统计图形和模拟视角下的模型理论解析

谢益辉

中国人民大学统计学院

2010年5月14日

大纲

解释理论：k-Means聚类

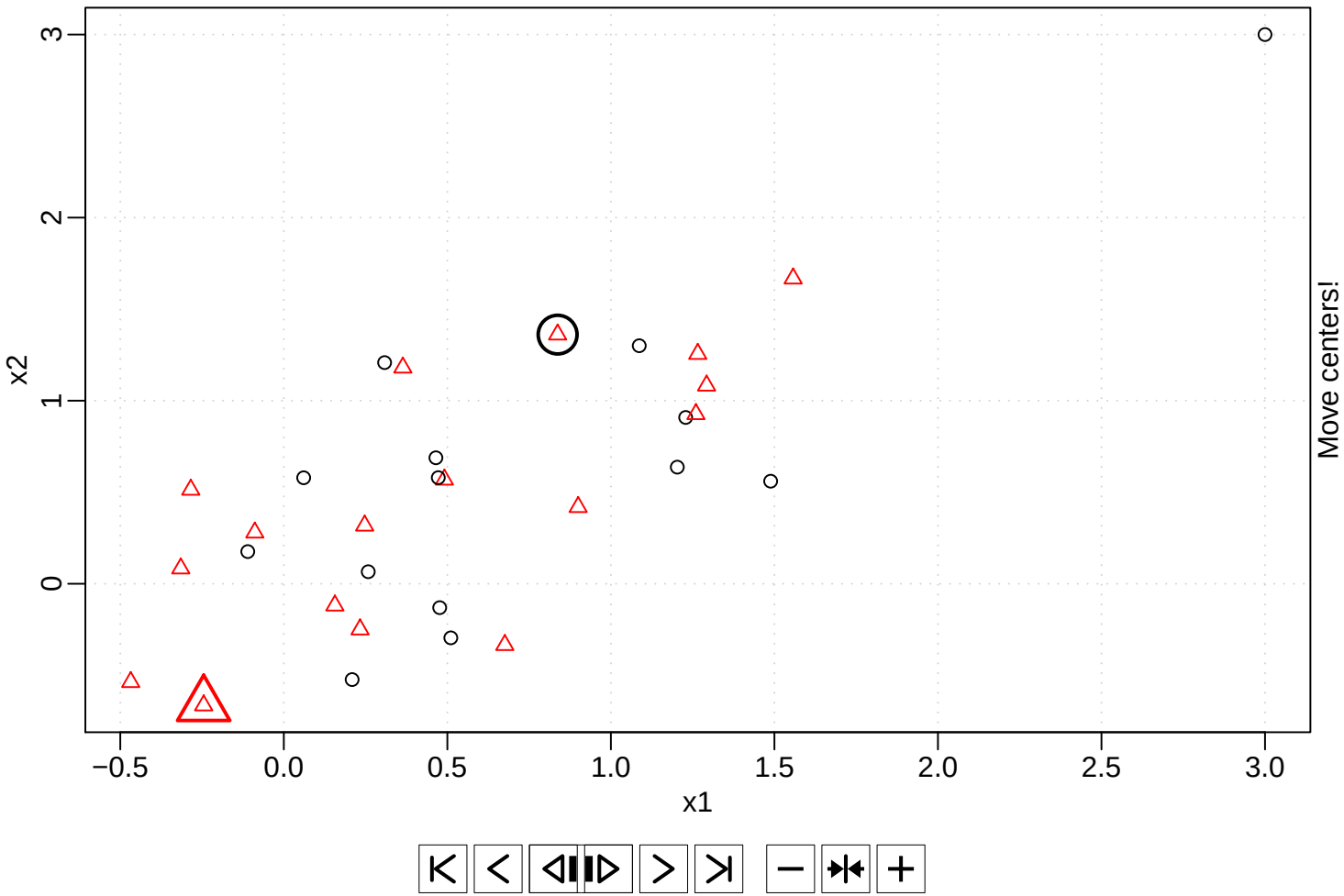

Figure 1: *k*-Means聚类的过程及离群点的影响（R包`kmeans.ani()`函数）

研究背景

- 根据Friendly and Denis (2001)的记录，世界上最早的统计图形主要起源于地图：地理位置的导航和探索
- 统计理论兼具复杂性和实际意义：如何探索模型的理论？
- 统计图形：直观呈现理论解释、快速反映关键信息
- 统计模拟：通过计算的方式进行“推导”

