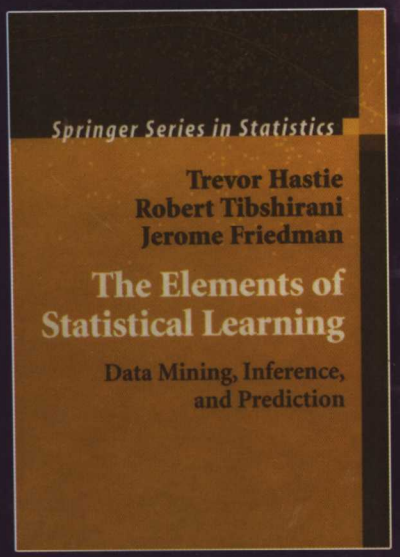
国外计算机科学教材系列

统计学习基础

——数据挖掘、推理与预测

The Elements of Statistical Learning

Data Mining, Inference, and Prediction



Trevor Hastie
[美] Robert Tibshirani 著
Jerome Friedman
范 明 柴玉梅 昝红英 等译

電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

经典教材

统计学习基础
——数据挖掘、推理与预测

The Elements of Statistical Learning
Data Mining, Inference, and Prediction

随着计算机和信息时代的到来，统计问题的规模和复杂性都有了急剧增加。数据存储、组织和检索领域的挑战导致一个新领域“数据挖掘”的产生。数据挖掘是一个多学科交叉领域，涉及数据库技术、机器学习、统计学、神经网络、模式识别、知识库、信息提取、高性能计算等诸多领域，并在工业、商务、财经、通信、医疗卫生、生物工程、科学等众多行业得到了广泛的应用。

本书试图将学习领域中许多重要的新思想汇集在一起，并且在统计学的框架下解释它们。尽管有些数学细节是必要的，但本书强调的是方法和它们的概念基础，而不是理论性质。本书内容广泛，从有指导的学习（预测）到无指导的学习，应有尽有。包括神经网络、支持向量机、分类树和提升等主题，是同类书籍中介绍得最全面的，适合从事数据挖掘和机器学习研究的读者阅读。

作者简介

Trevor Hastie, Robert Tibshirani 和 Jerome Friedman 都是斯坦福大学统计学教授，并在这个领域做出了杰出的贡献。Hastie 和 Tibshirani 提出了广义加法模型，并出版了专著“Generalized Additive Models”。Hastie 的主要研究领域为：非参数回归和分类、统计计算以及生物信息学、医学和工业的特殊数据挖掘问题。他提出了主曲线和主曲面的概念，并用 S-PLUS 编写了大量统计建模软件。Tibshirani 的主要研究领域为：应用统计学、生物统计学和机器学习。他提出了套索的概念，还是“An Introduction to the Bootstrap”一书的作者之一。Friedman 是 CART、MARS 和投影寻踪等数据挖掘工具的发明人之一。他不仅是位统计学家，而且是物理学家和计算机科学家，并先后在物理学、计算机科学和统计学的一流杂志上发表了论文 80 余篇。

ISBN 7-5053-9331-6

9 787505 393318 >

责任编辑：杜闽燕
封面设计：毛惠庚

ISBN 7-5053-9331-6　定价：45.00 元

国外计算机科学教材系列

统计学习基础

——数据挖掘、推理与预测

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Trevor Hastie

[美] Robert Tibshirani 著

Jerome Friedman

范 明 柴玉梅 昝红英 等译

電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内容简介

计算和信息技术的飞速发展带来了医学、生物学、财经和营销等诸多领域的海量数据。理解这些数据是一种挑战，这导致了统计学领域新工具的发展，并延伸到诸如数据挖掘、机器学习和生物信息学等新领域。许多工具都具有共同的基础，但常常用不同的术语来表达。本书介绍了这些领域的一些重要概念。尽管应用的是统计学方法，但强调的是概念，而不是数学。许多例子附以彩图。本书内容广泛，从有指导的学习（预测）到无指导的学习，应有尽有。包括神经网络、支持向量机、分类树和提升等主题，是同类书籍中介绍得最全面的。

本书可作为高等院校相关专业本科生和研究生的教材，对于统计学相关人员、科学界和业界关注数据挖掘的人，本书值得一读。

版权贸易合同登记号　图字：01-2002-4937

图书在版编目（CIP）数据

统计学习基础——数据挖掘、推理与预测/（美）黑斯蒂（Hastie, T.）等著；范明等译.
-北京：电子工业出版社，2004.1
（国外计算机科学教材系列）
书名原文：The Elements of Statistical Learning: Data Mining, Inference, and Prediction
ISBN 7-5053-9331-6

Ⅰ.统… Ⅱ.①黑… ②范… Ⅲ.统计学-教材 Ⅳ.C8

中国版本图书馆CIP数据核字（2003）第124311号

责任编辑：杜闽燕
印　　刷：北京兴华印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　邮编：100036
经　　销：各地新华书店
开　　本：787×1092　1/16　　印张：24.75　字数：634千字　彩插：22
印　　次：2004年1月第1次印刷
定　　价：45.00元

凡购买电子工业出版社的图书，如有缺损问题，请向购买书店调换；若书店售缺，请与本社发行部联系。联系电话：（010）68279077。质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

图 1.3

DNA 微阵列数据：人体瘤数据6830个基因（行）和64个样本（列）的表达水平矩阵。只显示100行的随机选样。显示的是热度图，从鲜绿（负，低显性）到鲜红（正，高显性）。遗漏的值为灰色。行和列以随机次序显示

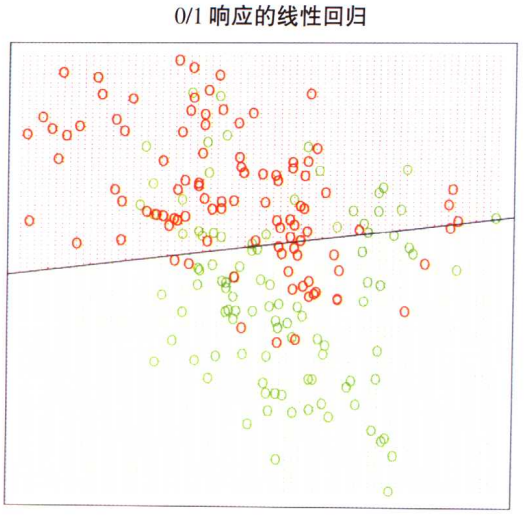


图 2.1

一个二维空间上的分类例子。类用二元变量编码（GREEN = 0，RED = 1），并且用线性回归拟合。直线是 $x^T\hat{\beta}=0.5$ 定义的判定边界。红色区域表示输入空间被分类为 RED 的部分，而绿色区域被分类为 GREEN

