R 中的函数型数据展示

邱怡轩[1]　魏太云　熊熹

[1] 统计之都

The 4th Chinese R Conference(Shanghai), 2011

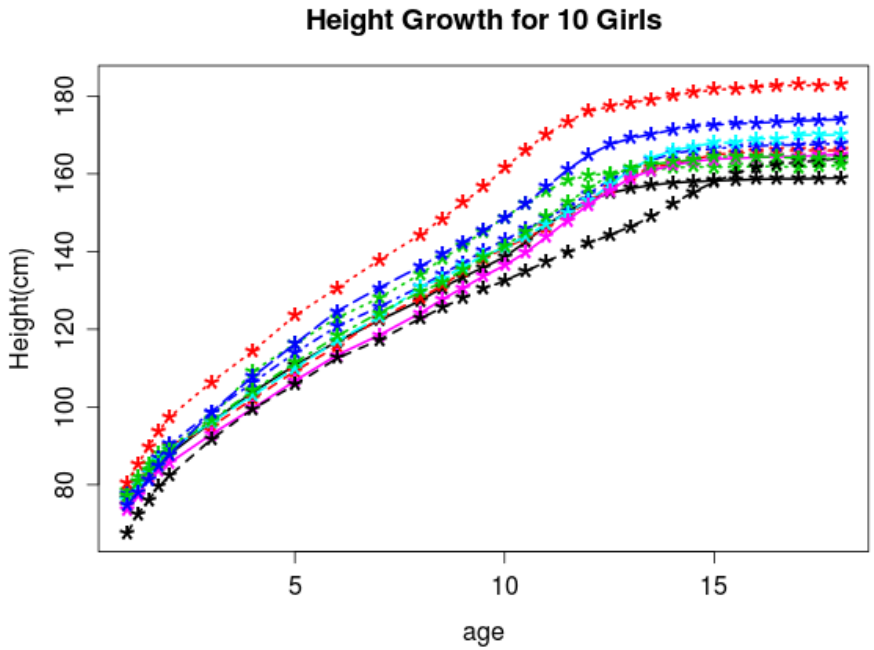
Outline

Outline

1. 函数型数据简介
   - 什么是函数型数据
   - 一般处理步骤
2. 函数型数据可视化
   - 曲线排齐
   - 数据模式的展示
   - 盒状图（boxplot）
   - 袋状图（bagplot）
   - 奇异值分解彩虹图（SVD rainbow plot）
   - 异常值展示

Functional data is multivariate data with an ordering on the dimensions. (Müller, (2006))

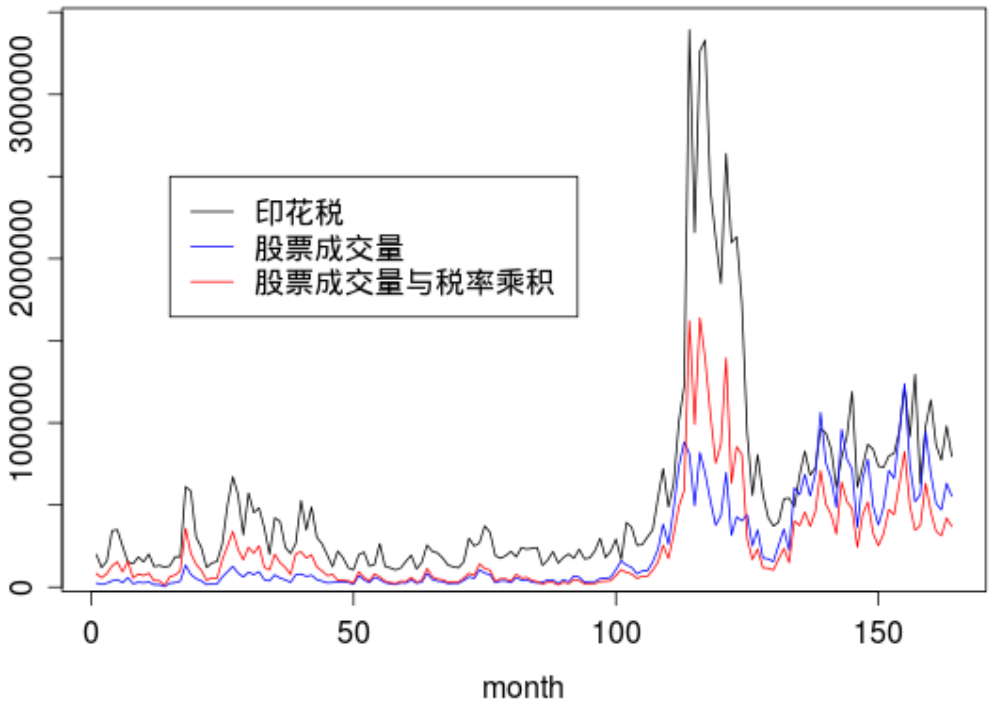
函数型数据意味着一个变量是多个有序观测（一般以时间作为顺序）的集合，每一次观测也可能是多个维度的。由于在建模中可以把一个序列的观测转化为一个函数表示，故得此名。人文社会科学中常用的面板数据（panal data）是函数型数据的特例。同样还有生物中所谓的纵向数据（Longitudinal Data）



