HZ BOOKS
华章IT

CAMBRIDGE

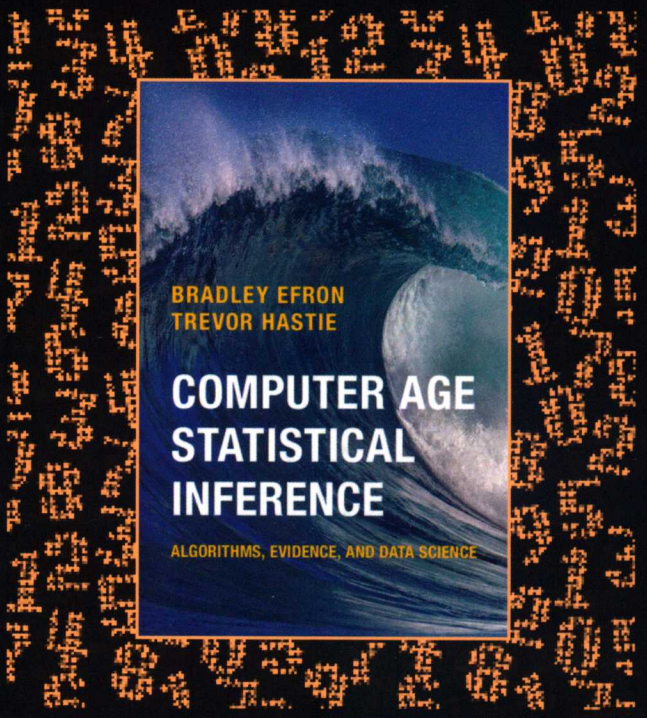
数据科学与工程技术丛书

计算机时代的统计推断

算法、演化和数据科学

[美] 布拉德利·埃夫隆（Bradley Efron） 斯坦福大学
特雷福·黑斯蒂（Trevor Hastie） 斯坦福大学 著

杭汉源 译



COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE

非外借

机械工业出版社
China Machine Press

数据科学与工程技术丛书

COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE

计算机时代的统计推断

算法、演化和数据科学

[美] 布拉德利·埃夫隆（Bradley Efron） 斯坦福大学
特雷福·黑斯蒂（Trevor Hastie） 斯坦福大学 著

杭汉源 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

计算机时代的统计推断：算法、演化和数据科学 /（美）布拉德利·埃夫隆（Bradley Efron），（美）特雷福·黑斯蒂（Trevor Hastie）著；杭汉源译 . —北京：机械工业出版社，2019.5
（数据科学与工程技术丛书）
书名原文：Computer Age Statistical Inference: Algorithms, Evidence, and Data Science

ISBN 978-7-111-62752-4

I. 计… II. ①布… ②特… ③杭… III. 计算机应用－统计推断 IV. 0212-39

中国版本图书馆 CIP 数据核字（2019）第 091087 号

本书版权登记号：图字 01-2017-8953

This is a Simplified-Chinese edition of the following title published by Cambridge University Press:

Bradley Efron, Trevor Hastie, Computer Age Statistical Inference: Algorithms, Evidence, and Data Science 978-1-107-14989-2.

© Bradley Efron and Trevor Hastie 2016.

This Simplified-Chinese edition for the People's Republic of China (excluding Hong Kong, Macau and Taiwan) is published by arrangement with the Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

© Cambridge University Press and China Machine Press in 2019.

This Simplified-Chinese edition is authorized for sale in the People's Republic of China (excluding Hong Kong, Macau and Taiwan) only. Unauthorized export of this simplified Chinese is a violation of the Copyright Act. No part of this publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of Cambridge University Press and China Machine Press.

本书原版由剑桥大学出版社出版。

本书简体字中文版由剑桥大学出版社与机械工业出版社合作出版。未经出版者预先书面许可，不得以任何方式复制或抄袭本书的任何部分。

此版本仅限在中华人民共和国境内（不包括香港、澳门特别行政区及台湾地区）销售。

本书以丰富的案例介绍了计算机时代下的统计推断的发展脉络，从理论的角度剖析统计推断的各类算法、证据等，揭示统计推断如何推动当今大数据、数据科学、机器学习等领域的快速发展并引领数据分析的变革，最后展望了统计学和数据科学的未来方向。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：陈佳媛　　责任校对：殷 虹
印　刷：北京瑞德印刷有限公司　　版　次：2019 年 6 月第 1 版第 1 次印刷
开　本：185mm × 260mm 1/16　　印　张：20.75（含 1.5 印张彩插）
书　号：ISBN 978-7-111-62752-4　　定　价：119.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88379833　　投稿热线：（010）88379604
购书热线：（010）68326294　　读者信箱：hzjsj@hzbook.com

版权所有 · 侵权必究
封底无防伪标均为盗版
本书法律顾问：北京大成律师事务所 韩光 / 邹晓东

表 8.2　细胞输注数据；人类细胞群落在 1 至 5 天内以五种比率灌注小鼠细胞核，并观察它们是否生长旺盛。绿色数据是逻辑回归模型的估计值 $\hat{\pi}_{ij}$。例如，在最低“比例 / 天数”类别中，31 个群落中的 5 个生长旺盛，观察比例为 5/31 = 0.16，逻辑回归估计值为 $\hat{\pi}_{11} = 0.11$

		时间				
		1	2	3	4	5
比例	1	5/31 0.11	3/28 0.25	20/45 0.42	24/47 0.54	29/35 0.75
	2	15/77 0.24	36/78 0.45	43/71 0.64	56/71 0.74	66/74 0.88
	3	48/126 0.38	68/116 0.62	145/171 0.77	98/119 0.85	114/129 0.93
	4	29/92 0.32	35/52 0.56	57/85 0.73	38/50 0.81	72/77 0.92
	5	11/53 0.18	20/52 0.37	20/48 0.55	40/55 0.67	52/61 0.84

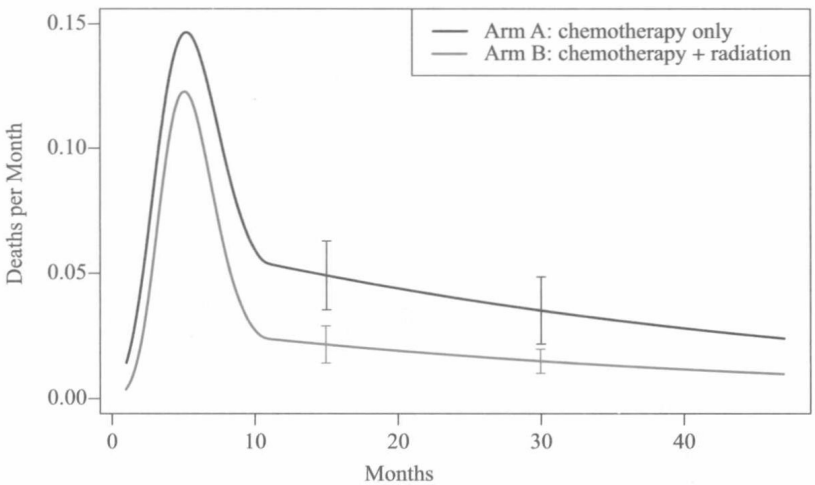


图 9.2　NCOG 研究的参数风险率估计。A 组的黑色曲线比 B 组在治疗后多于一年的所有时间内具有约 2.5 倍的风险。在 15 个月和 30 个月时标注出标准误差