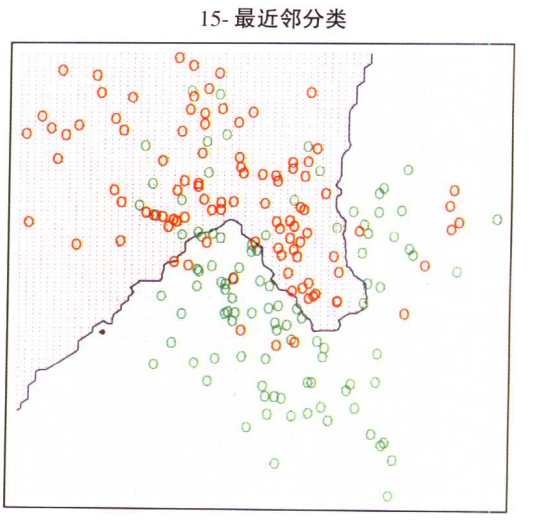


图 2.2

与图 2.1 相同的二维分类例子。类用二元变量编码（GREEN = 0，RED = 1），并用式（2.8）的 15- 最近邻平均拟合。因此，预测类用 15- 最近邻的多数表决确定

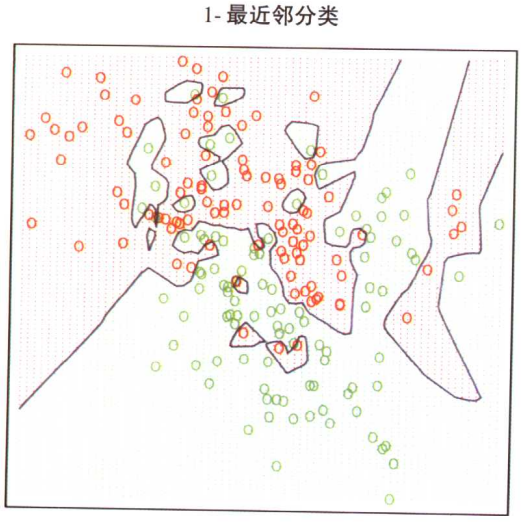


图 2.3

与图 2.1 相同的二维分类例子。类用二元变量编码（GREEN = 0，RED = 1），并用 1- 最近邻分类预测

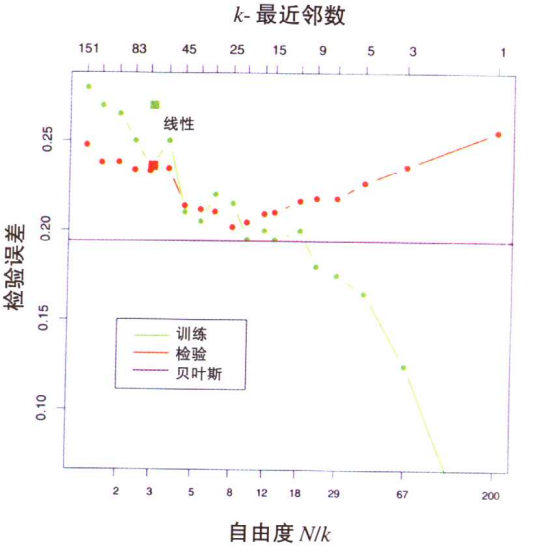


图 2.4

图 2.1、图 2.2 和图 2.3 使用的模拟例子的误分类曲线。使用一个规模为 200 的训练样本和一个规模为 10 000 的检验样本。红色曲线是 k- 最近邻分类的检验误差，绿色曲线是训练误差。线性回归的结果是三自由度上较大的红色和绿色方块。紫色直线是最优的贝叶斯误差率

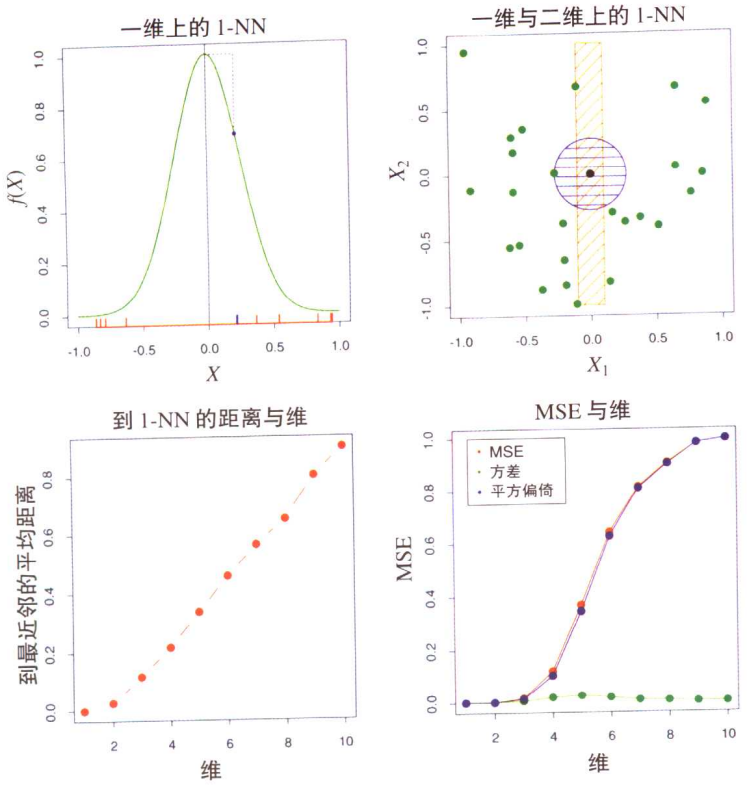


图 2.7

一个模拟例子，表明维灾难及其对MSE、偏倚和方差的影响。对于 $p=1, \cdots, 10$，输入特征值在$[-1, 1]^p$上均匀分布。左上角的图显示IR上的(无噪声)目标函数：$f(X)=e^{-8\|X\|^2}$，并图示1-最近邻对$f(0)$估计所产生的误差。训练点用蓝色粗体标记。右上角的图展示1-最近邻域的半径随维数p增加的原因。左下角的图展示1-最近邻域的平均半径。右下角的图显示作为维数p的函数，MSE、平方偏倚和方差的曲线