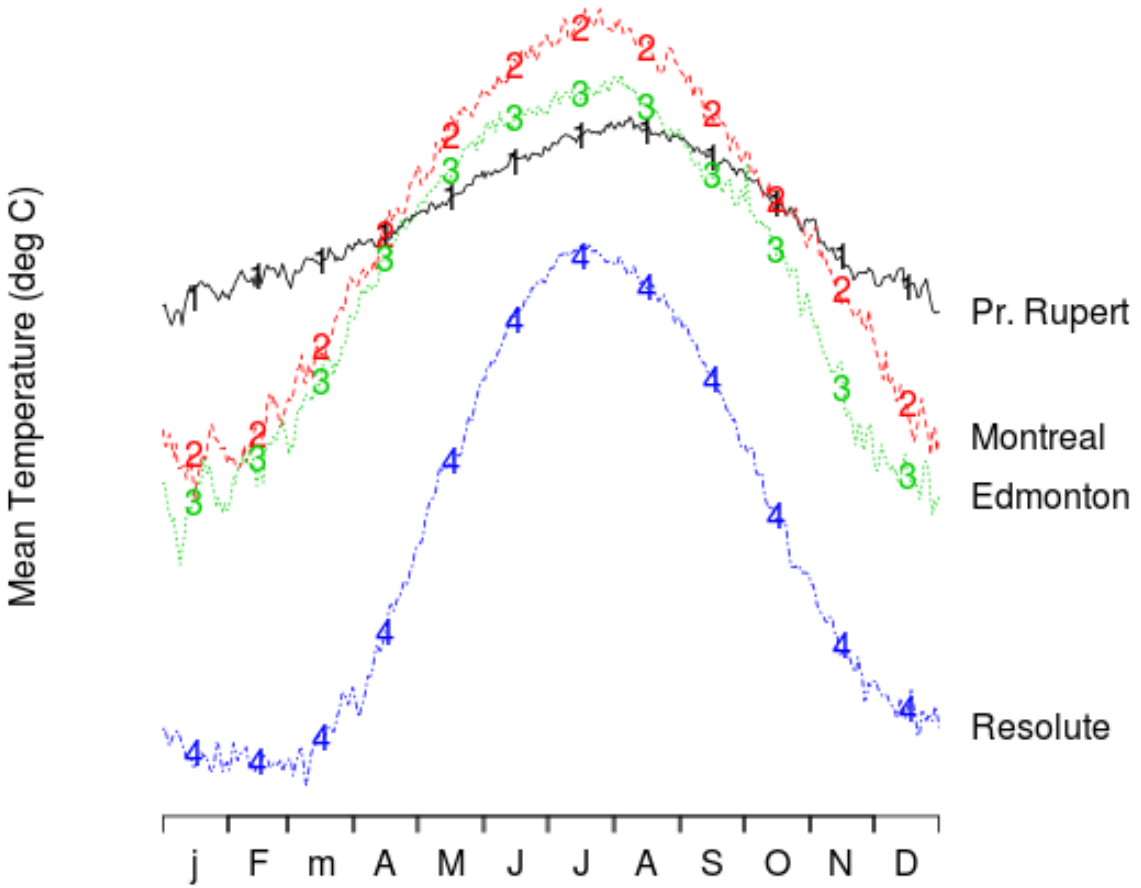
数据特性

- Quantity.
- Smooth, but may complex processes.
- Repeated observations with Similarity .
- Multiple dimensions
- High-frequency

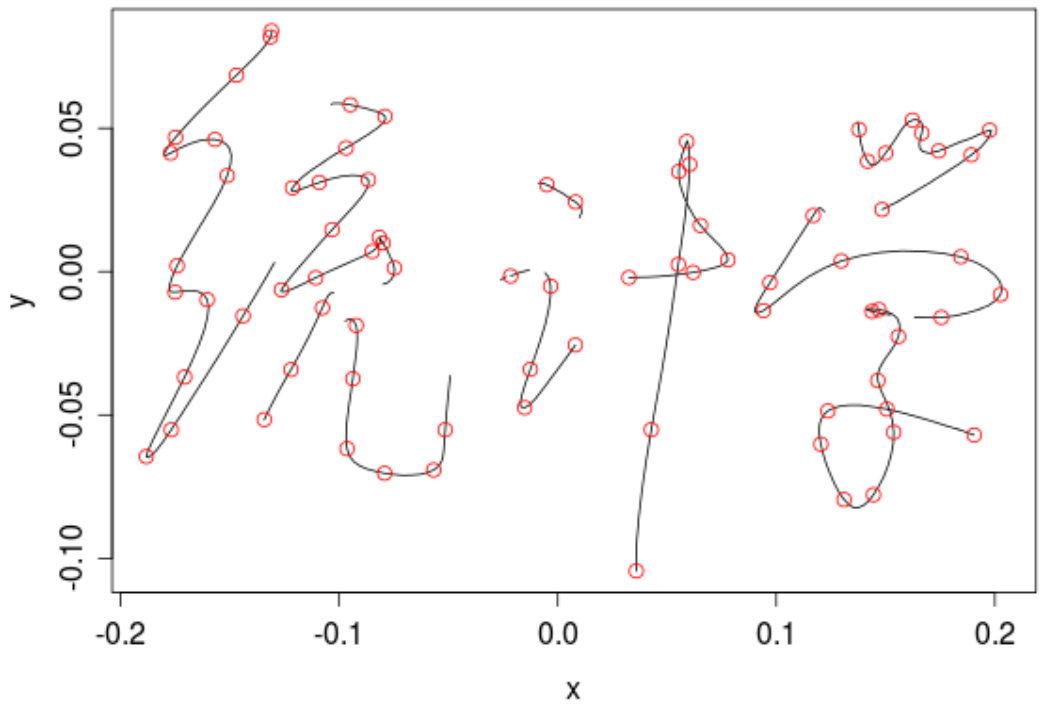
More examples.



More examples.



More examples.



Outline

1. 函数型数据简介
   - 什么是函数型数据
   - 一般处理步骤
2. 函数型数据可视化
   - 曲线排齐
   - 数据模式的展示
   - 盒状图（boxplot）
   - 袋状图（bagplot）
   - 奇异值分解彩虹图（SVD rainbow plot）
   - 异常值展示

