific, Continental, Atlantic, and Arctic climates? How does the weather pattern in say, Lund, differ from the typical pattern in the rest of the Scandinavian?

Statisticians, on the other hand, will have different questions to ask of the data. For example:

Statisticians, on the other hand, will have different questions to ask of the data. For example:

- How can we represent the temperature pattern of a specific city over the entire year instead of just looking at the twelve discrete points? Should we just “connect the dots”, or is there a better way to do this?

Statisticians, on the other hand, will have different questions to ask of the data. For example:

- How can we represent the temperature pattern of a specific city over the entire year instead of just looking at the twelve discrete points? Should we just “connect the dots”, or is there a better way to do this?
- Do the summary statistics “mean” and “covariance” have any meaning when we’re dealing with curves?

Statisticians, on the other hand, will have different questions to ask of the data. For example:

- How can we represent the temperature pattern of a specific city over the entire year instead of just looking at the twelve discrete points? Should we just "connect the dots", or is there a better way to do this?
- Do the summary statistics "mean" and "covariance" have any meaning when we're dealing with curves?
- How can we determine the primary modes of variation in the data? How many typical modes can summarize these thirty-five curves?

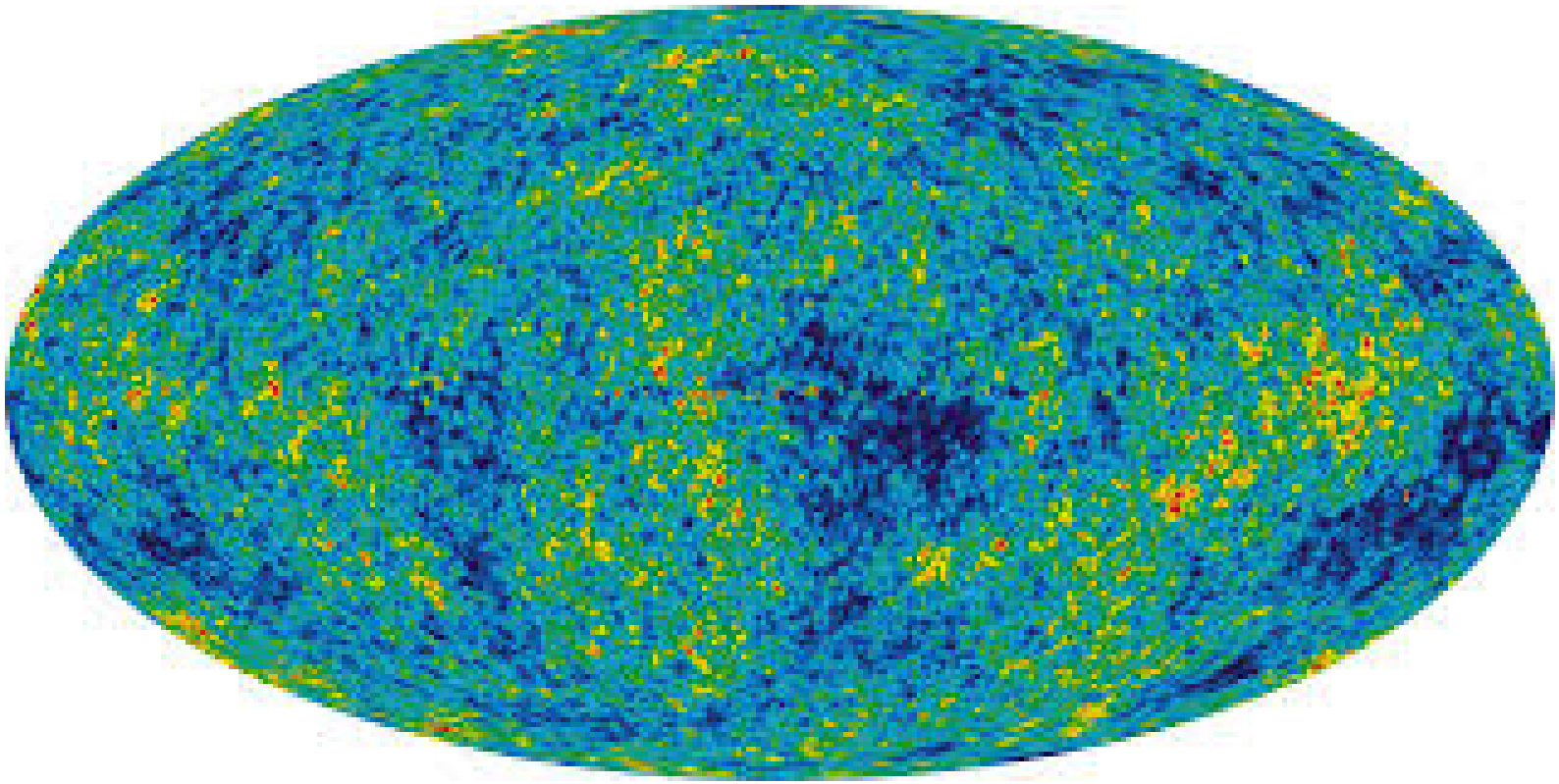
Statisticians, on the other hand, will have different questions to ask of the data. For example:

- How can we represent the temperature pattern of a specific city over the entire year instead of just looking at the twelve discrete points? Should we just "connect the dots", or is there a better way to do this?
- Do the summary statistics "mean" and "covariance" have any meaning when we're dealing with curves?
- How can we determine the primary modes of variation in the data? How many typical modes can summarize these thirty-five curves?
- Do these curves exhibit strictly sinusoidal behavior?

Statisticians, on the other hand, will have different questions to ask of the data. For example:

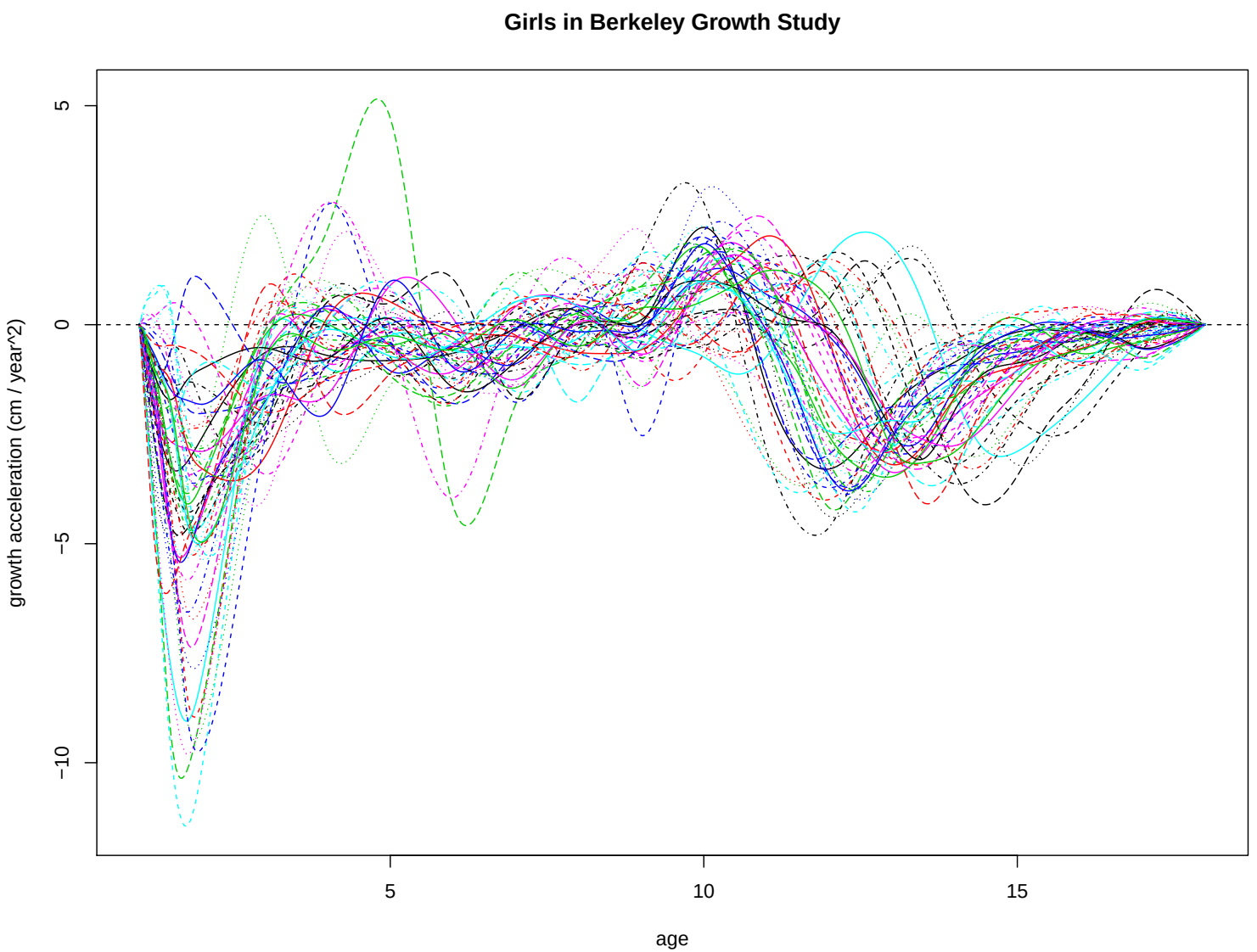
- How can we represent the temperature pattern of a specific city over the entire year instead of just looking at the twelve discrete points? Should we just "connect the dots", or is there a better way to do this?
- Do the summary statistics "mean" and "covariance" have any meaning when we're dealing with curves?
- How can we determine the primary modes of variation in the data? How many typical modes can summarize these thirty-five curves?
- Do these curves exhibit strictly sinusoidal behavior?
- Can we create an analysis of variance (ANOVA) or linear model with the curves as the response and the climate as the main effect?