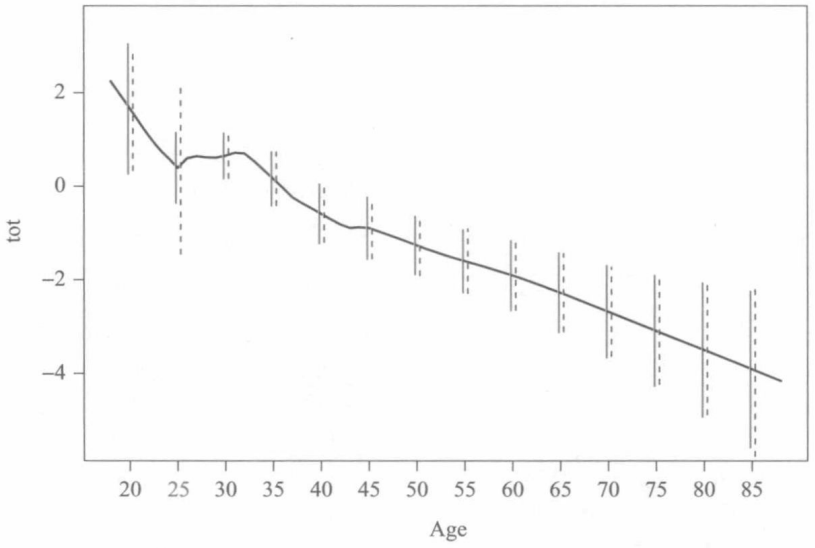


图 10.1　图 1.2 中肾脏数据的 lowess 曲线。竖直的线条显示 ±2 标准差：蓝色虚线为刀切法（10.6），红色实线为自助法（10.16）。刀切法在年龄为 25 时对于变化估计过高

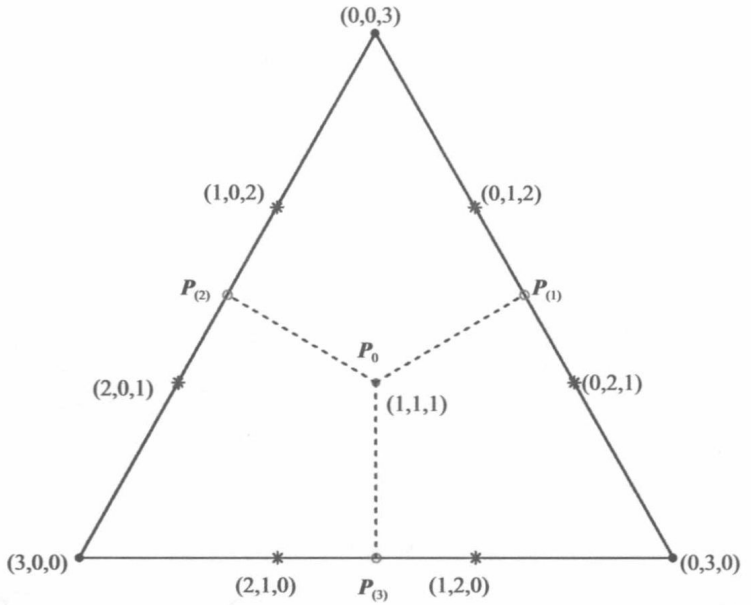


图 10.3　样本数 $n = 3$ 时的重抽样。中心点是 P_0（10.26）；绿圈为刀切法的点 $P_{(i)}$（10.28）；三对显示出自助重抽样数（N_1，N_2，N_3）（10.29）。自助法概率为，P_0 为 6/27，每个拐角点为 1/27，每个六角点为 3/27

试读结束：需要全本请在线购买：　www.ertongbook.com

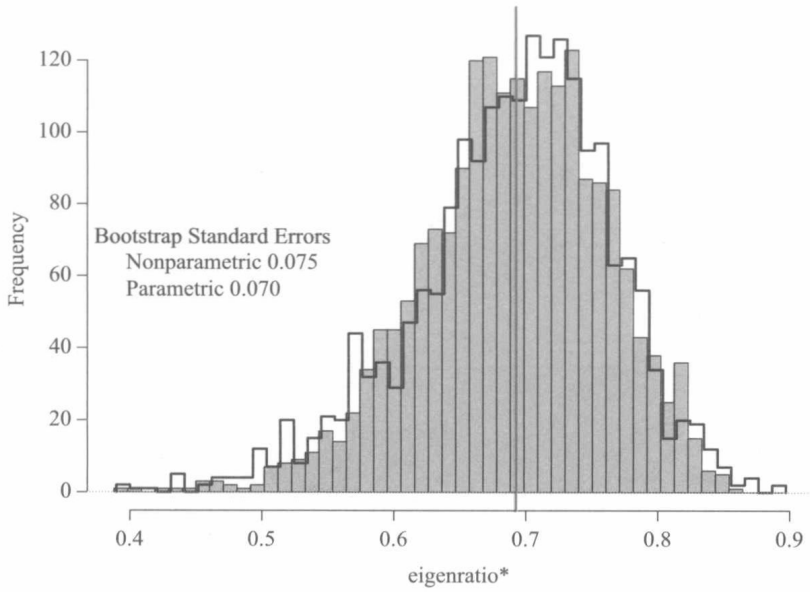


图 10.6　特征比的例子，学生分数数据。实线直方图对应 $B = 2000$ 来自五维正态 MLE 的参数自助法复制$\hat{\theta}$；细线直方图对应 2000 次图 10.2 的非参数复制。MLE $\hat{\theta} = 0.693$ 时竖直的红线

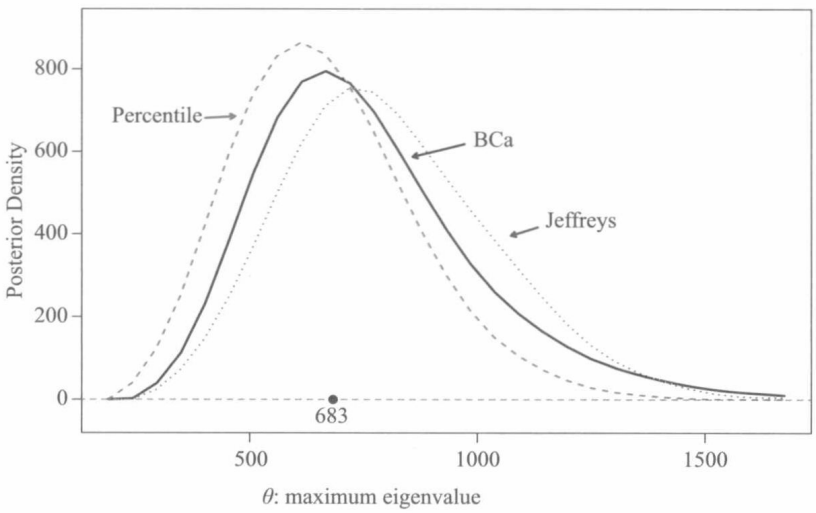


图 11.7　最大特征值参数（11.71）的置信密度，使用多元正态模型（11.70）作为学生评分数据。红色虚线为百分位数法，BCa 为纯黑（其中 $(z_0, a) = (0.178, 0.093)$）。蓝色虚线表示使用 Jeffreys 先验（11.72）的 θ 的贝叶斯后验密度

