现代统计图形

谢益辉

爱荷华州立大学统计系

2010年6月14日

第三届中国R语言会议（北京）培训@中国人民大学

大纲

图形的推断功能：霍乱传染之谜

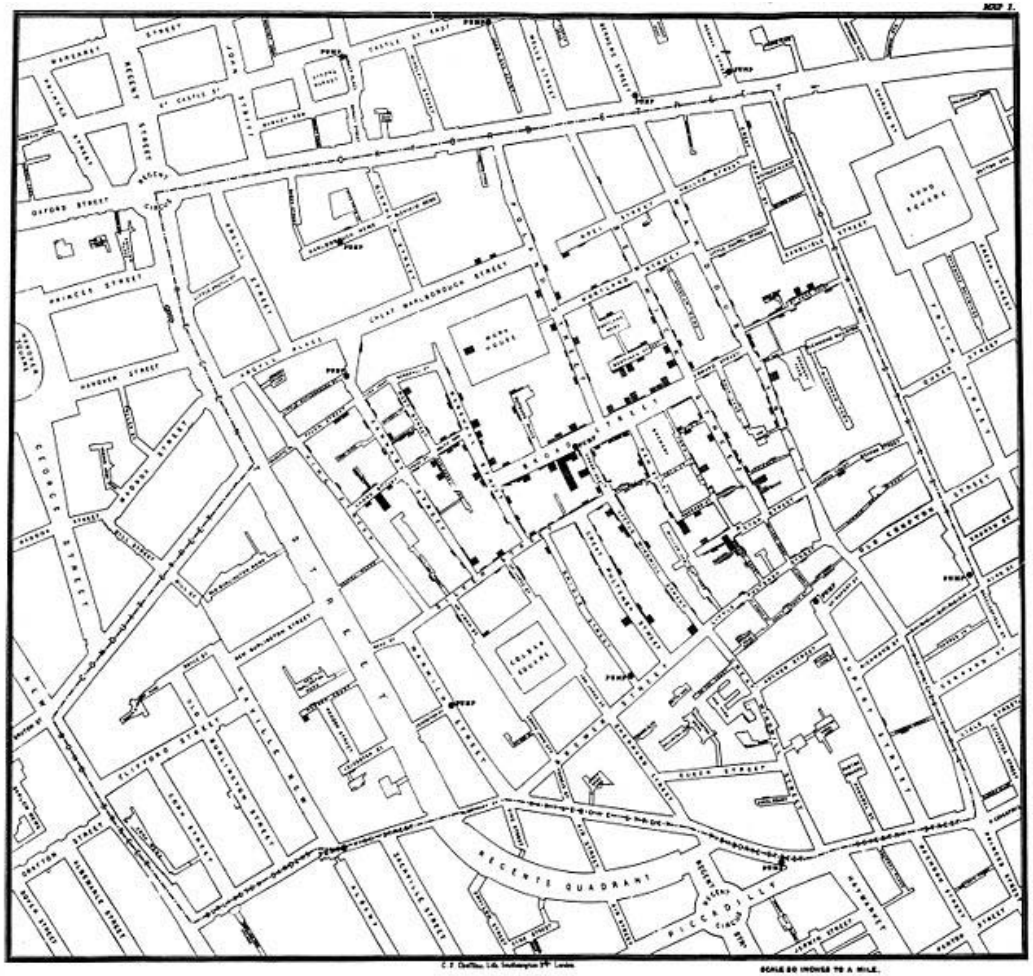

Figure 1: John Snow的霍乱传染原因探索图

图形的描述功能：拿破仑的远征

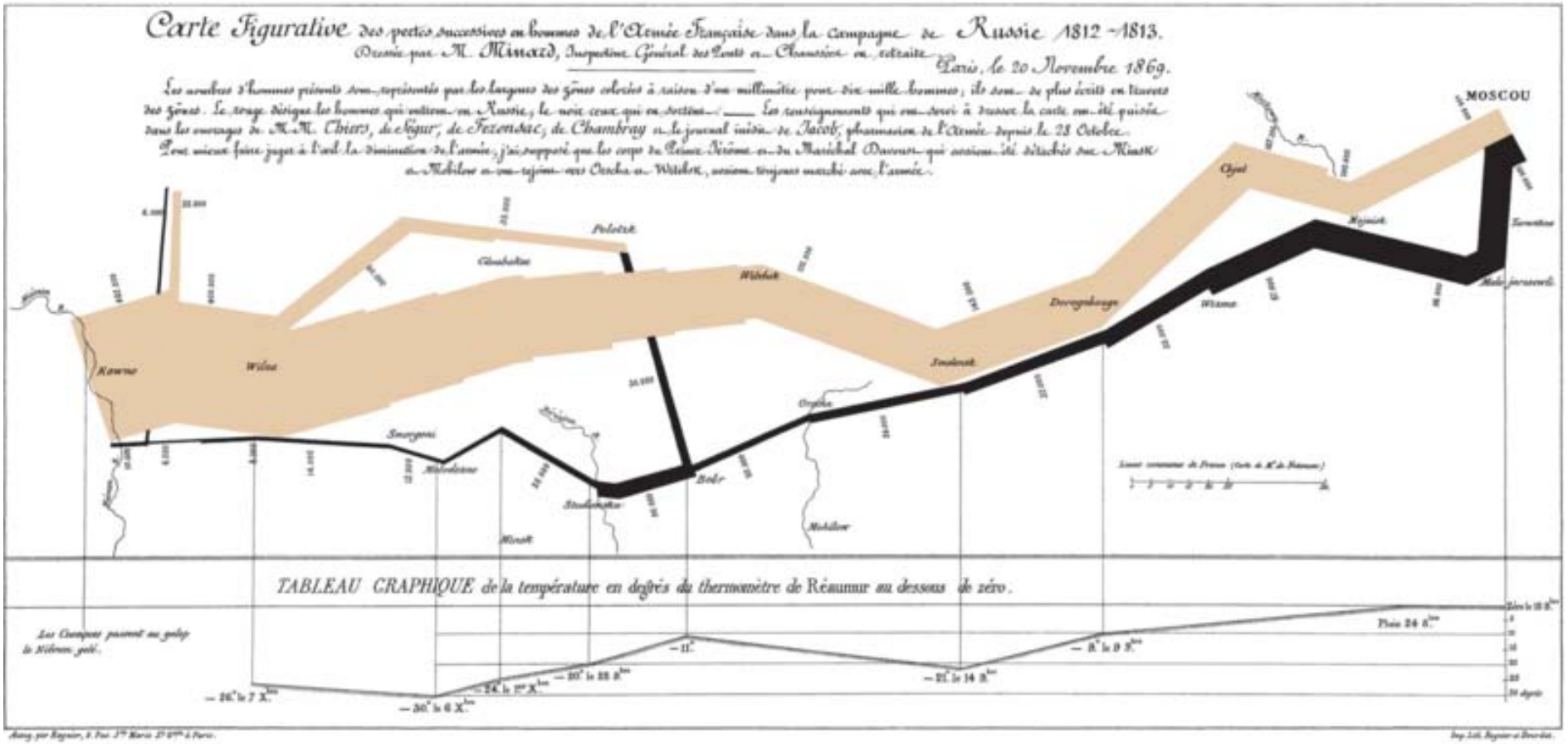

Figure 2: Charles Joseph Minard的拿破仑远征图

基础图形系统（base graphics）

- 灵活
  - 任何图形无非就是点线面构成的
  - 基础图形系统为图形基础元素提供了详尽的设置
- 繁琐
  - 挑剔的用户会觉得基础图形系统繁琐到无法忍受
  - par()就是这样一个恶魔

par()的功能简介

- 参见《现代统计图形》书稿3.1节
- 演示示例

最常用的一个泛型函数：plot()