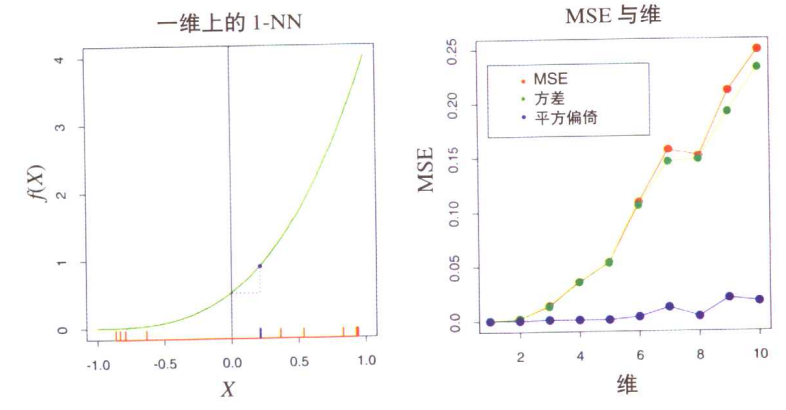


图 2.8

一个与图 2.7 具有相同设置的模拟例子。这里，除一个维为$f(X)=\frac{1}{2}(X_1+1)^3$外，函数为常数。方差占支配地位

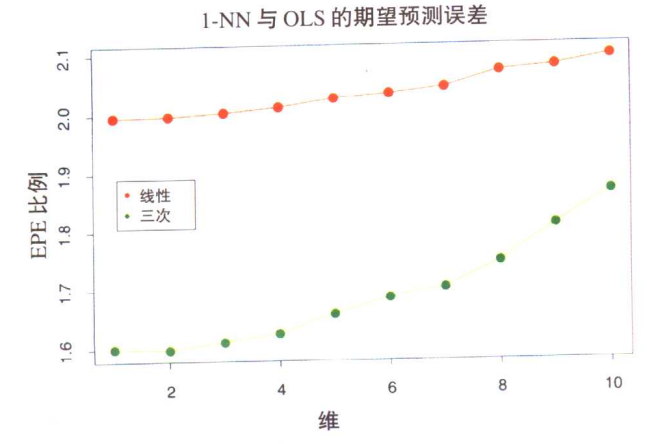


图 2.9

显示1-最近邻相对于最小二乘方关于模型$Y=f(X)+\varepsilon$的期望预测误差的曲线(在x_0=0)。对于红色曲线，$f(x)=x_1$；而对于绿色曲线，$f(x)=\frac{1}{2}(x_1+1)^3$

图 3.5

前列腺癌例子所有可能的子集模型。对每个子集容量，显示该容量的每个模型的残差平方和

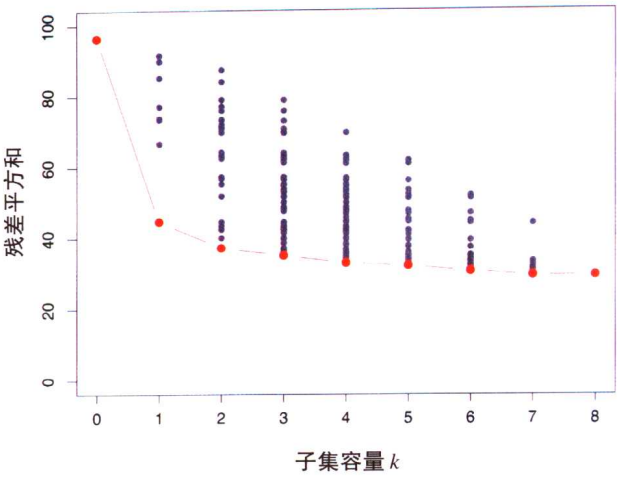
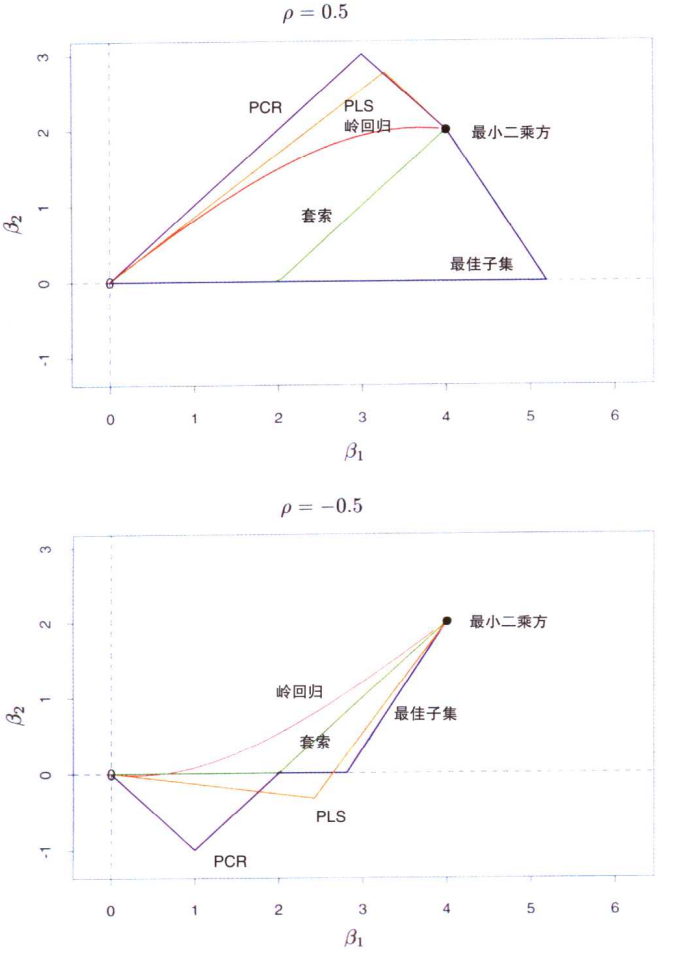


图 3.11

对于一个简单例子，不同方法的系数曲线图：两个具有相关度 ± 0.5 的输入，而实际的回归系数是 $\beta = (4, 2)$



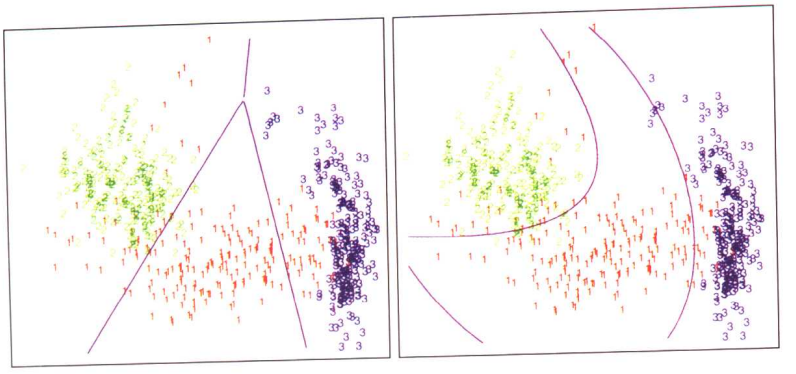


图 4.1

左图显示取自三个类的一些数据点，以及由线性判别分析找出的线性判定边界。右图显示二次判定边界。这些边界通过找出 5 维空间 $X_1, X_2, X_1X_2, X_1^2, X_2^2$ 中的线性边界得到。在该空间上的线性不等式是原空间中的二次不等式