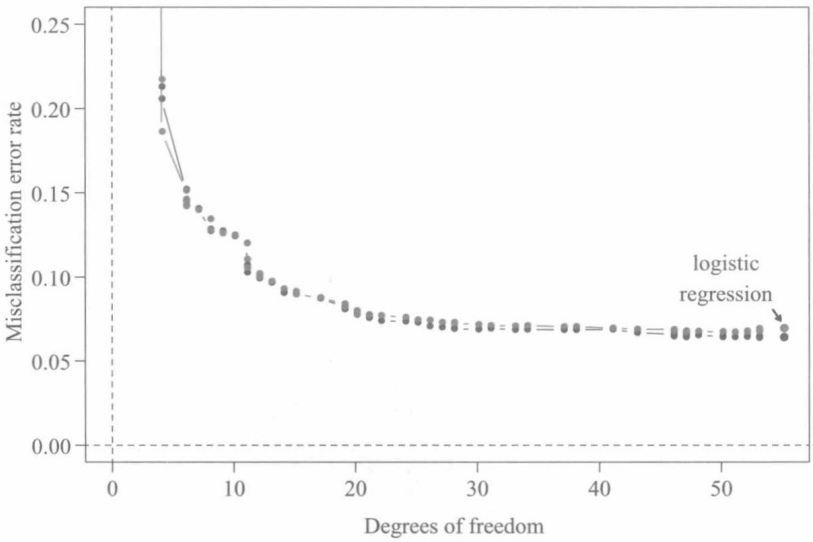


图 12.2　垃圾邮件数据。表观误差为 err（蓝色），由 glmnet 生产的一列预测规则的交叉验证估计（红色）。自由度为非零回归系数的个数：df = 57 即为一般的逻辑回归，相应的表观误差为 0.064，交叉验证率为 0.069。最小交叉验证错误率为 0.067

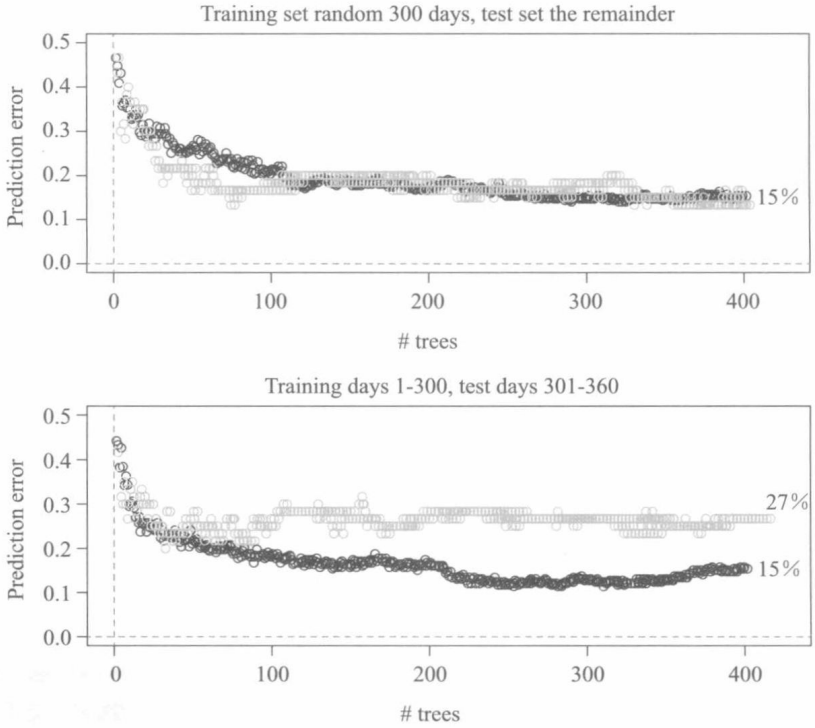


图 12.6　使用虚拟药物研究（12.81）和（12.82）的随机森林预测，测试误差（蓝点）与交叉验证训练误差（黑点）。顶部：随机选择 300 天的训练集，剩余的 60 天作为测试集。底部：前 300 天作为训练集，后 60 天作为测试集