统计图形的发展

- 200多年前饼图诞生[1]
- 一个里程碑：Tukey (1977)的探索性数据分析，图形种类大大扩充，以探索数据为主
- 接下来是计算机语言的发展，S语言为另一个里程碑：快捷的交互式数据分析（图形为一大支柱）
- 其它独立图形软件或R包
- 图形的计算机支持已非常便利（简单演示：`library(rgl); demo(bivar)`）

[1]不过各种统计图形中，饼图可算是最糟糕的图形

统计模拟的意义

- Bootstrap方法的开篇作(Efron, 1979)
- Simon (1997)的5000美金赌注
- 统计教学：模拟与计算能更快获得答案

模型本身的局限

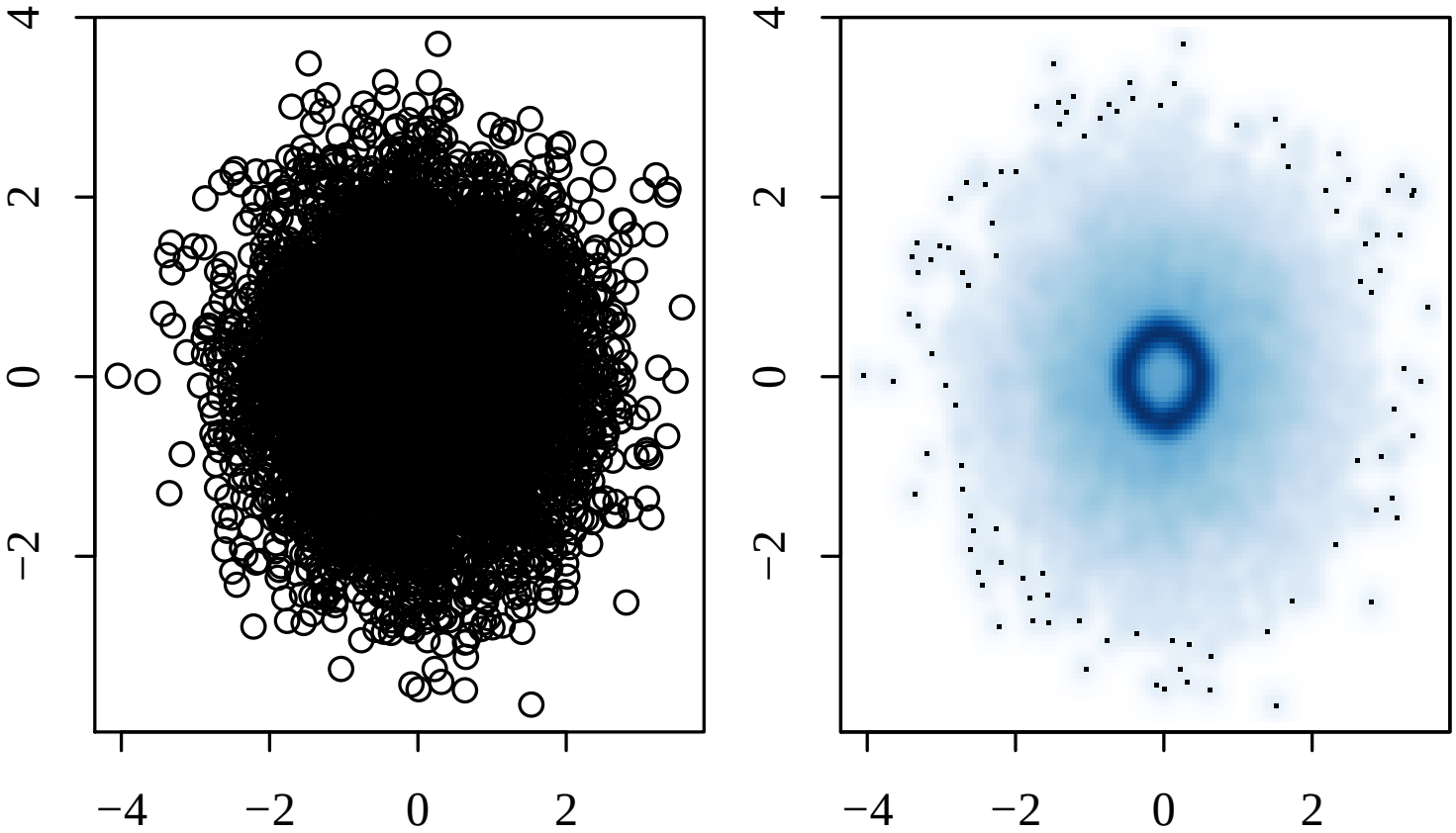

Figure 2: 寻找二维大数据中隐藏的特征（更多演示：
http://yihui.name/en/2008/09/to-see-a-circle-in-a-pile-of-sand/）

异方差t检验中的自由度校正

- 两样本t检验原本假设等方差，但在异方差的时候可用Welch校正调整t统计量的自由度，近似解决异方差问题
- 问题：使用或不使用Welch校正对检验有多大影响？
- 从理论推导入手也许能得到答案，而模拟会更快给出一组具体的答案：两组样本量的差异会影响检验结果的差异
- 过程：生成两组异方差的随机数，分别用等方差和异方差的公式算P值，然后对比之

异方差t检验中的自由度校正（续）

- 结果见论文中图2（样本量相等）、图3（样本量不等）和图4（样本量差异对P值差异的影响）
- t检验的等方差假设条件在样本量差异较大的时候稳健性很差
- 模拟的思路清楚，操作简单易行：从假设条件直接计算结果

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：论文图5（及GGobi演示）

最小中位数平方回归的性质

- 稳健回归：避免离群点的影响
- 最小中位数平方回归：
  $\hat{\beta} = \arg\min_{\beta} \text{median}\left\{(y_i - \hat{y}_i)^2\right\},\ i = 1, 2, \cdots, n$
- 它不总是优于最小二乘回归：对大量集中在数据中心的数据点非常敏感
- 模拟：生成大量集中在数据中心的数据
- 效果：论文图7
- 模拟一步到位，没有数学推导

多个离群点的诊断：Cook距离的局限

- 问题：若数据中存在多个离群点，它们会互相掩护，传统的删除单个样本看拟合值或系数变化的测度将失效