- 什么是泛型函数（generic function）?
- 两个数值向量—散点图：plot.default()
  - 定制外观，演示示例
- 其它数据类型
  - data.frame：散点图矩阵，例`plot(iris)`
  - 数值型对因子型：箱线图，例`plot(Petal.Width ~ Species, data = iris)`
  - 因子型：条形图，例`plot(factor(c(1, 1, 2)))`
  - 很多R包都创造特有数据类型，扩充`plot()`函数，用户使用起来更统一，例MASS包中的岭回归

颜色

- colors()
- 颜色可以通过三种方式传达：
  - 颜色名称，如`red`
  - 整数：对应当前调色板`palette()`
  - 16进制的三个数字表示三原色：如`#FF0000`
  - 简单示例
- 特定主题调色板，如`heat.colors()`

点

- points()
- 两种表达点的形状的方式：
  - 整数
  - 单个字符

折线、直线、线段、曲线

- 给定x、y坐标连折线：lines()
- 给定斜率和截距连直线：abline()
- 给定x、y坐标连线段：segments()（与折线的区别？）
- 给定x、y坐标连光滑曲线：xspline()

矩形、多边形

- rect()、polygon()
  - 填充颜色、样式
  - 边线

文本

- text()
- mtext()
- title()

图例

- legend()

坐标轴

- axis()

各类基础图形

- 一维原始数据：条形图、饼图、Cleveland点图、坐标轴须、带状图
- 散点图：散点图、向日葵散点图
- 曲线：函数曲线
- 密度和分布：直方图、茎叶图、QQ图
- 汇总：箱线图、因素效应图
- 分类数据关联：关联图、马赛克图
- 分类对连续：条件密度图、棘状图
- 分类画图：协同图
- 三维图形：颜色图、等高线图、三维透视图、平滑散点图
- 高维散点图：散点图矩阵、符号图

其它图形（附加包）

- 地图
- 小提琴图（vioplot包，或lattice）
- 脸谱图（TeachingDemos包）
- 平行坐标图（lattice，或ggplot2）
- 调和曲线图（http://cos.name/2009/03/parallel-coordinates-and-andrews-curve/）
- 二维箱线图（aplpack包，bagplot()）

lattice

- 重要思想：根据变量分类画图
- 统一使用方法：参数为formula类型
- 设置繁琐无比，个人认为不方便、也不美观

ggplot2

- 图层叠加的概念，如同魔方
  - 几何单位（Geom，点？线？光滑？）+统计变换（Stat，直方图？QQ图？）+尺度表示（Scale，颜色渐变？元素大小？）+坐标系（Co-ord，笛卡尔？极坐标？）+面板分类（Facet，根据分类变量分别画图）+元素位置调整（Position，条形图并列或堆积？散点随机微小打乱？）
  - 扩展了泛型函数：+（使用非常形象）
- 细节设置自动化，例如图例