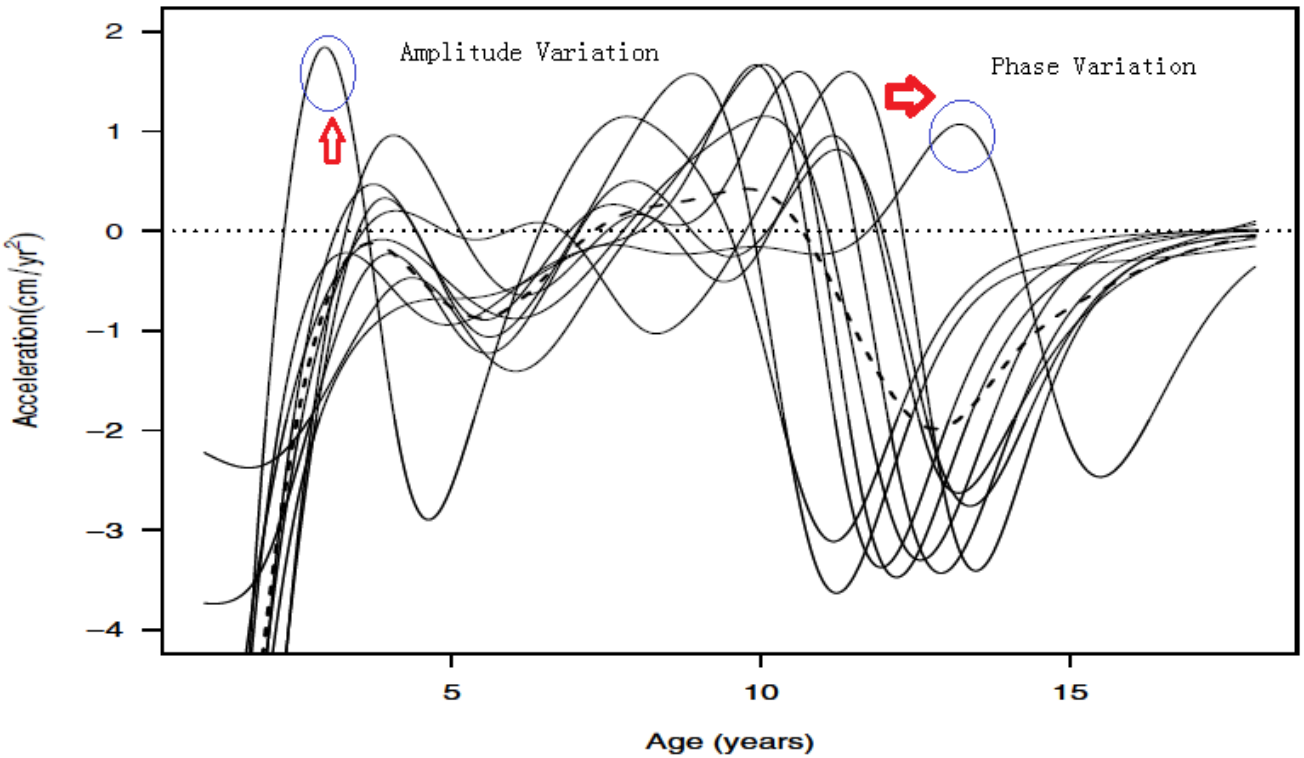
What do we do with functional data

- 从离散型数据点转化为函数
  - 基函数展开（插值 Interpolation）
  - 平滑处理（罚方法）
- 曲线排齐（Registration）
- 描述与探索性分析
  - Functional PCA
  - Functional Clusters
  - Functional Differential Analysis

Outline

1. 函数型数据简介
   - 什么是函数型数据
   - 一般处理步骤
2. 函数型数据可视化
   - 曲线排齐
   - 数据模式的展示
   - 盒状图（boxplot）
   - 袋状图（bagplot）
   - 奇异值分解彩虹图（SVD rainbow plot）
   - 异常值展示

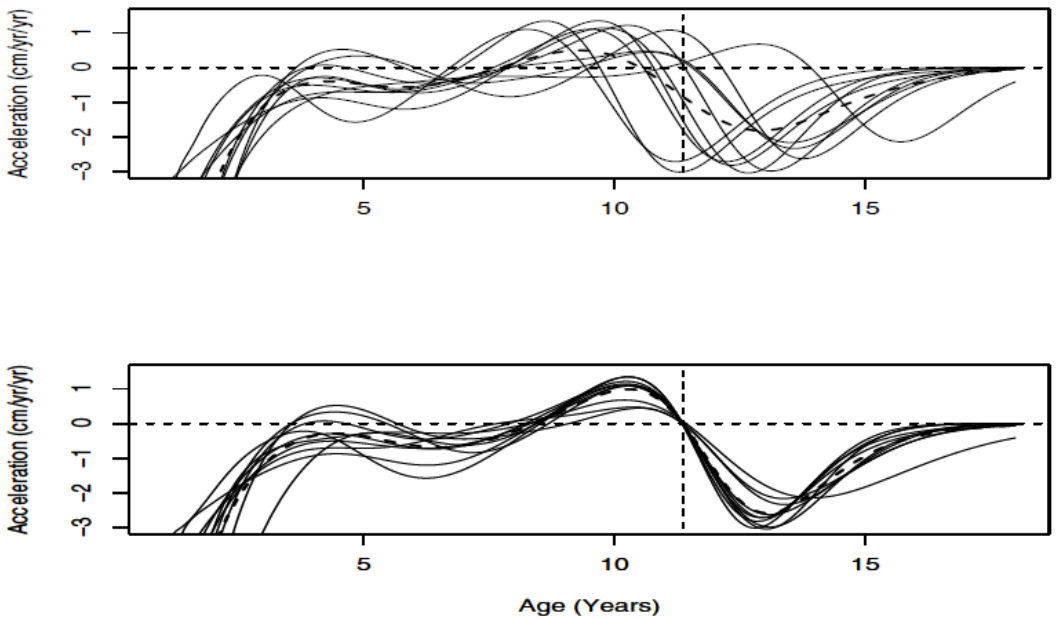
函数型观测值的变化既含有振幅 (Amplitude) 的变化, 也含有位相 (Phase) 的变化。每个函数的振幅和位相可能都不一样。如果将两者混在一起, 会导致许多的问题。如下例中，每个人的青春期增长突增出现的时间都不一样。



里程碑排齐

为了使跨截面曲线的显著特征都在大体相同的变量值 t 处显现出来，需要对曲线进行一定变换，有移位排齐、里程碑排齐以及更一般化的排齐。

下图使用里程碑排齐，使得所有观测的女孩儿的发育速度二次导数为 0（青春期生长突增）的时间一致。这样我们就可以得到一个更有代表性也更符合实际情况的平均发育趋势图，至少在我们的例子中如此。