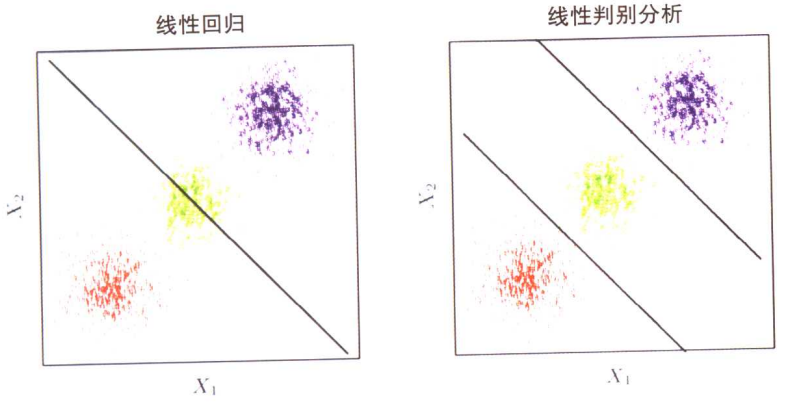


图 4.2

数据取自 IR^2 中的三个类，并容易被线性判定边界分开。右图显示被线性判别分析找到的边界。左图显示被指示响应变量的线性回归找出的边界。中间类完全被屏蔽（不占支配地位）

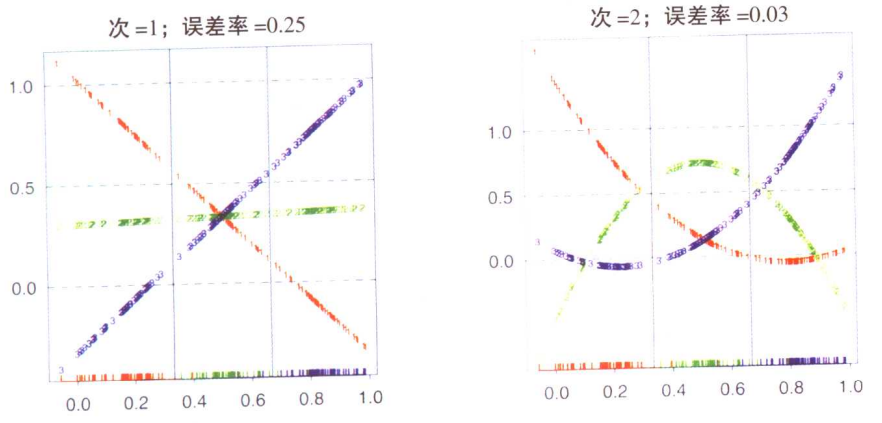


图 4.3

对于一个 3- 类问题，IR 上的线性回归的屏蔽作用。底部的底线图（rug plot）指示每个观测的位置和类隶属关系。每幅图上的三条曲线是 3- 类指示变量的拟合回归；例如，对于红色类，红色观测的 y_{red} 为 1，而绿色和蓝色观测的 y_{red} 为 0。每幅图的上方是训练误差率。对于该问题，贝叶斯误差率为 0.025，与 LDA 误差率一样

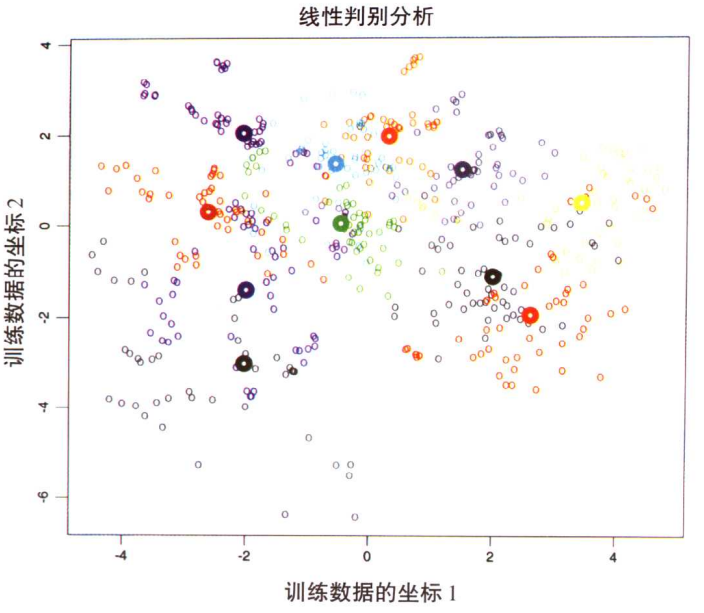


图 4.4

元音训练数据的二维图。有 11 个类，$X \in \mathrm{IR}^{10}$。这是 LDA 模型（见第 4.3.3 节）下的最佳视图。加重的圆是每个类的投影均值向量。类的重叠相当多

图 4.5

左图显示三个高斯分布，它们具有相同的协方差和不同的均值。图中包含的是每种情况围绕概率 95% 的常量密度围线。图中显示了每两个类之间的贝叶斯判定边界（虚线），而分离所有三个类的贝叶斯判定边界是粗实线（前者的子集）。在右图中，我们看到取自每个高斯分布的容量为 30 的样本，以及拟合的 LDA 判定边界

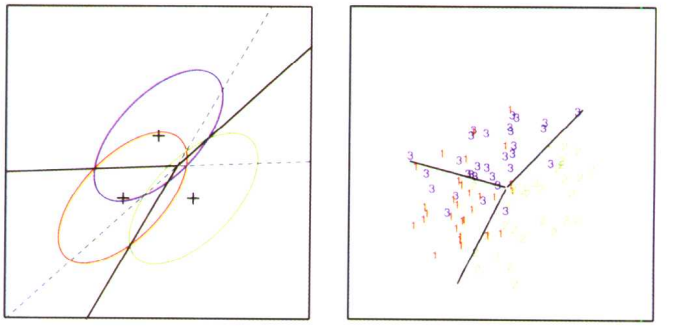
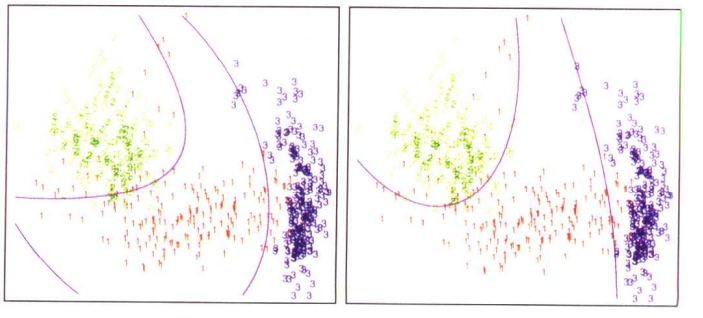


图 4.6

拟合二次边界的两种方法。对于图 4.1 的数据，左图显示二次判定边界（使用 5 维空间 X_1, X_2, X_1X_2, X_1^2, X_2^2 上的 LDA 得到）。右图显示QDA发现的二次判定边界。差别很小，通常也是如此