- 从重抽样或部分抽样的角度解决：既然一次删一个点不行，那么何不抽取样本的子集再拟合回归模型？
- 效果：论文图8（及网页动画）
- 计算的思路易于实施，在推公式之前不妨一试

关于神奇的87.53%

- 用图形发现数据的特征，论文图9
- 理论与模型可后行

LOWESS平滑

- 线性模型带有很强的假设
- 通常的非线性模型仍然带有很强的假设
- 在这些模型之前，可以让数据“自己说话”：LOWESS是一种方法
- 通过简单的计算和图形，可以让数据更有效地“说话”
- 植物数目案例，论文图10

假设检验之外？

- 假设检验本身是非常低效的数据分析工具：统计分析不是仅仅为了一个P值
- 可画图：箱线图、小提琴图等（论文图11、图12）
- 可模拟：重抽样，计算任意我们想知道的统计量（数学推导可能极其复杂）

Tukey首尾计数

- 手指头计算假设检验的方法（检验两样本均值的差异）
- 某工厂的6-sigma黑带极其重视
- 而一则模拟说明，它的稳健性可能很差（论文图14）
- 计算机如此发达，是否有必要推广手指头式的计算？

小结

- 模拟：快速得到答案
- 图形：直观反映事实
- 现实：统计软件在输出报表、报表、报表……
- 问题：统计理论可否通过模拟和图形变得“有趣”?

展望

- 统计计算和模拟的潜力有待大力发掘
- 除了探索数据，我们也可以并且应该探索理论
- R语言？

参考文献

Efron B (1979). "Bootstrap methods: another look at the jackknife." *The Annals of Statistics*, **7**(1), 1–26.

Friendly M, Denis DJ (2001). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Accessed: March 18, 2010, URL http://www.math.yorku.ca/SCS/Gallery/milestone/.

Simon JL (1997). *Resampling: The New Statistics*. 2nd edition. Resampling Stats. URL http://www.resample.com/content/text/index.shtml.

Tukey JW (1977). *Exploratory data analysis*. Massachusetts: Addison-Wesley.