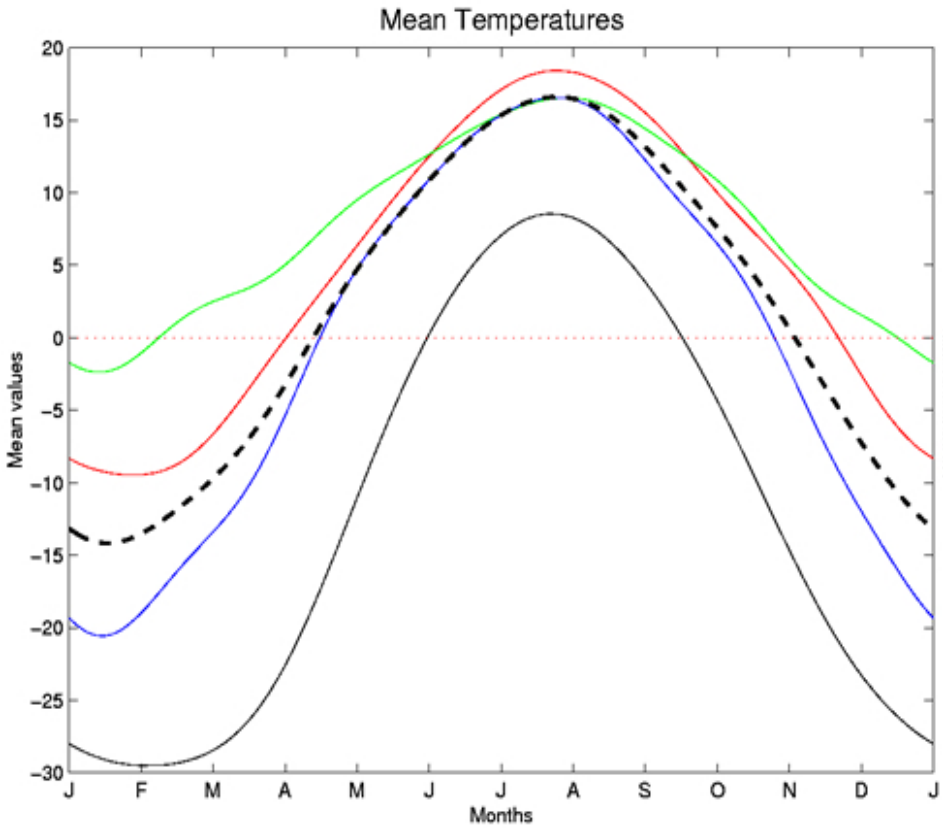


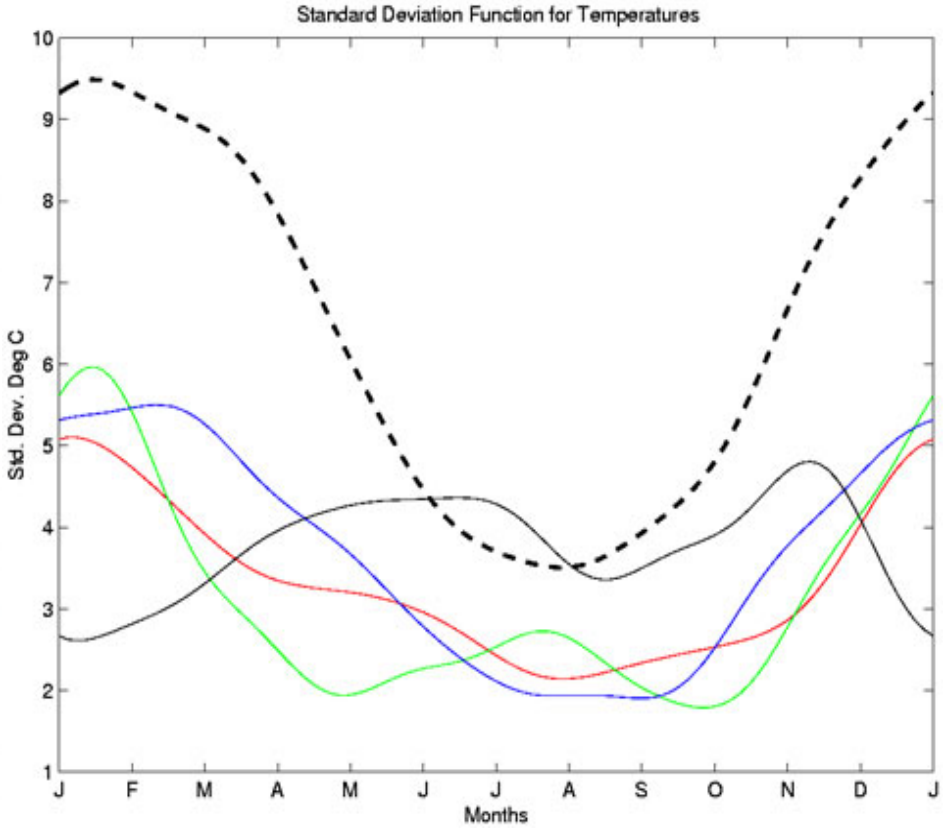
Outline

1. 函数型数据简介
   - 什么是函数型数据
   - 一般处理步骤
2. 函数型数据可视化
   - 曲线排齐
   - 数据模式的展示
   - 盒状图（boxplot）
   - 袋状图（bagplot）
   - 奇异值分解彩虹图（SVD rainbow plot）
   - 异常值展示

均值与方差图

函数型数据的均值可以直接从传统的一维数据的均值中直接化用过来，如果观测是同时的，则每个时间点的不同观测取均值，再绘制曲线即可，同理可得方差图。



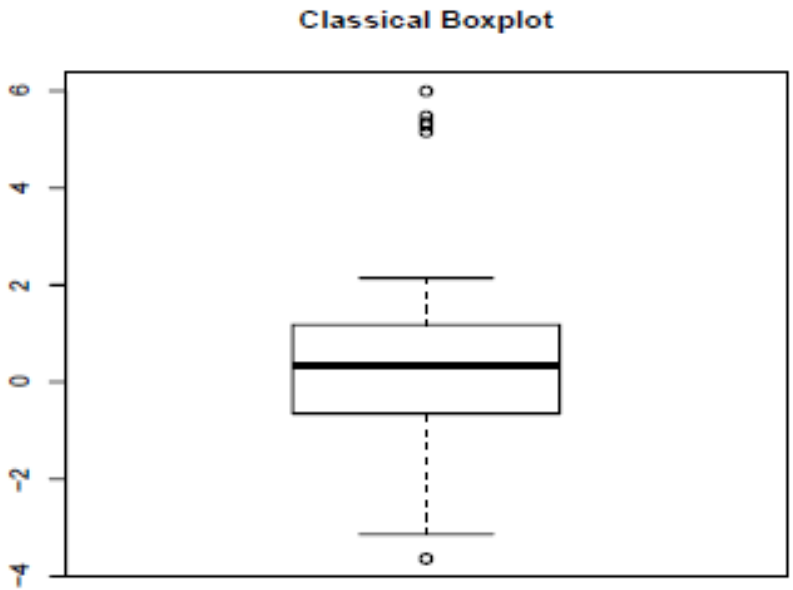


Outline

1. 函数型数据简介
   - 什么是函数型数据
   - 一般处理步骤
2. 函数型数据可视化
   - 曲线排齐
   - 数据模式的展示
   - 盒状图（boxplot）
   - 袋状图（bagplot）
   - 奇异值分解彩虹图（SVD rainbow plot）
   - 异常值展示