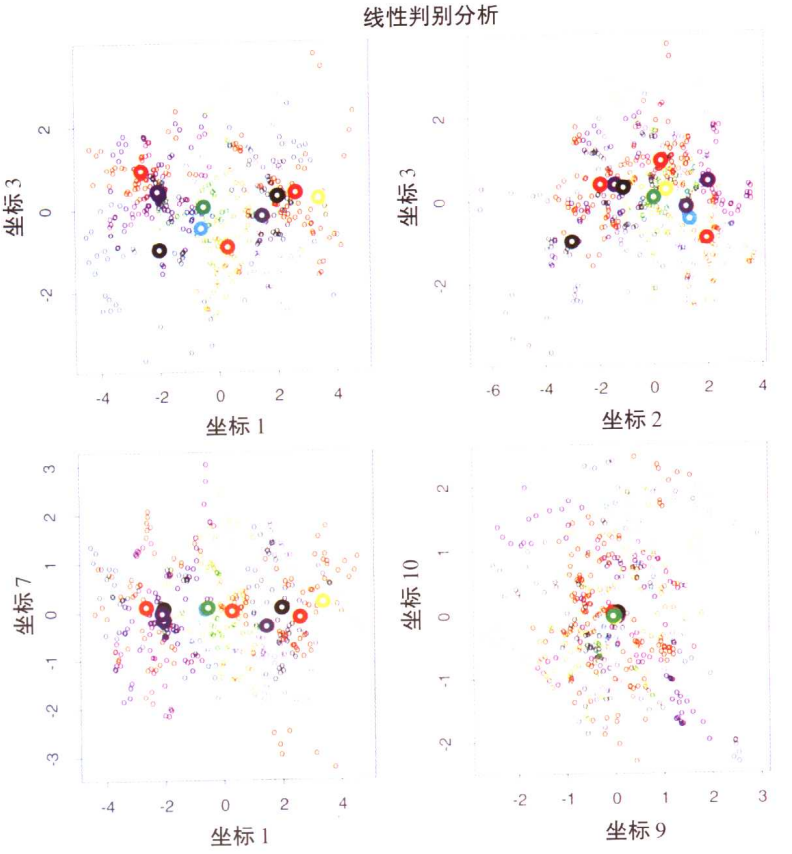


图 4.8
到标准变量对上的4个投影。注意，随标准变量的秩增加，形心变得集中。在右下图，它们似乎被叠加，并且类变得最乱

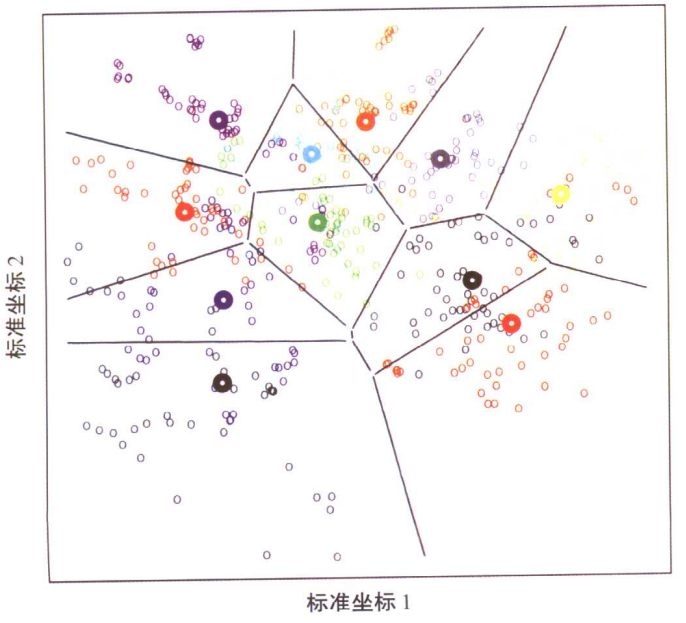


图 4.11
在前两个标准变量生成的二维子空间中元音训练数据的判定边界。注意，在任意较高维的子空间中，判定边界是较高维的仿射平面，不能用线表示

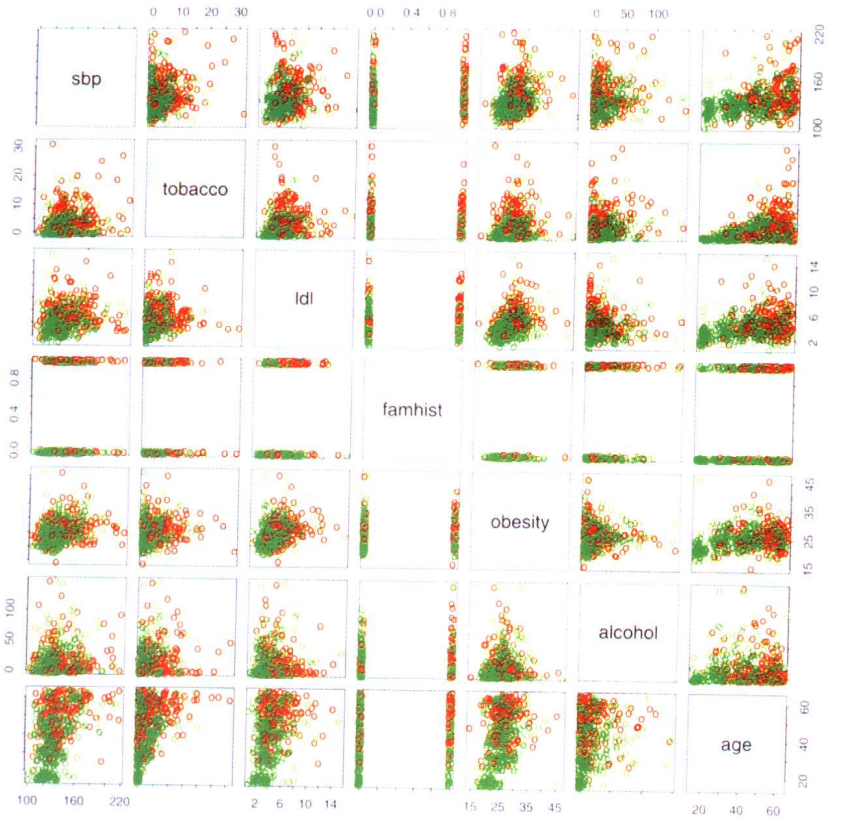


图 4.12

南非心脏病数据的散点图。每幅图显示一对风险因素，并且病例和控制用颜色编码（红色是病例）。心脏病家族史变量（famhist）是二元的（yes 或 no）

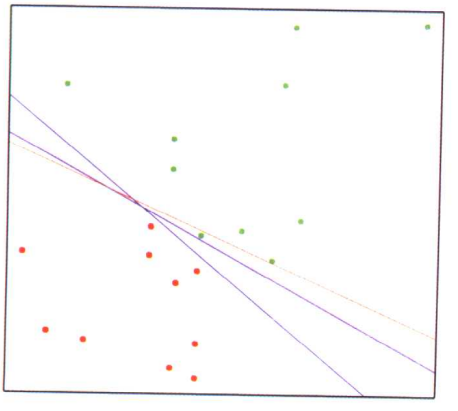


图 4.13

一个小例子，包含两个可被超平面分隔的类。橙色线是最小二乘方解，它将一个训练数据误分类。图中还显示了两个蓝色分隔超平面，它们被以不同的随机初始化的感知器学习算法找出