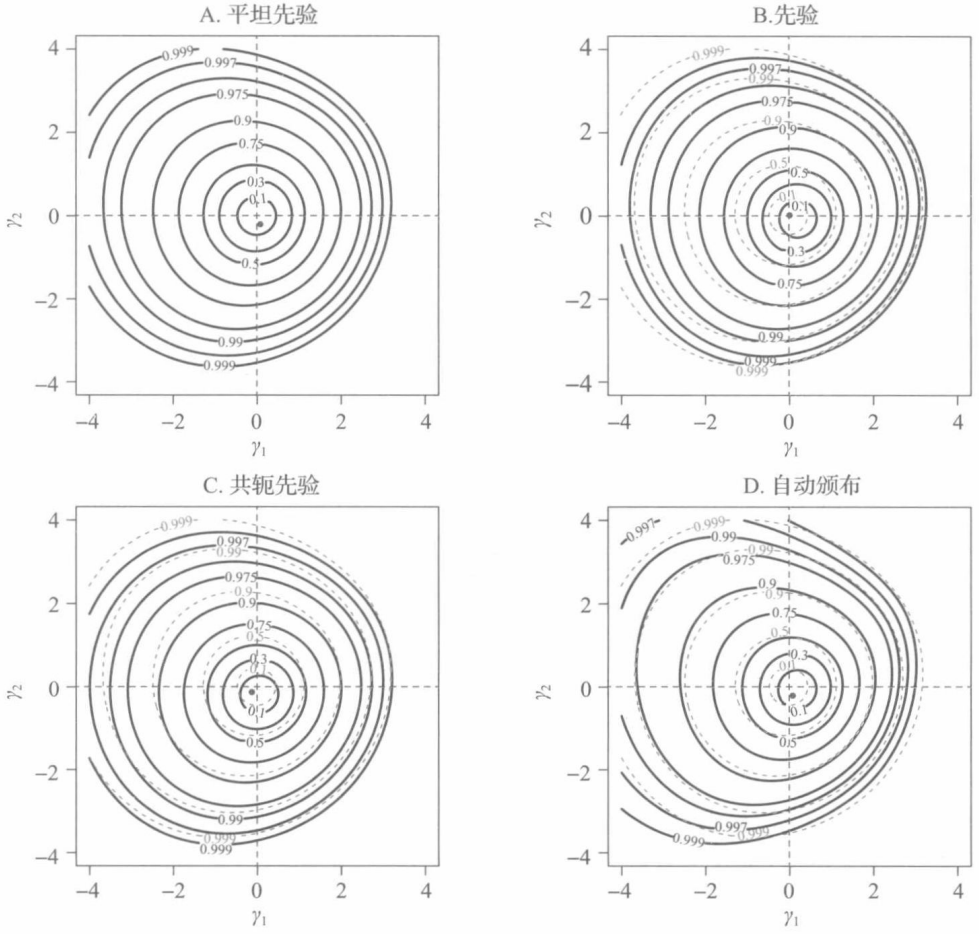


图 13.2　血管收缩数据：如文中所述，来自四个无信息先验的 γ 的相等后验密度（13.27）的等高线。数字表示等高线内的概率；*A* 中的浅虚等高线是平坦先验

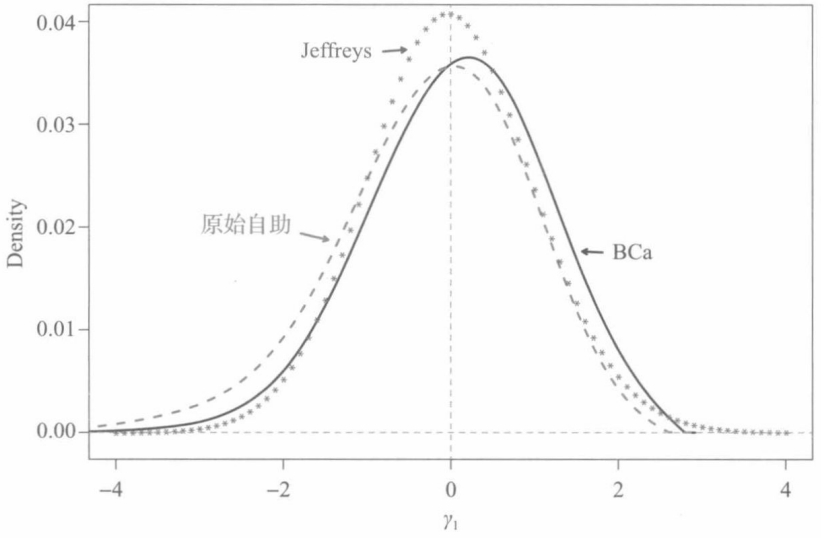


图 13.3　γ 在（13.27）的第一个坐标 γ_1 的血管收缩数据的后验密度。红色曲线：来自模型（13.24）和（13.25）的 $B = 8000$ 参数重复的原始（未加权）分布；实心黑色曲线：BCa 密度（11.68）（$z_0 = 0.123$，$a = 0.053$）；虚线蓝色曲线：使用 Jeffreys 多参数先验的后验密度（11.72）

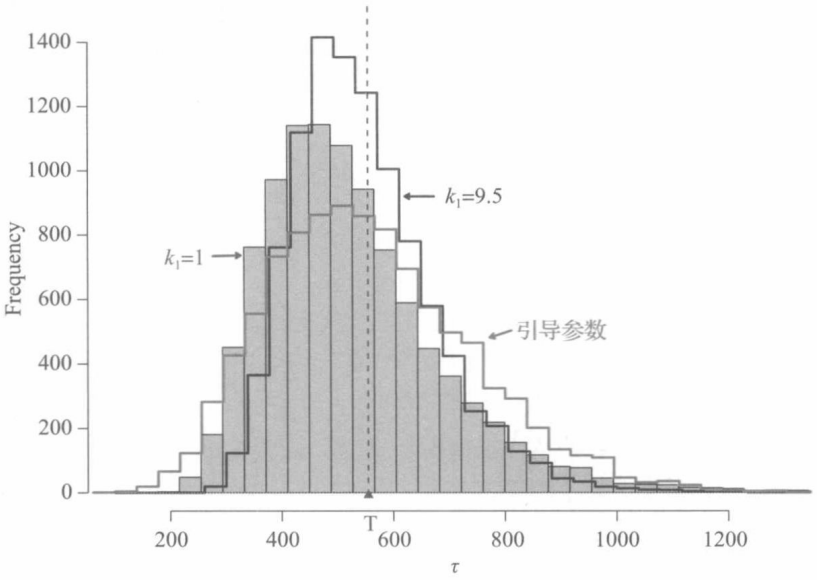


图 13.5　方差参数的后验分布，模型（13.63）至（13.65），来自表 13.2 的血管收缩 $y = 1$ 组的空气体积。实心蓝绿色直方图：$B = 10\ 000$ 个具有 $k_1 = 1$ 的 Gibbs 样本；黑线直方图：$B = 10\ 000$ 个样本，其中 $k_1 = 9.5$；红线直方图：10 000 个参数自助样本（13.72）表明即使 $k_1 = 1$ 先验也有相当大的后验效应

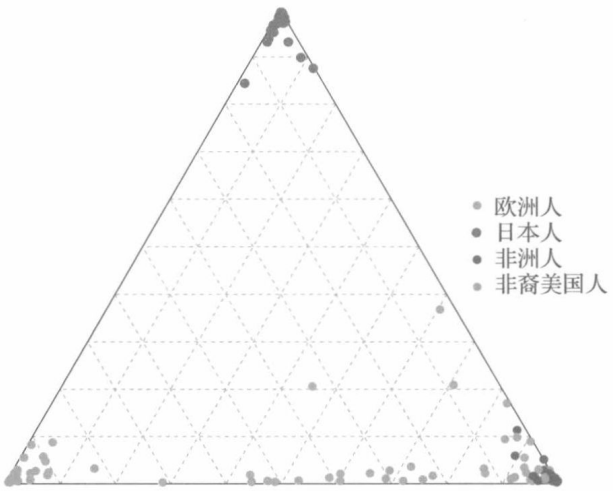


图 13.7　基于 MCMC 采样估计的 Q_i 的后验均值的重心坐标图

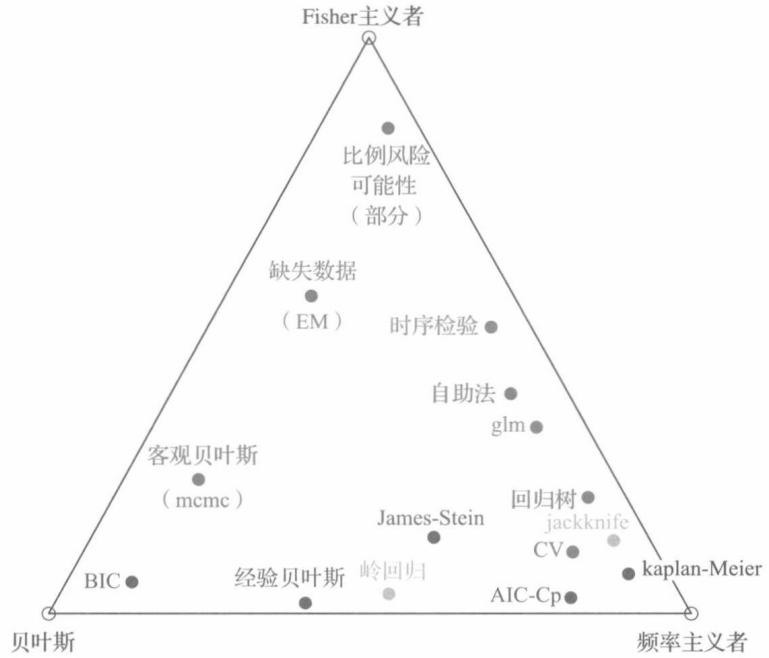


图 14.1　正如文中所描述的那样，贝叶斯、频率主义者和 Fisher 主义者影响了 20 世纪 50 年代到 90 年代的 15 个主要方法。颜色表明电子计算在其发展中的重要性：红色，至关重要；紫罗兰，非常重要；绿色，重要；淡蓝色，不太重要；蓝色，可以忽略不计