传统盒状图

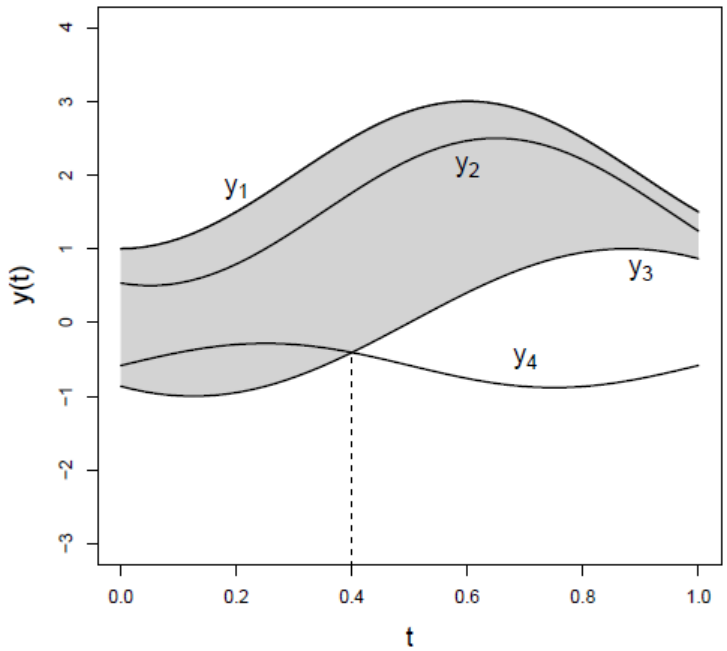
大家肯定对 boxplot 无限熟悉（此处基本假设您学过统计学或者做过量化数据分析，当然我不保证这个原假设是否成立：p）



上图很直观的展示了“五数”。怎么把函数型数据也由 boxplot 展示呢？显然我们需要知道函数型数据的“特征数”，而且把“秩”的概念推广到函数或者曲线中。

曲线的深度

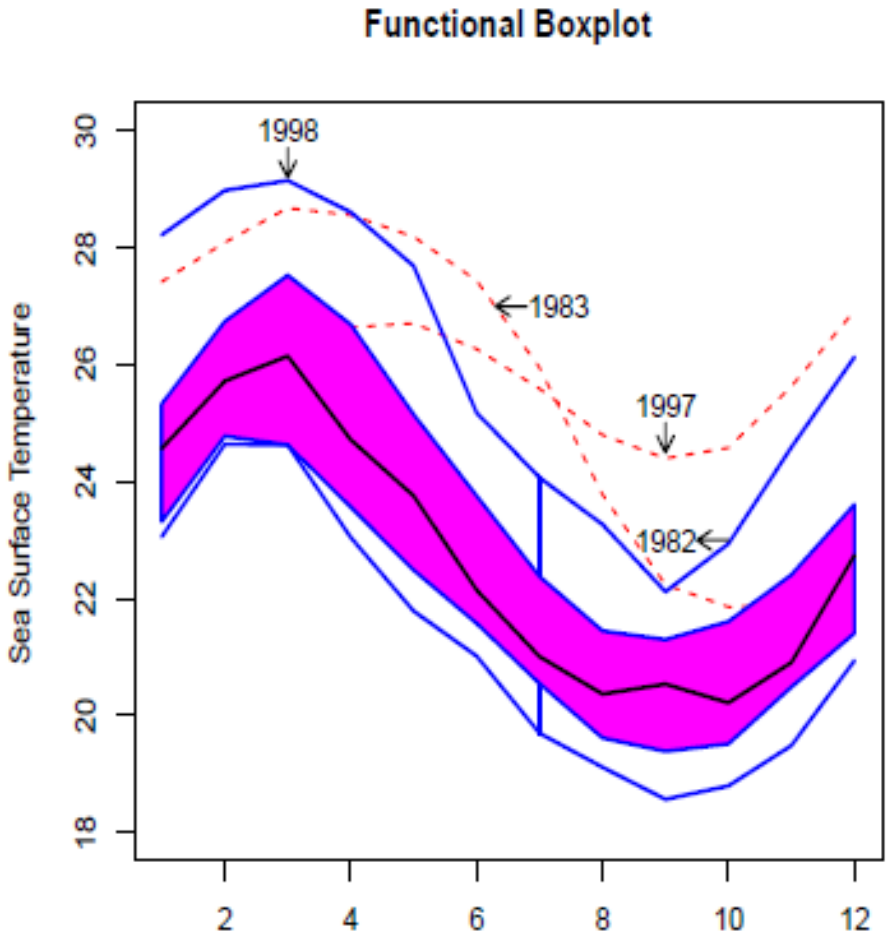
曲线的深度是由统计学家引进的，对应于数据的排序，或曰秩，的对应概念。深度可以明显由下例展示：



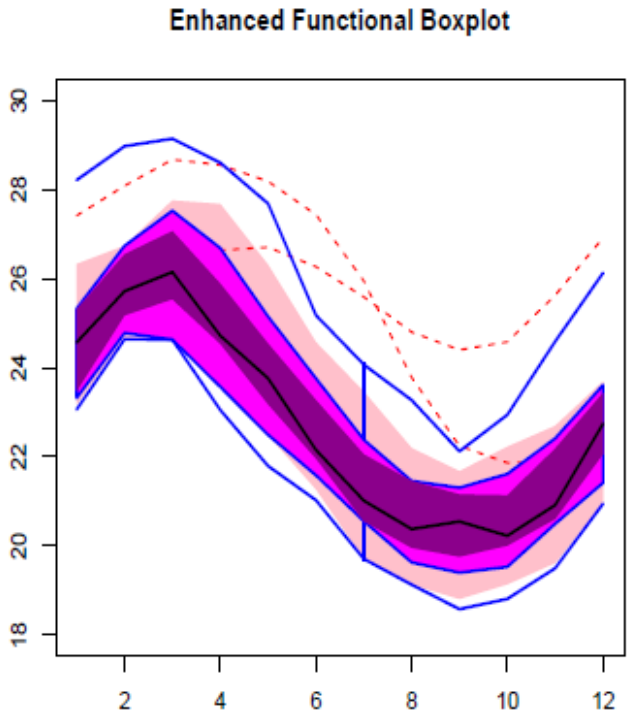
有了深度的定义，可以方便计算每条曲线，也就是每个观测对象在样本总体中的排序，从而按照排序获得“五数”并展示出来。