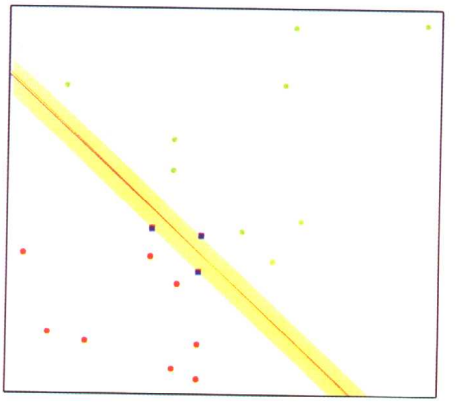


图 4.15

与图4.13相同的数据。阴影区域描述分离两个类的最大边缘。有三个支撑点，它们在边缘的边界上，而最佳分离超平面（蓝线）将隔离带一分为二。图中还显示了逻辑斯缔回归找出的边界（红线），它非常接近最佳分离超平面（见第12.3.3节）

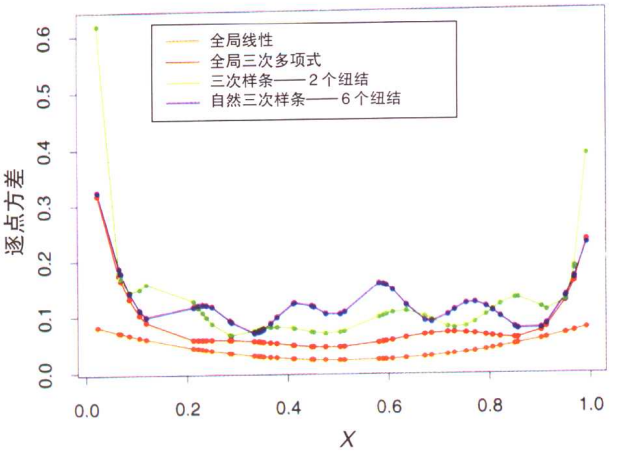


图 5.3

4个不同模型的逐点方差曲线。X 包含50个点，随机地取自 $U[0, 1]$，一个假定的误差模型具有常数方差。线性和三次多项式拟合分别具有2个和4个自由度，而三次样条和自然三次样条具有6个自由度。三次样条在0.33和0.66有两个纽结，而自然样条在0.1和0.9具有边界纽结，并且有4个内部纽结均匀地散布在它们中间

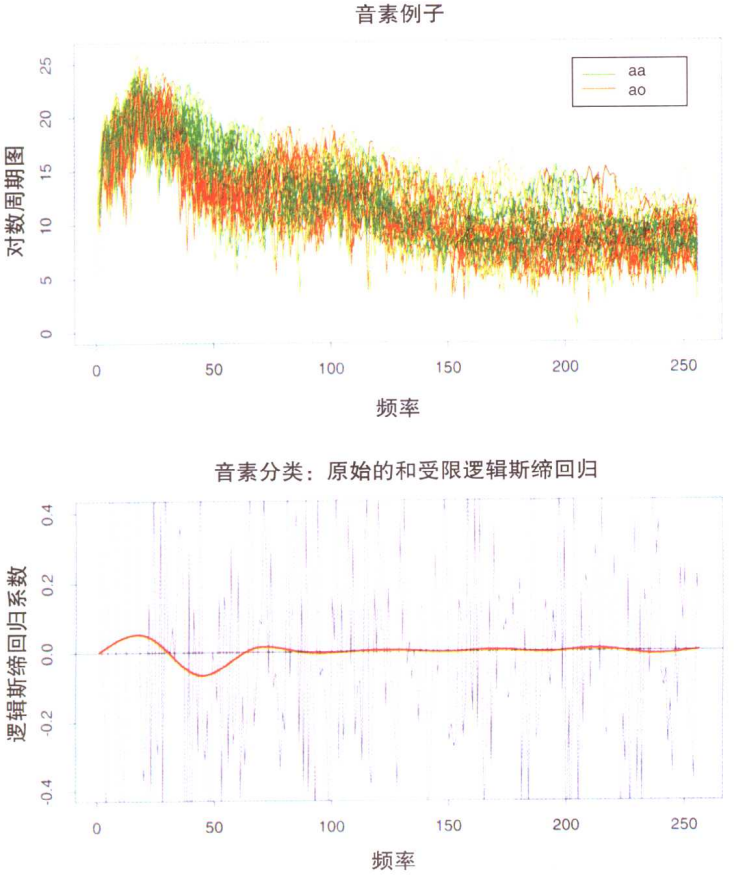


图 5.5

上图显示对数周期图；对于15个例子，每个音素“aa”和“ao”从695个“aa”和1022个“ao”中选样，对数周期图作为频率的函数显示。每个对数周期在256个均匀分布的频率点上测量。下面的图显示逻辑斯缔回归系数（作为频率的函数），使用256个对数周期图作为输入值，通过极大似然拟合数据。在红色曲线上，限制系数是光滑的，而在锯齿状灰色曲线上没有限制