模型本身可能的局限

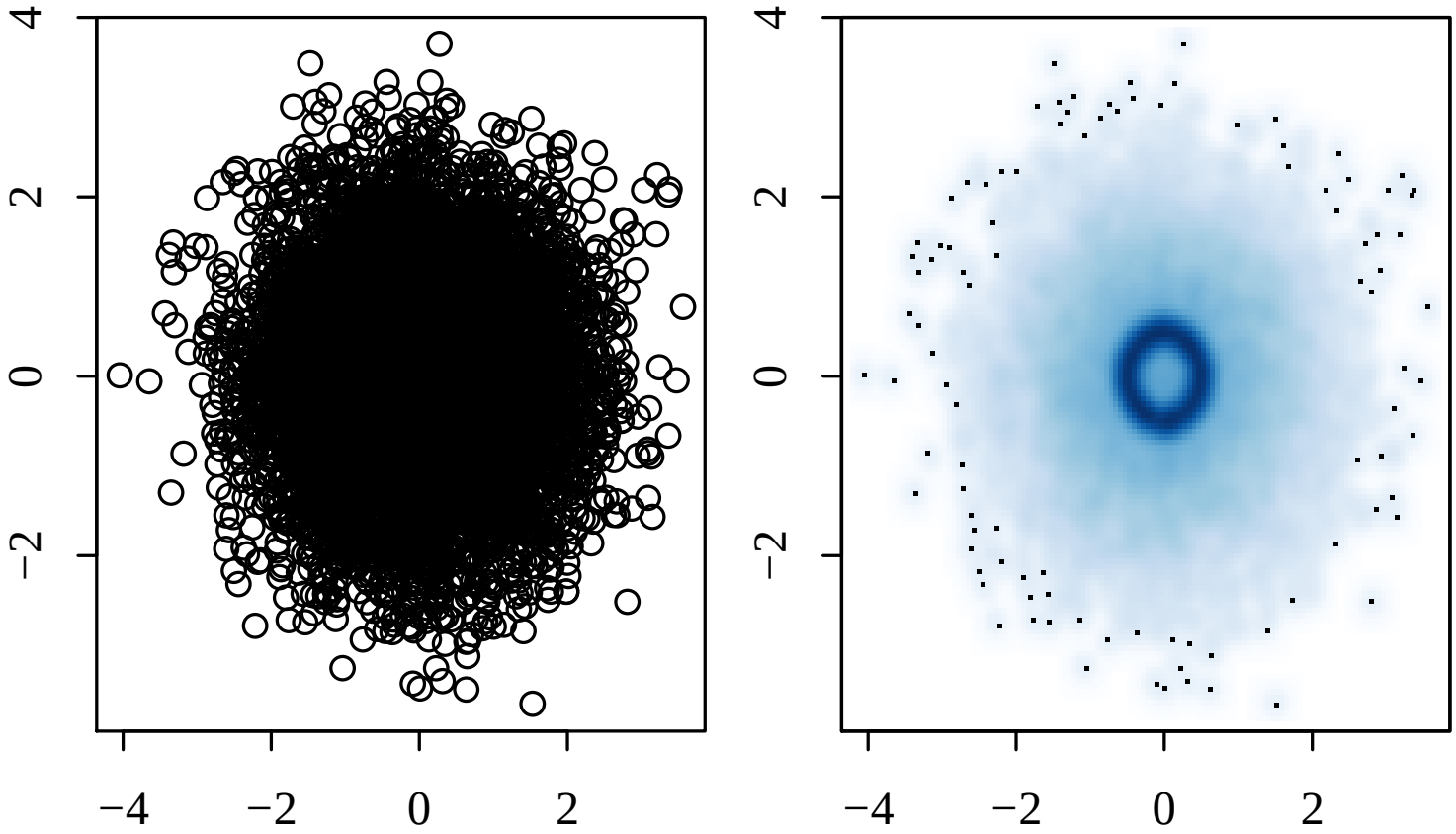

Figure 3: 寻找二维大数据中隐藏的特征（更多演示：
http://yihui.name/en/2008/09/to-see-a-circle-in-a-pile-of-sand/）

线性模型

- 一元回归：散点图，回归直线
- 多元回归
- 回归诊断

光滑方法

- 必杀技：lowess()/loess()
- 地图上的光滑

主成分分析和判别分析

- 得分散点图

多维标度分析

- 多维数据降维画散点图，保持距离的一致性

分类与回归树

- rpart包

图形中的数学公式标注

- expression()
- demo(plotmath)