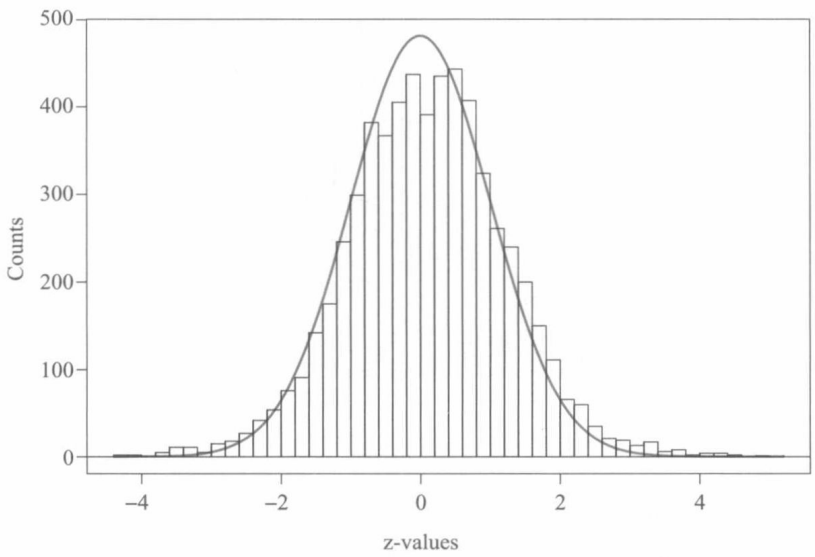


图 15.1　$N = 6033$ 的 z 值的柱状图，每一个都来自前列腺癌研究中的一个基因。如果所有基因都符合原假设（15.3），则直方图将贴近红色曲线。对于哪些基因我们可以拒绝原假设

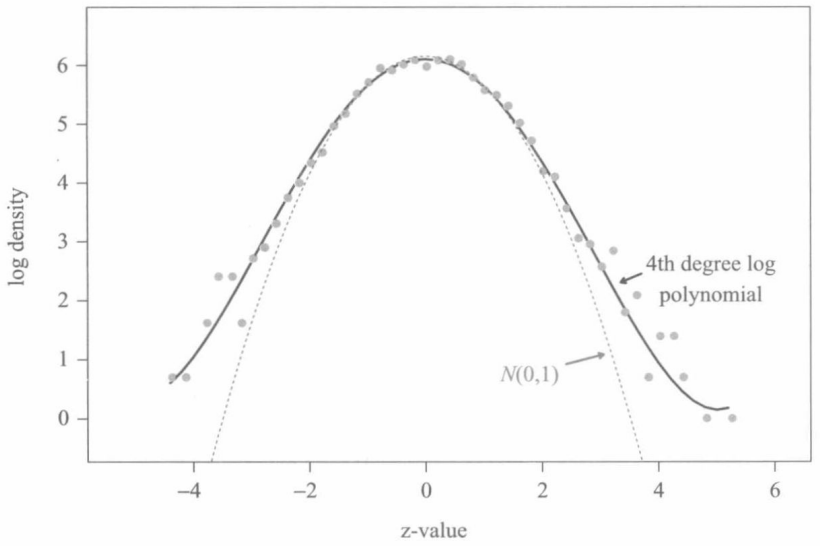


图 15.6　点是图 15.1 的直方图的对数箱子计数。纯黑色曲线是用于计算图 15.5 中的 fdr（z）的四阶对数多项式拟合。虚线的红色曲线，对数原假设密度（15.41）在 $|z| \leqslant 2$ 时拟合得很好

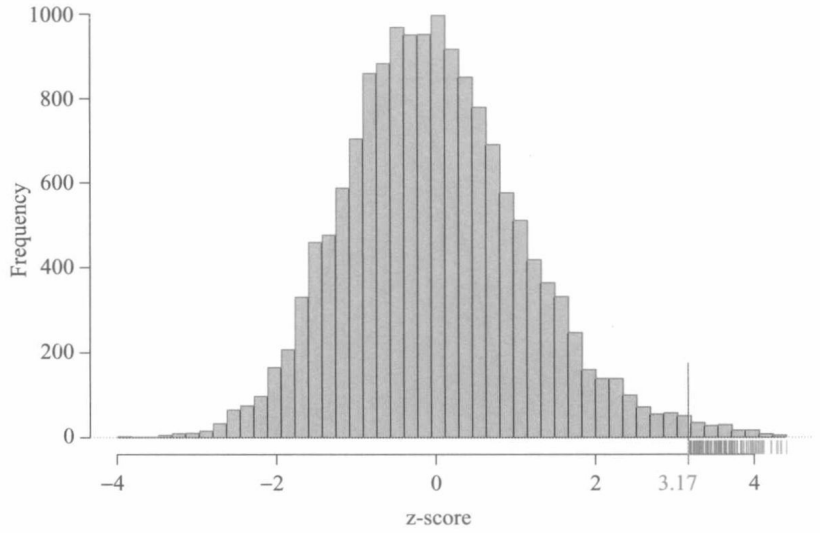


图 15.9　DTI 研究的 z 分数直方图，比较阅读障碍儿童和 15 443 个脑部位置的正常对照者。基于经验零分布的 FDR 分析给出了满足 fdr（z_i）$\leqslant 0.20$ 的 149 个体素。那些有 $z_i \geqslant 3.17$ 的情形（用红色虚线表示）