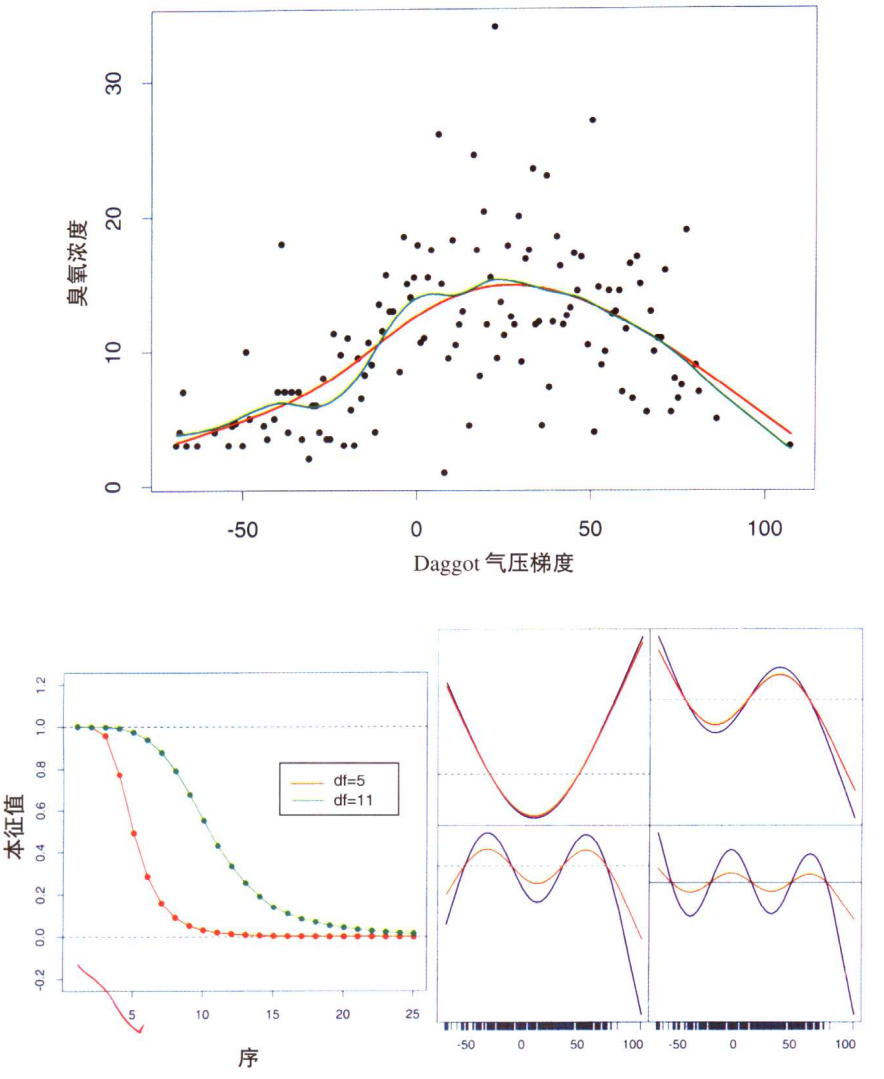


图 5.7

[上图]臭氧浓度作为Daggot气压梯度的函数的光滑样条拟合。两个拟合对应于光滑参数的不同值。光滑参数的选取是为了得到5个和15个由 $df_\lambda = \text{trace}(\mathbf{S}_\lambda)$定义的有效自由度。[左下图]两个光滑样条矩阵的前25个本征值。前两个本征值恰为1，并且所有的本征值都大于或等于零。[右下图]样条光滑子矩阵的第三个和第六个本征向量。在每条曲线中，$\mathbf{u}_k$ 都对照 $\mathbf{x}$ 绘制，并因此视为 x 的函数。图底部的底线指示数据点的出现。阻尼函数表示这些函数的光滑版本（使用5df光滑子）

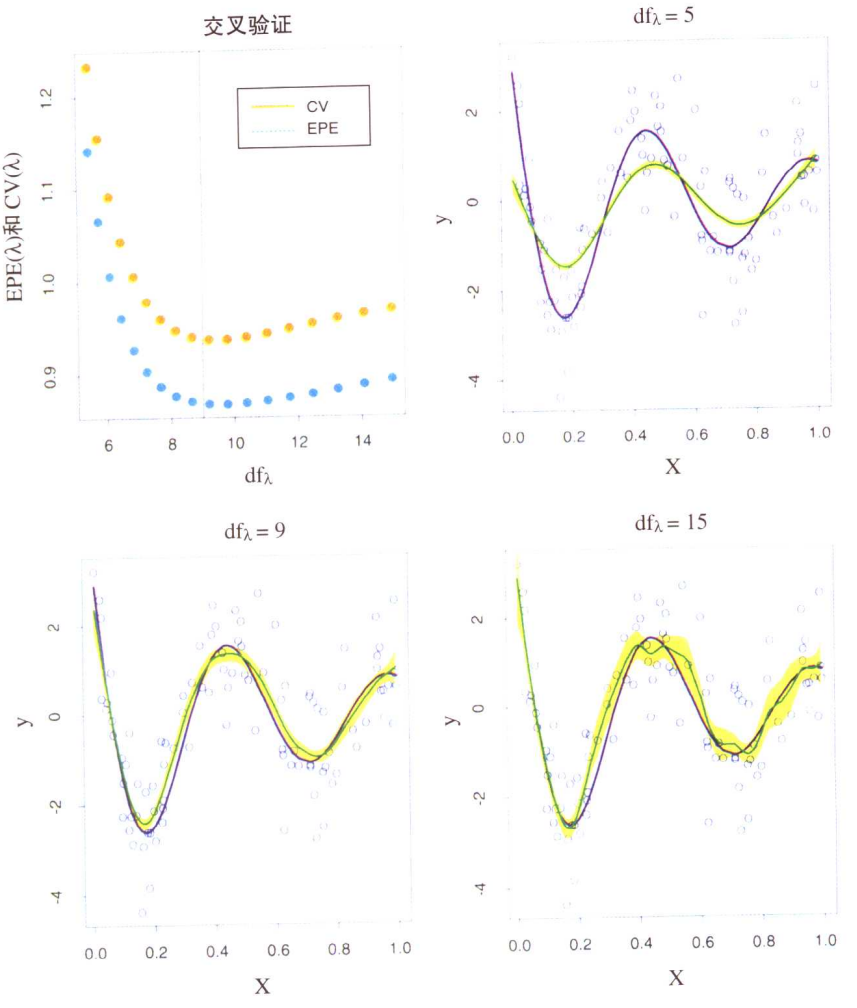


图 5.9

非线性加法误差模型（5.22）的实现，左上图显示其EPE(λ)和CV(λ)曲线。其他图对于不同的df_λ，显示数据、真实函数（紫色）和拟合曲线（绿色），其中黄色阴影是拟合曲线 ± 2 倍标准误差频带