图形文件输出

- pdf(), postscript()
- png(), jpeg()
- 高质量输出Cairo

k-Means聚类动态演示

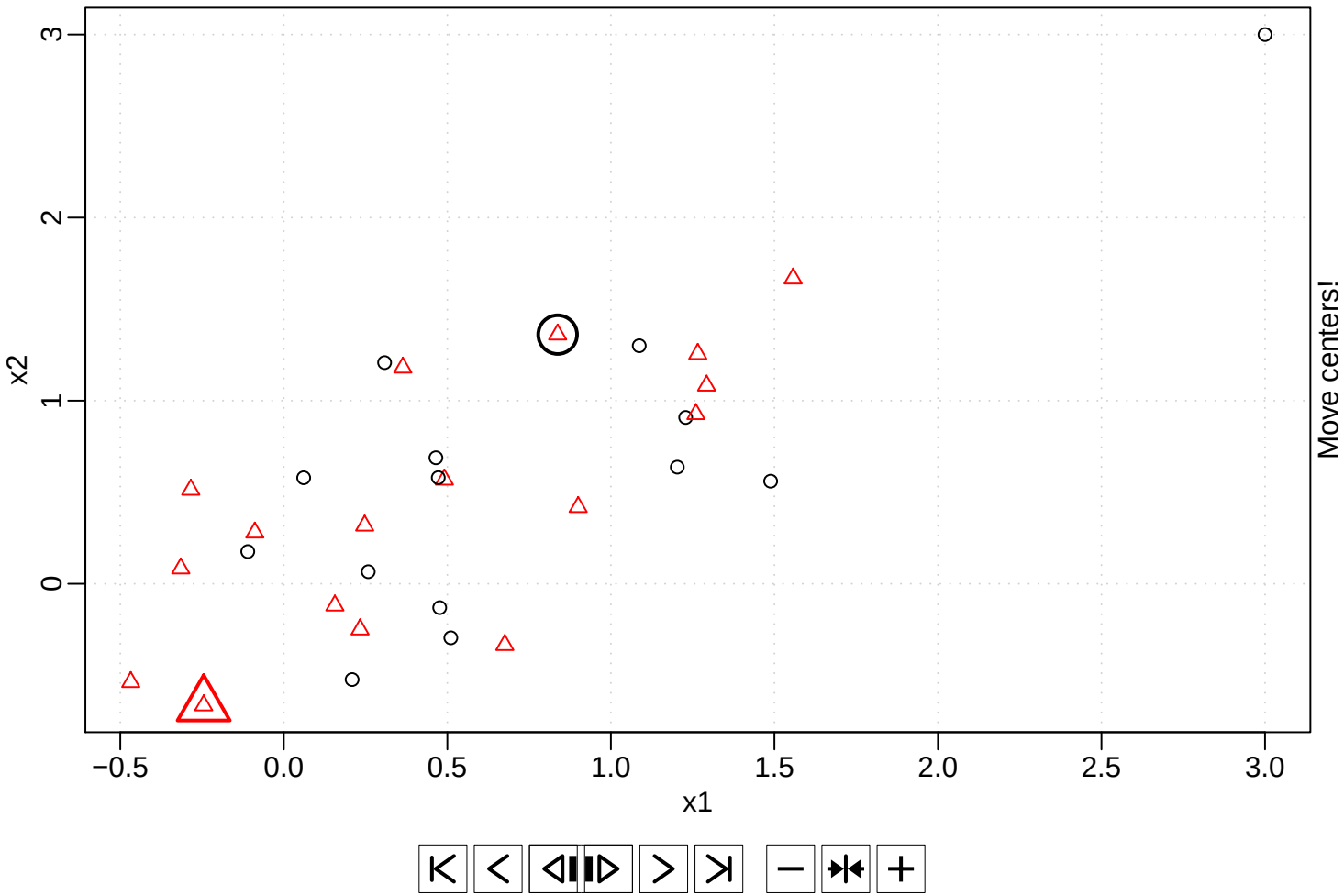

Figure 4: *k*-Means聚类过程及离群点的影响（`animation`包中`kmeans.ani()`）

多元回归的控制变量

- 回归初学者问题：为什么不拿因变量对每个自变量分别做回归？什么叫“控制变量”？
- 构造一个模拟的例子，看控制与不控制的效果，一目了然
- 思路：y本来随x增大而减小，但加入控制变量z之后y看起来随x增大而增大
- 效果：GGobi演示

关于作者

- 主页：http://yihui.name
- Email：xie@yihui.name
- COS论坛R版块：http://cos.name/cn/forum/15（若非与我个人相关的问题，请尽量发论坛）
- 谢谢各位!