试读结束：需要全本请在线购买：　www.ertongbook.com

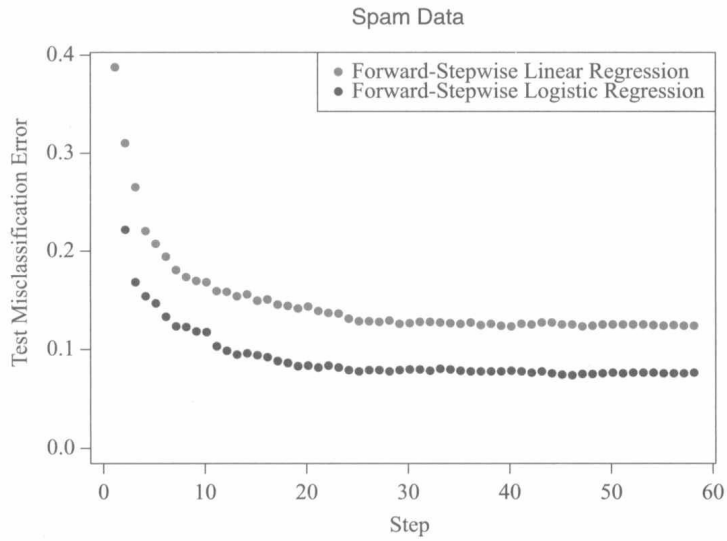


图 16.2　垃圾邮件数据的前向逐步回归。显示的是测试数据上的分类错误，作为步骤数量的函数。棕色点对应于线性回归，其中响应编码为 –1 和 +1; 大于零的预测被分类为 +1，小于零的分类为 –1。蓝点对应于逻辑回归，表现更好。我们看到这两条曲线基本上在 25 步后达到最小值

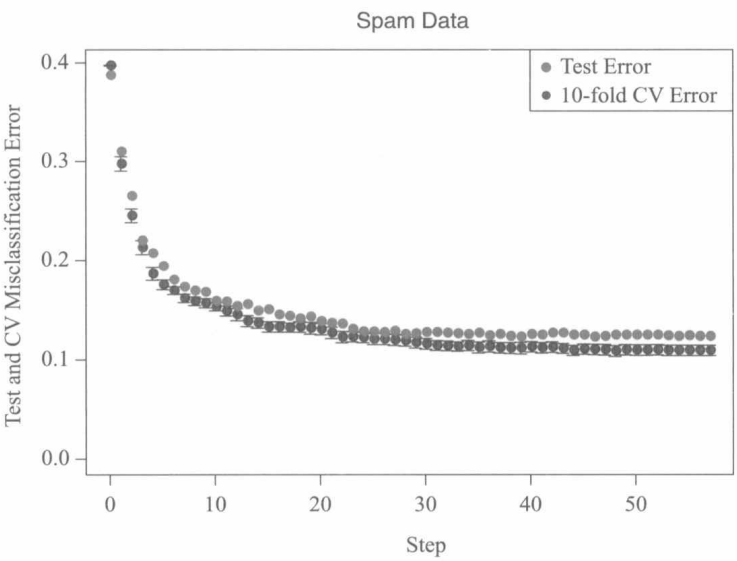


图 16.3　作为步骤的函数，对垃圾数据进行前向逐步回归的十折交叉验证分类错误（绿色）。由于每个误差为 10 个数字的平均，我们可以计算一个标准误差；包含在图中的是标准误差带。棕色曲线是测试数据中的分类错误

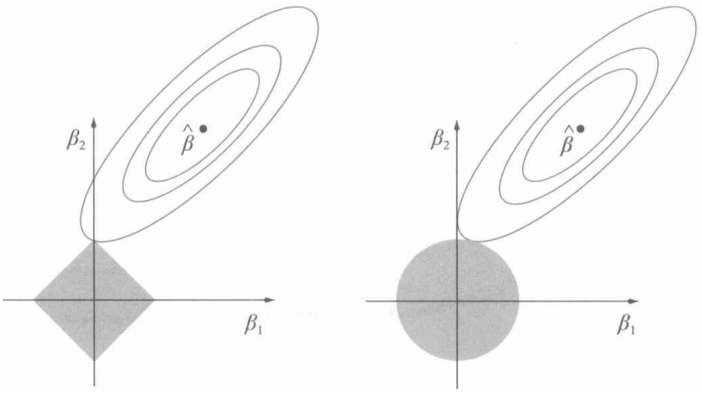


图 16.4　用 $\beta \in \mathbb{R}^2$ 来举例说明岭回归和套索之间的区别。在这两幅图中，红色等高线与平方误差损失函数相对应，无限制最小二乘估计 $\hat{\beta}$ 在中心。蓝色区域显示约束条件，左边是套索和右边是岭估计。约束问题的解答对应于扩大损失等值线首先接触约束区域的值。由于套索约束的形状，这通常会在角落（或更一般地说是边缘）处，这意味着在这种情况下，最小化 β 具有 $\beta_1 = 0$。对于岭估计，这不太可能发生

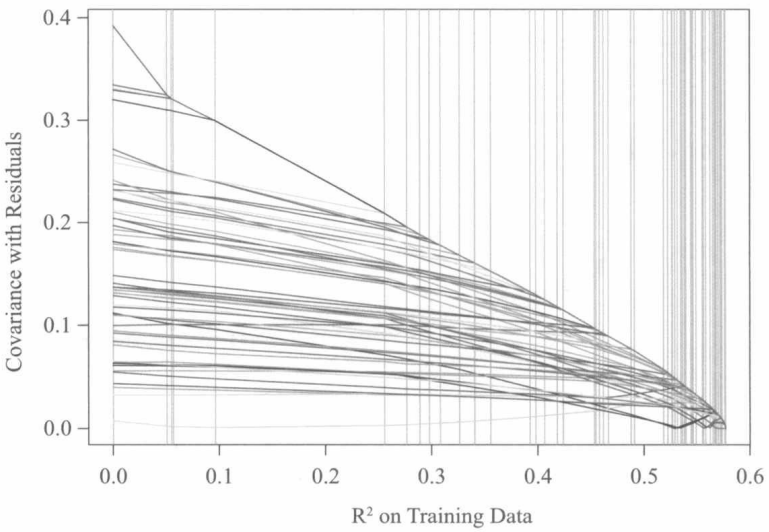


图 16.7　垃圾邮件数据的协方差演变。由于变量与最大协方差相关，它们成为有效集的一部分。这些时间由竖直的灰色条表示再次按照图 16.5 对训练 R^2 进行绘制

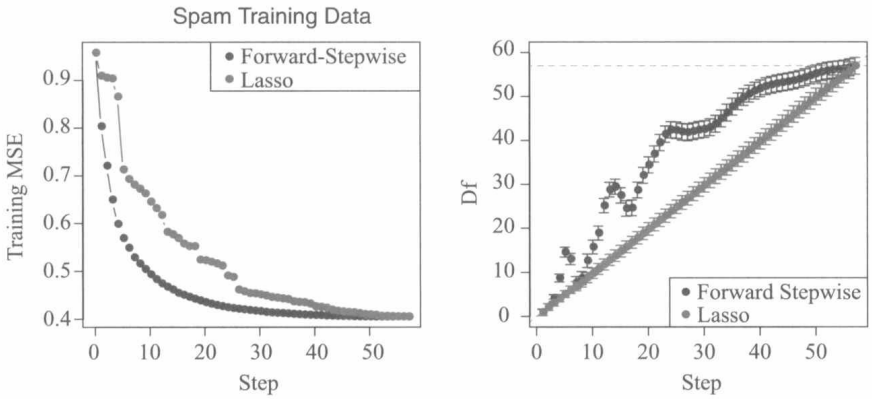


图 16.8　左图：对于前向逐步回归和套索，对垃圾数据的训练均方差（MSE）作为有效集大小的函数。前向逐步比套索更具侵略性，因为它（过度）更快地适应训练数据。右图：模拟显示前向逐步回归与套索的自由度。套索每步使用一个自由度，而前向逐步更贪婪，并且使用更多自由度，特别是在早期阶段。由于这些自由度是使用 5000 个随机模拟数据集计算的，因此我们在估计中包含标准误差带