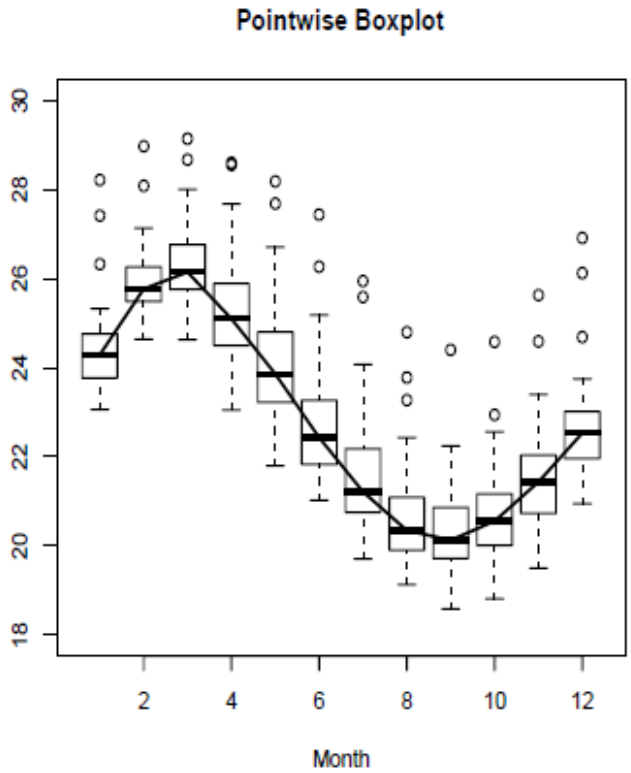
函数型盒状图

函数型盒状图: 中位线, 50% 中心区域, 正常取值范围, 经验异常.



有对比才有鉴别


Enhanced Functional Boxplot
30
28
26
24
22
20
18
2
4
6
8
10
12

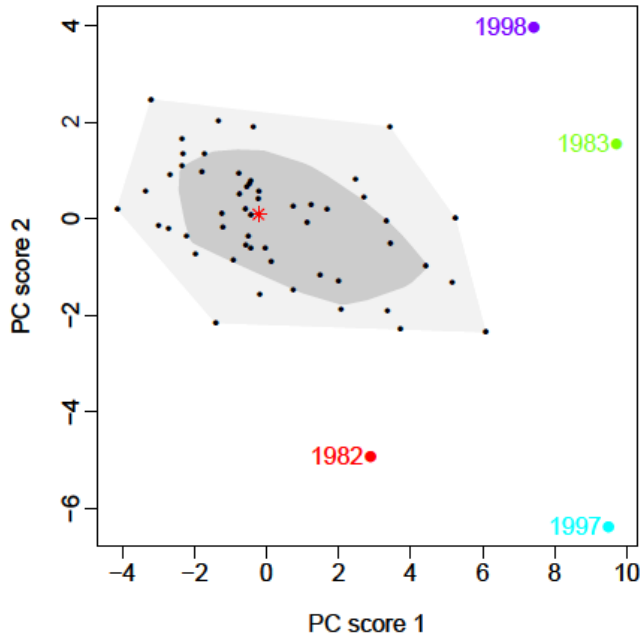

Pointwise Boxplot
30
28
26
24
22
20
18
2
4
6
8
10
12
Month

Outline

1. 函数型数据简介
   - 什么是函数型数据
   - 一般处理步骤
2. 函数型数据可视化
   - 曲线排齐
   - 数据模式的展示
   - 盒状图（boxplot）
   - 袋状图（bagplot）
   - 奇异值分解彩虹图（SVD rainbow plot）
   - 异常值展示

和 boxplot 一致，袋状图（bagplot）也是想展示数据的中位数以及异常值等需要用不同曲线的“秩”来表示的值，但是它采用的是完全不同的思路定义曲线的顺序。

首先，对数据进行稳健主成分分析，得出前两个主成分，从而每个对象都可以由两个主成分的不同取值表示，由一条数据线变成一个坐标点。下一步是如何展示数据点并排序。目前使用的最多的是双变量得分深度法，思想来自于 halfspace location depth (Tukey, 1974)。或者也可以采用双变量核密度估计，以估计的核函数在每点的取值作为该点的深度。