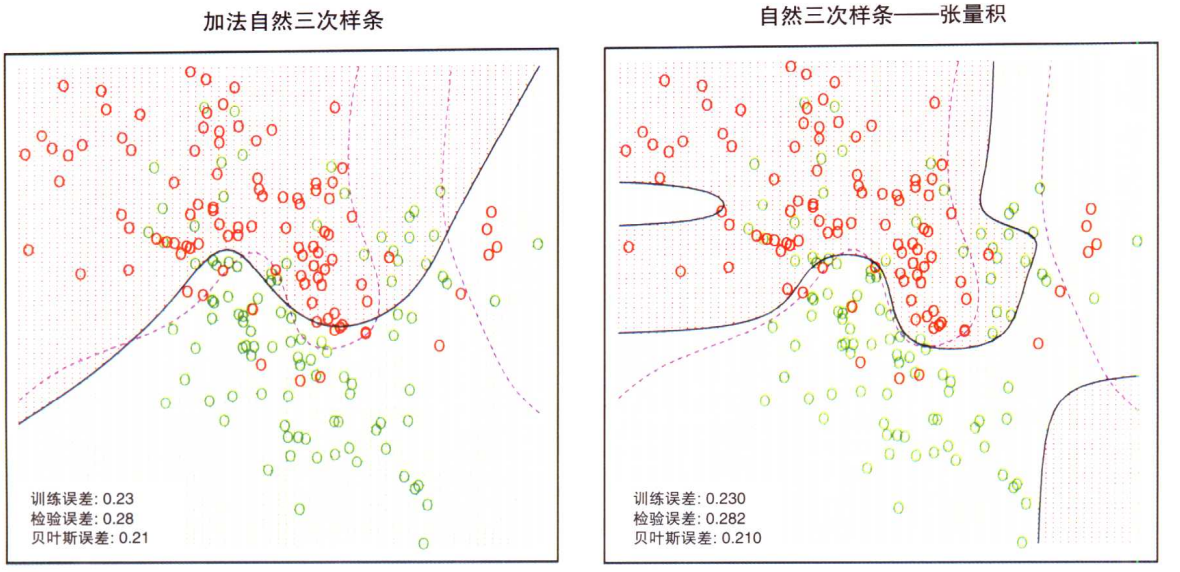


图 5.11

图 2.1 的模拟例子。左图显示加法逻辑斯缔回归模型的判定边界，在两个坐标上都使用自然样条（全部 df = 1 + (4 − 1) + (4 − 1) = 7）。右图显示在每个坐标上使用自然样条基张量积的结果（全部 df = 4 × 4 = 16）。紫色虚线边界是该问题的贝叶斯判定边界

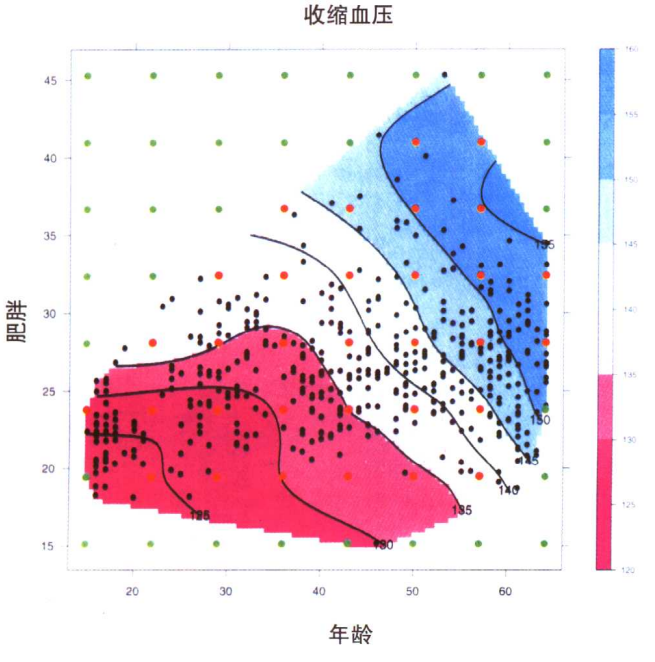


图 5.12

用围线图显示薄板样条拟合心脏病数据。响应是收缩血压（sbp），它作为年龄（age）和肥胖（obesity）的函数建模。图中标定有数据点以及用做纽结的点的格。谨慎使用来自数据凸包之内（红色）格的纽结，并忽略数据凸包之外（绿色）格的纽结

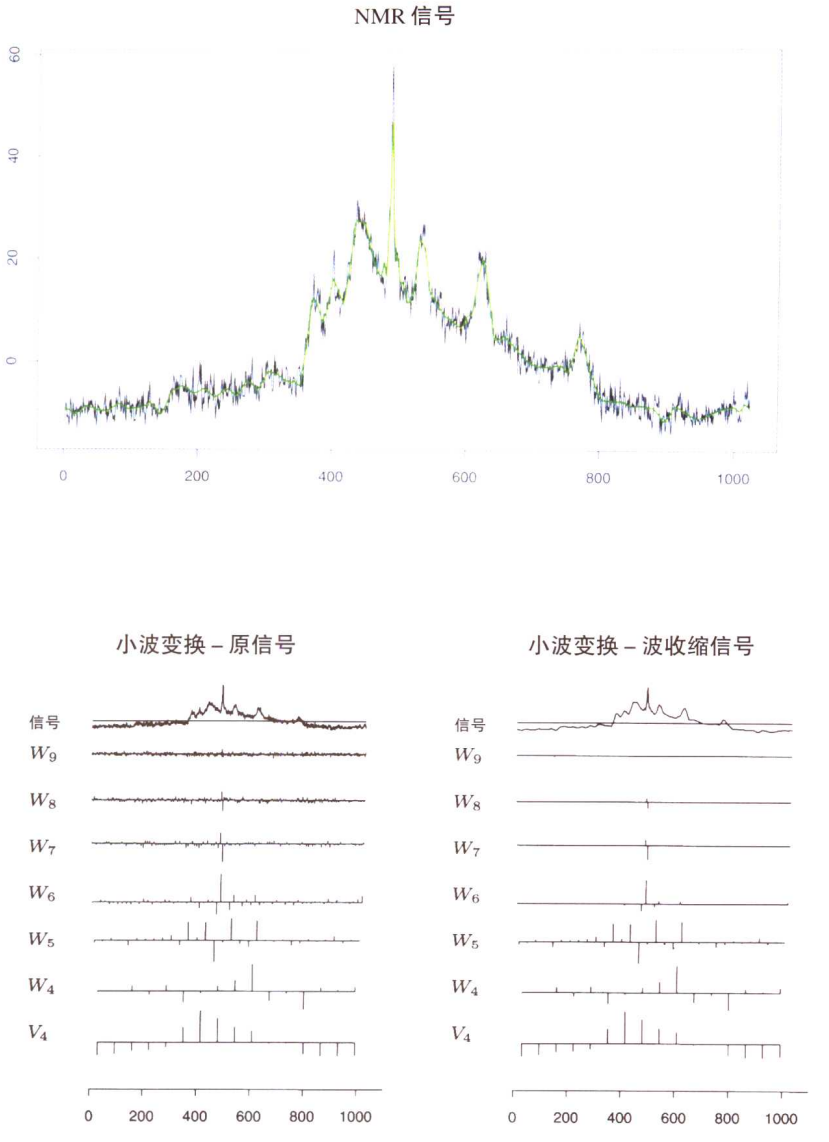


图 5.14

上图显示一个NMR信号，小波收缩版本以绿色叠加。左下图表示原信号的小波变换，使用symmlet-8基函数，取到V_4。每个系数用垂直条的高度（正的或负的）表示。右下图表示使用S-PLUS中的waveshrink函数收缩后的小波系数。waveshrink函数实现了Donoho和Johnstone的小波自适应SureShrink方法