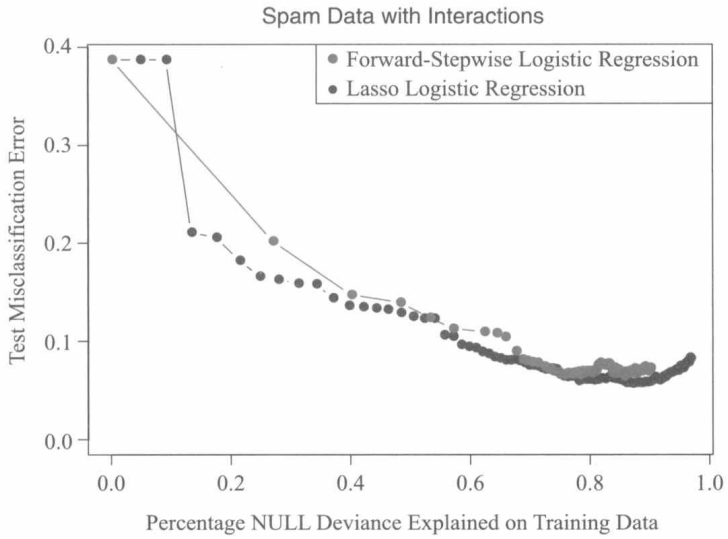


图 16.9　测试套索对分类误差与垃圾数据的前向逐步逻辑回归，我们在这里考虑了成对相互作用和主效应（总共 3061 个预测因子）。这里套索的最小误差为 0.057，逐步逻辑回归的为 0.064，而 0.071 为仅考虑主效应的套索逻辑回归模型。逐步模型在遇到收敛问题之前达到了 134 个变量，而套索的最大有效集为 682 个

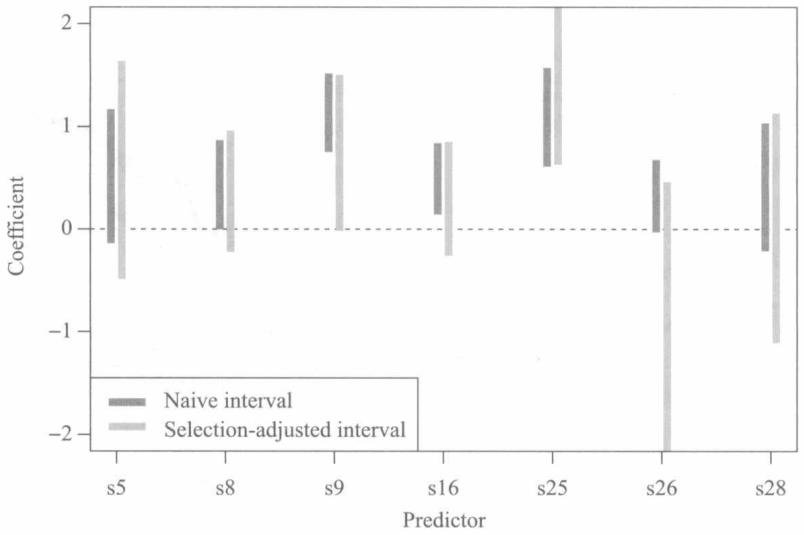


图 16.10　艾滋病毒数据。HIV 阳性患者 7 个位点的耐药性线性回归，特定基因组位置的突变指标。这 7 个地点从总共 30 名候选人中选出，使用套索。简单的 95% 置信区间（黑暗）使用标准线性回归推断，忽略选择事件。明亮间隔为 95% 置信区间，使用线性回归，但是以选择事件为条件

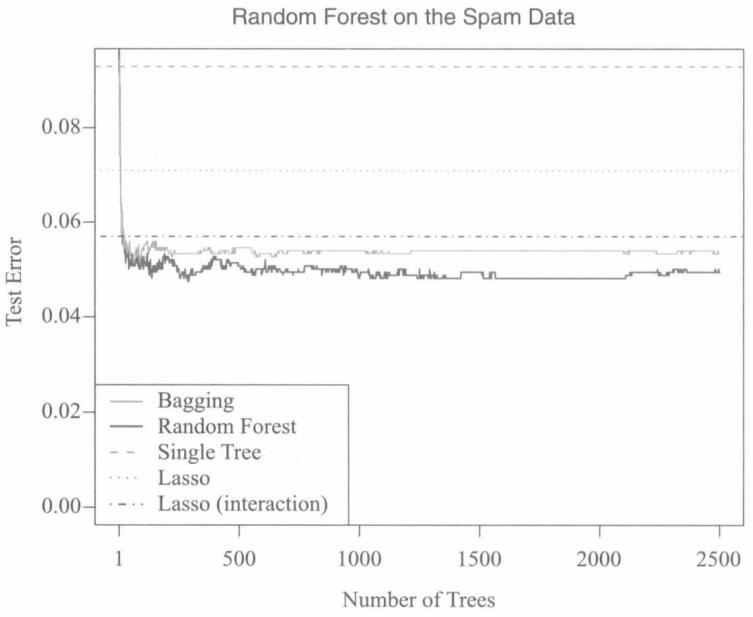


图 17.2　垃圾邮件数据上随机森林的测试误分类误差，作为树的函数红色曲线随机选择 $p = 57$ 个特征的 $m = 7$ 作为分裂变量的候选项，每次进行分裂。蓝色曲线使用 $m = 57$，因此相当于装袋。装袋和随机森林都胜过套索方法和单一树