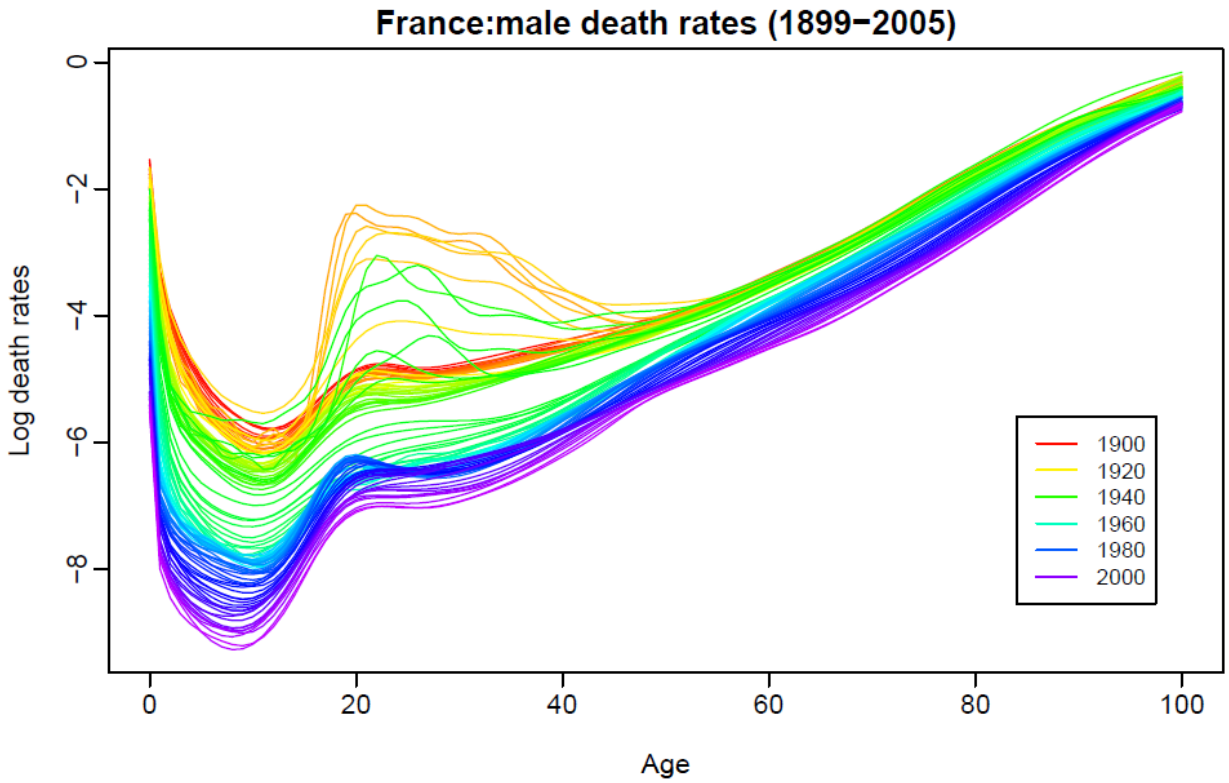


Outline

1. 函数型数据简介
   - 什么是函数型数据
   - 一般处理步骤
2. 函数型数据可视化
   - 曲线排齐
   - 数据模式的展示
   - 盒状图（boxplot）
   - 袋状图（bagplot）
   - 奇异值分解彩虹图（SVD rainbow plot）
   - 异常值展示

彩虹图

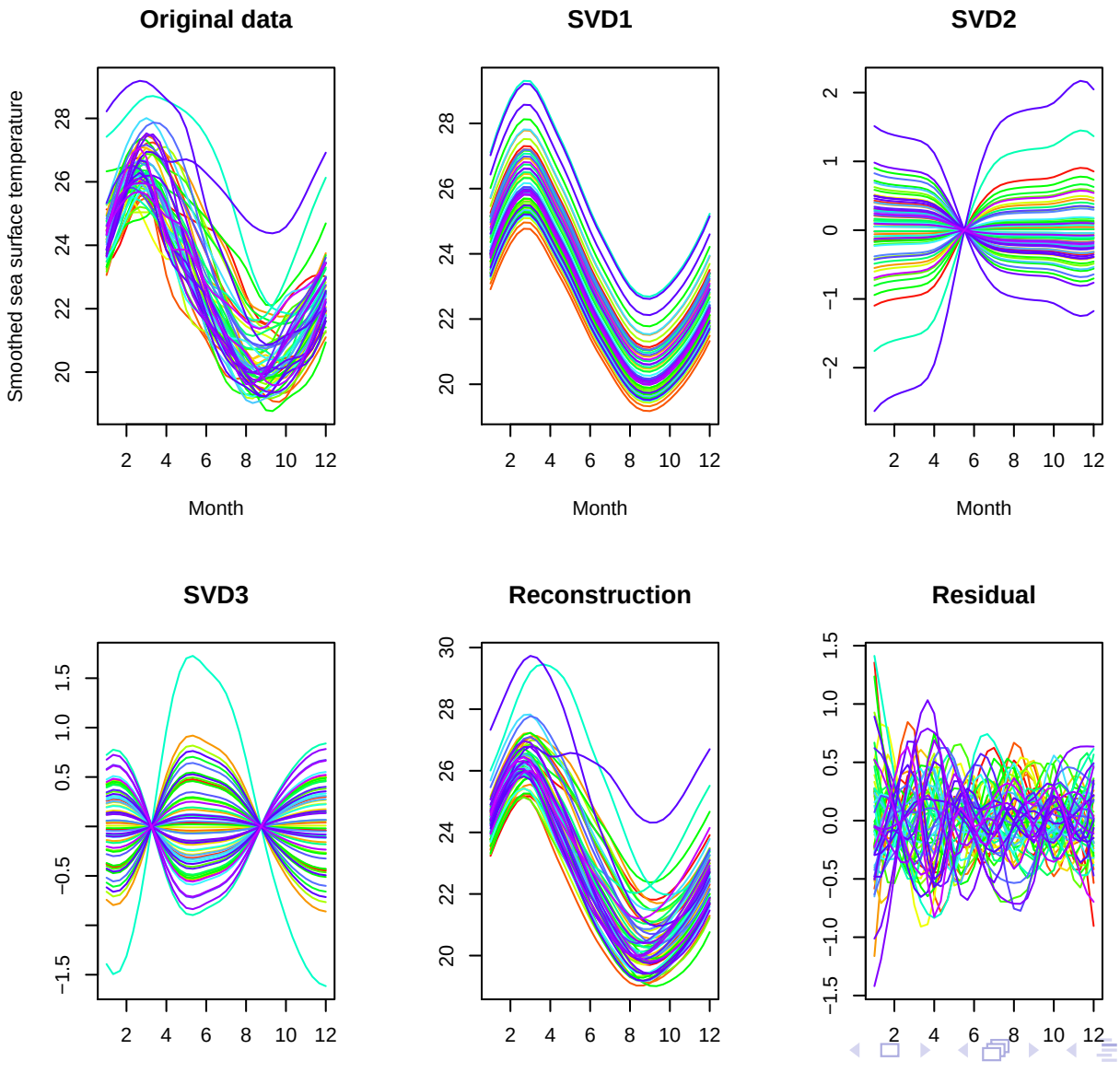
彩虹图的概念比较简单，就是把不同的对象的曲线按照彩虹的颜色变化标注，最为常见的做法是以时间顺序，也可以根据个人目的做出不同的标记，如下图：