Examples: Cosmic microwave background radiation

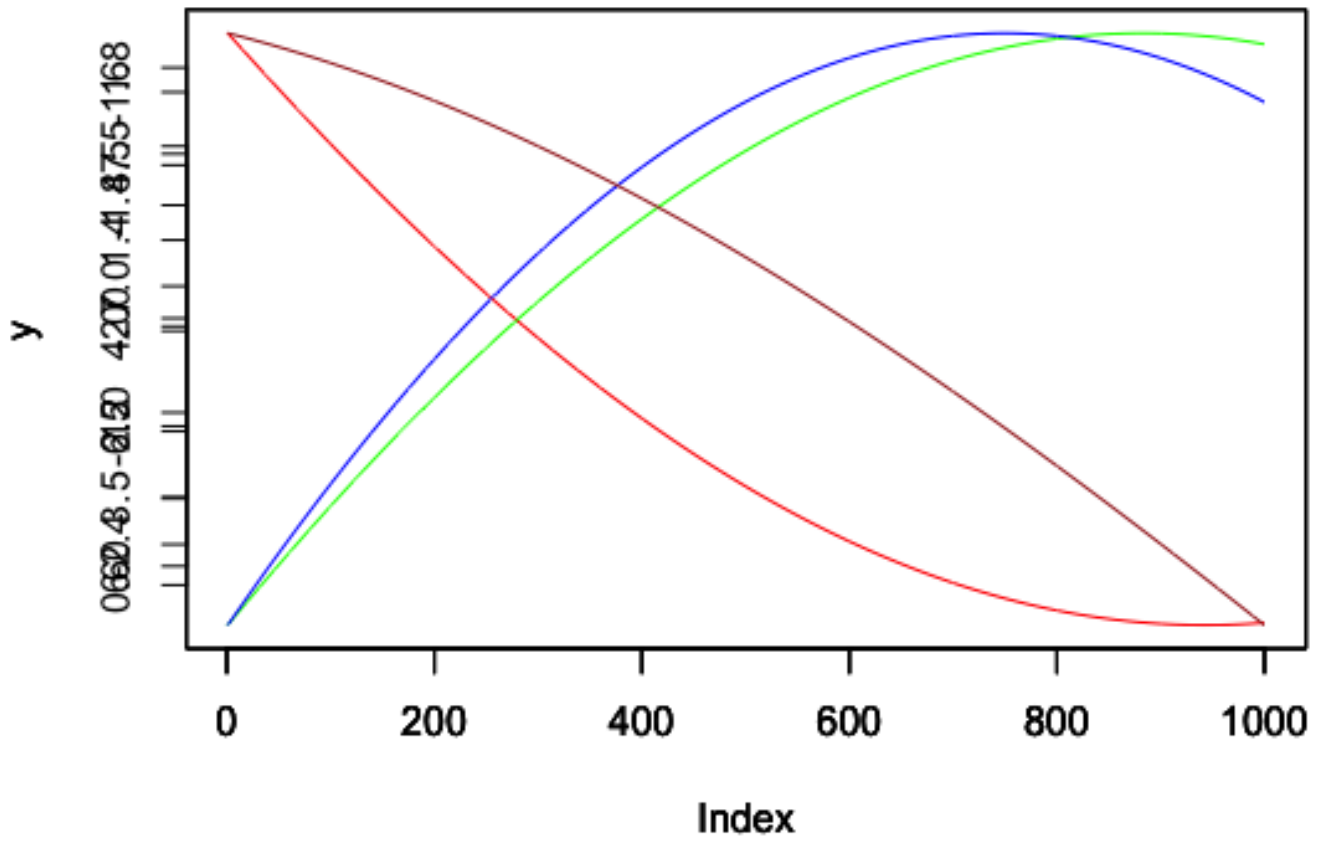


Interest is primarily modeling, and extracting features / patterns.

Examples: Growth data of girls



A fun example: finite dimensional functional data



A fun example: finite dimensional functional data

The figure corresponds to four realizations of

$$y(t) = \sum_{k=1}^{10} \xi_k \sin\left(\frac{\pi}{k} + t\right)$$

where ξ_k are i.i.d. $\mathcal{N}(0,1)$ random variables, and therefore, the real dimensionality of this data is 10 (far from continuum).

A fun example: finite dimensional functional data

The figure corresponds to four realizations of

$$y(t) = \sum_{k=1}^{10} \xi_k \sin\left(\frac{\pi}{k} + t\right)$$

where ξ_k are i.i.d. $\mathcal{N}(0,1)$ random variables, and therefore, the real dimensionality of this data is 10 (far from continuum).

NOTE: Brownian motion can be represented as an infinite series of the above kind. Thus, most of the physical processes we wish to model/analyse are infinite dimensional in nature.

Objective of FDA

Our interest is in...

- Representations of distribution of functions
  - mean
  - variation
  - covariation
- Relationships of functional data to
  - covariates
  - responses
  - other functions
- Relationships between derivatives of functions.
- Timing of events in functions.

... and the challenges are...

- Estimation of functional data from noisy, discrete observations.
- Numerical representation of infinite–dimensional objects
- Representation of variation in infinite dimensions.
- Description of statistical relationships between infinite dimensional objects.
- In case the covariates are larger than the observations? (regularisation and smoothness)
- Measures of variation and confidence in estimates.

Concluding thought

We are drowning in information and starving for knowledge

Rutherford D. Roger*

*American librarian (Yale University)