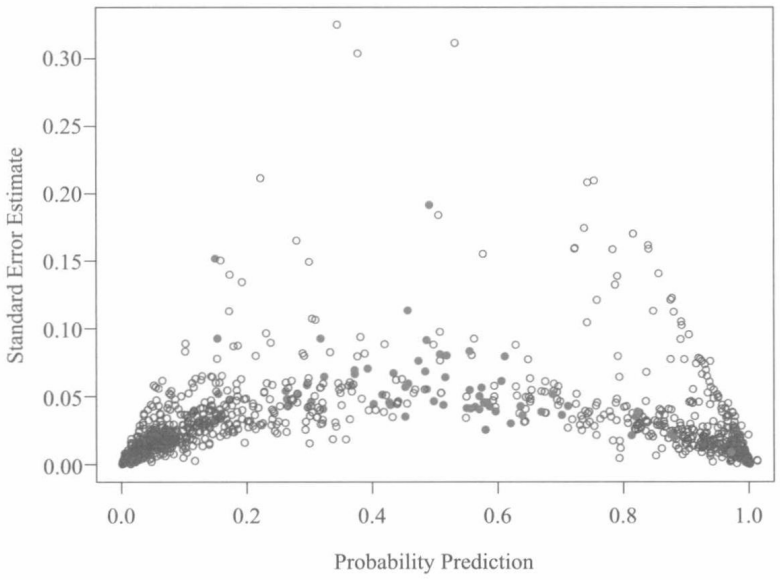


图 17.4　垃圾邮件测试数据中概率估计的刀切标准误差估计（包含偏差修正）。标记为红色的点是错误分类，并倾向于集中在决策边界附近（0.5）

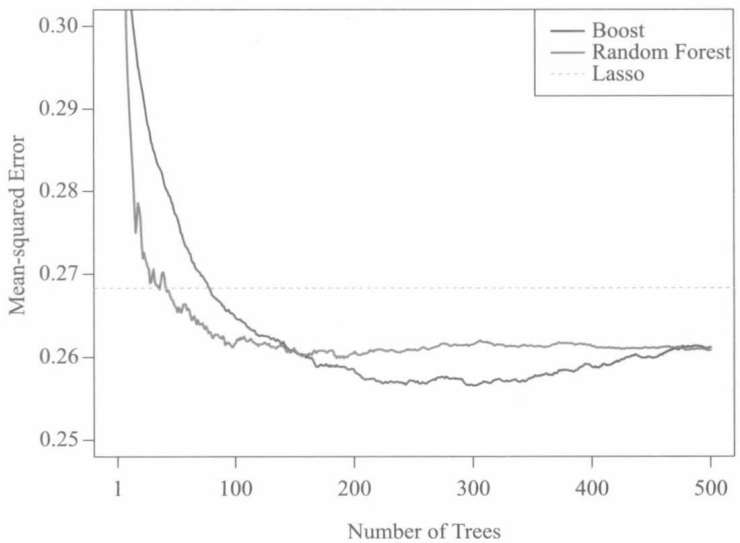


图 17.6　拟合 ALS 训练数据的增强回归树模型的测试性能，其中 $n = 1197$ 和 $p = 369$。所示为 625 个指定测试观测值上的均方根误差，作为树的个数的函数。这里深度 $d = 4$ 且 $\epsilon = 0.02$。与随机森林相比，提升可以实现更低的测试 MSE。我们看到，随着树的数目 B 变大，提升的测试误差开始增加——这是过度拟合的结果。随机森林不会过拟合。点的蓝色水平线显示线性模型的最佳性能，由套索拟合。由于垂直尺度不会延伸到零，所以差异不如它们看起来那么剧烈

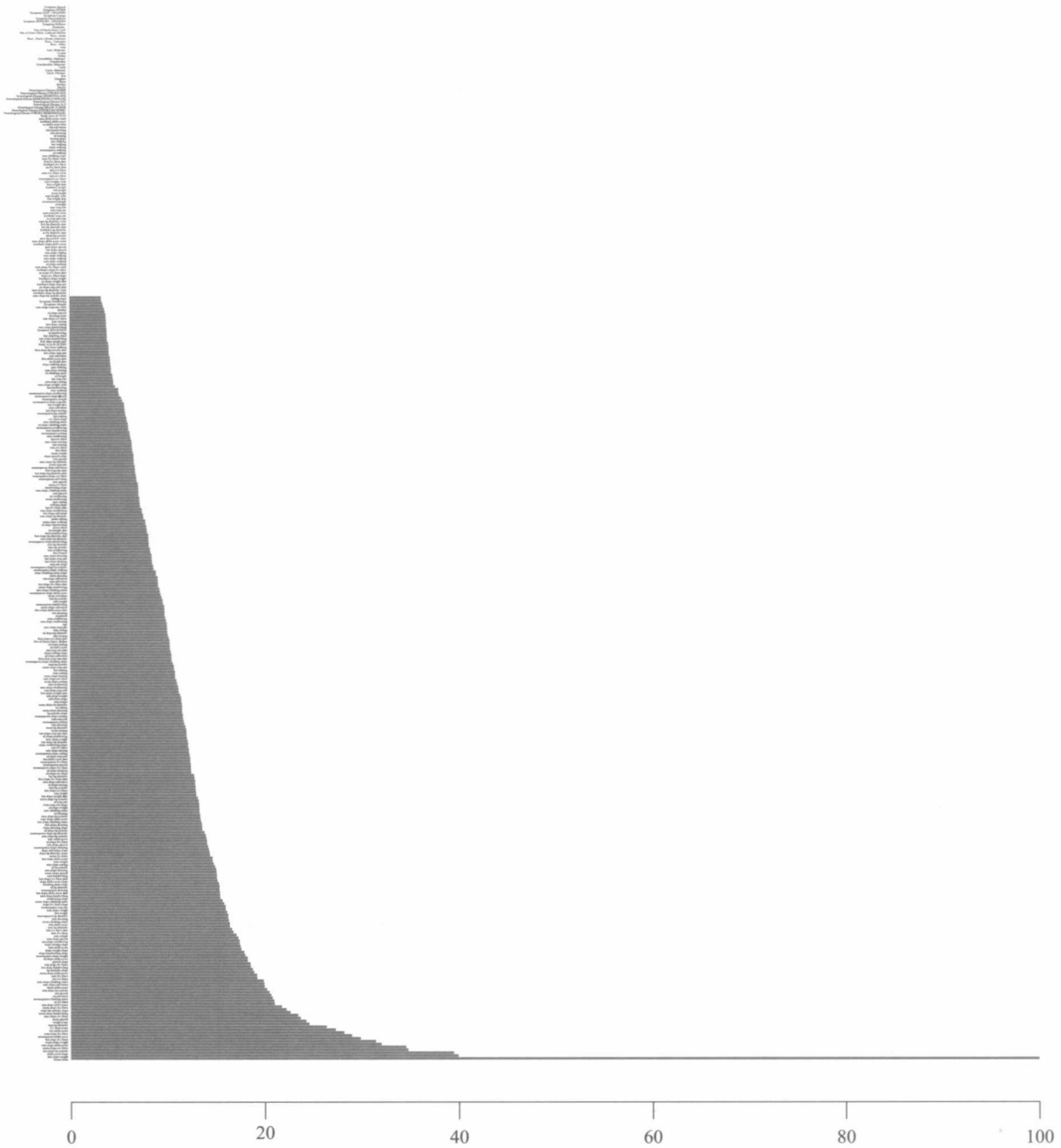


图 17.7　ALS 数据的变量重要性图。这 369 个变量中的 267 个被用于集成中。标签显示的变量太多，因此该图可作为视觉指南。变量 Onset.Delta 具有相对重要性 100（最低的红色栏），大约 40 左右，是接下来的两个变量（last.slope.weight 和 alsfrs.score.slope）的两倍以上。然而，重要性缓慢下降，这表明该模型需要大部分的变量

试读结束：需要全本请在线购买：　www.ertongbook.com