彩虹图展示 SVD

SVD，奇异值分解，是谱分析理论在任意矩阵上的推广。对于矩阵 A(m*n)，存在 U(m*m)，V(n*n)，S(m*n)，满足 A = U*S*V'。U 和 V 中分别是 A 的奇异向量，而 S 是 A 的奇异值。AA' 的正交单位特征向量组成 U，特征值组成 S'S，A'A 的正交单位特征向量组成 V，特征值（与 AA' 相同）组成 SS'。因此，奇异值分解和特征值问题紧密联系。显然奇异值分解也可以用来做主成分分析，而且可以直接用来展示！如果你觉得自己线性代数学的不够好，不喜欢对着矩阵绕来绕去，R 可以帮你解决一切……

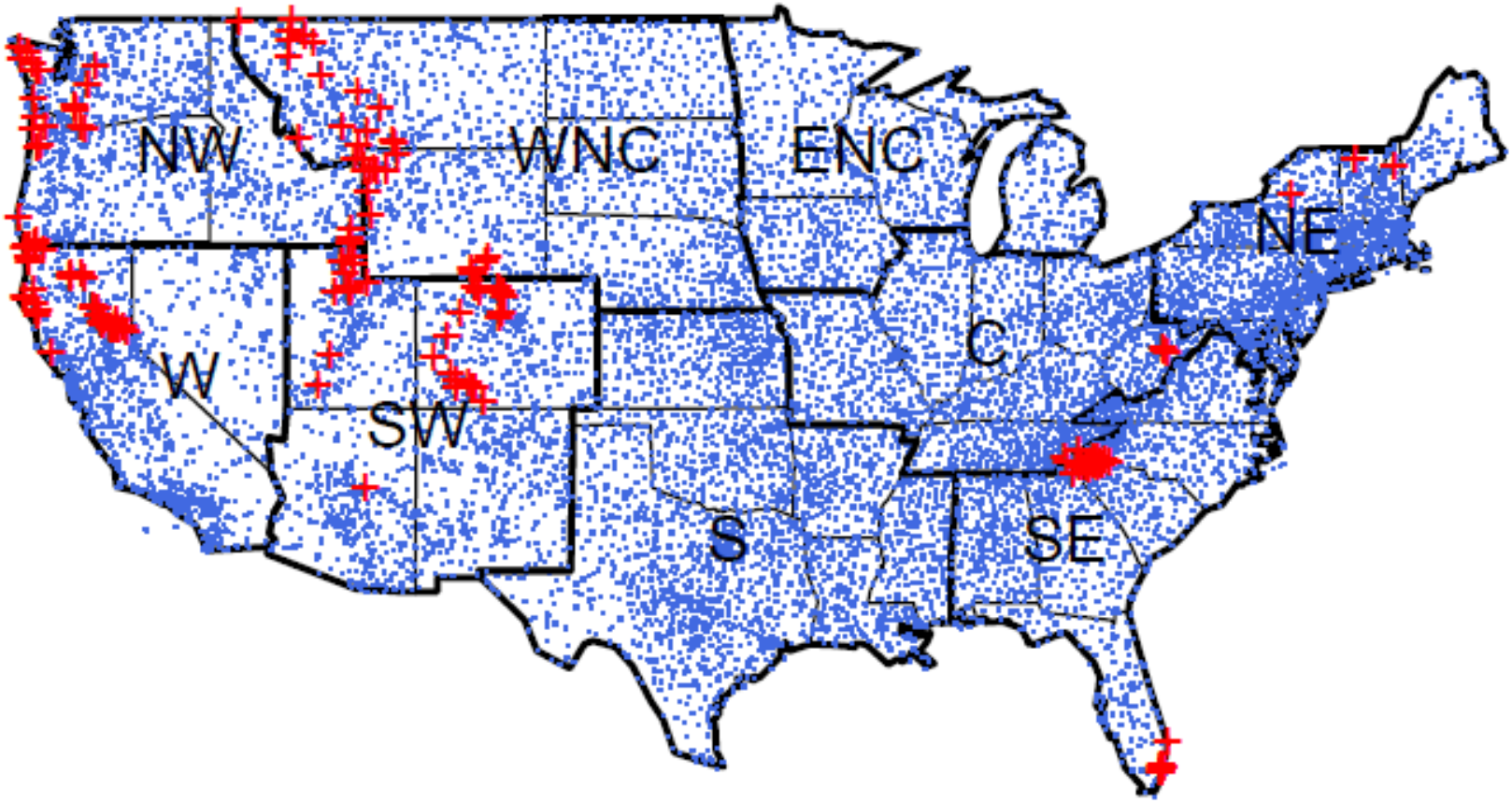
彩虹图展示 SVD


Original data
SVD1
SVD2
SVD3
Reconstruction
Residual
Smoothed sea surface temperature
Month

Outline

1. 函数型数据简介
   - 什么是函数型数据
   - 一般处理步骤
2. 函数型数据可视化
   - 曲线排齐
   - 数据模式的展示
   - 盒状图（boxplot）
   - 袋状图（bagplot）
   - 奇异值分解彩虹图（SVD rainbow plot）
   - 异常值展示

盒装图和袋状图都可以用来检测异常值，就是那些有最小深度或者最小密度的观测。很多统计学家很喜欢这个，因为他们可以不断 simulation 出各种数据，来对比和展示自己的方法对于异常值检测的优势（要发 paper 肯定要有 simulation 和对比分析啊亲 ~ 你让不让我们做 faculty 的人混饭吃呀！），可惜我一直木有明白，到底是检测出很多异常还是检测出很少异常才说明方法好呀……不过很多图还是很炫的


NW
WNC
ENC
NE
W
SW
C
S
SE

Thanks to Marc G. Genton from TAMU

For Further Reading

Books:

- Ramsay, J. O.; Hooker, Giles; and Graves, Spencer (2010) Functional Data Analysis with R and Matlab, Springer, New York.
- Ramsay, James O., and Silverman, BernardW. (2006), Functional Data Analysis, 2nd ed., Springer, New York.

Papers:

- R. J. Hyndman and H. L. Shang. (2010) "Rainbow plots, bagplots, and boxplots for functional data", Journal of Computational and Graphical Statistics, 19(1), 29-45.
- Ying Sun, Marc G. Genton. (2011) Functional Boxplots. Journal of Computational and Graphical Statistics 20:2, 316-334 .

Packages:

- fda http://cran.r-project.org/web/packages/fda/
- rainbow http://cran.r-project.org/web/packages/rainbow/