2025~2026 学年 春 季学期 多元统计分析 课程结课报告

学号：2024017349

姓名：李倩倩

中国石油大学 (北京) 克拉玛依校区 文理学院 数学与统计系

Sunday 7th June, 2026

摘要

结课报告分为两部分：

第一部分：课程内容总结 对课程所学内容进行系统、完整的总结，包括基本概念、基本理论、基本模型、基本结论、模型拟合的基本方法、用 R 如何实现以及实施的主要步骤和注意事项等. 第一部分满分 30 分.

第二部分：数据分析实战 对所给数据集应用所学方法完成分析任务. 第二部分满分 70 分.

请同学们各自独立完成结课报告，所有内容用中文书写，生成的报告应内容正确、排版规范、格式美观. 凡抄袭、雷同者，结课报告成绩记零分.

请用自己的“学号-姓名”建立一个文件夹，将你的结课报告 LaTeX 文件、生成的 pdf 文件、报告中用到的图形、数据文件等放入其中，同时将建模过程中使用到的全部 R 代码生成一个“.R”后缀的代码文件，一并放入该文件夹中，以便验证代码的运行结果. 然后将该文件夹压缩成以“学号-姓名”命名的一个文件，发送至 xiaolei@cup.edu.cn 邮箱.

截止日期：2026 年 6 月 7 日 24:00.

完成时间： Sunday 7th June, 2026 15:46

1 课程内容总结

通过本学期的学习，我对多元统计分析这门课的理解是：它不是把一元统计方法简单重复很多遍，而是在同一个研究对象上同时分析多个变量的位置、离散程度、相关结构、潜在结构和分类结构。为了把知识点整理得更清楚，我把所学内容概括为三条主线：第一条是多元数据的描述、矩阵表示和多元分布理论；第二条是估计、检验和数据矩阵分解的统计基础；第三条是主成分分析、因子分析、聚类分析、判别分析和对应分析等具体多元方法。多元数据通常可以写成

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

其中 n 为观测数，p 为变量数。我认为实际分析中比较稳妥的流程是：

(1) 明确变量类型和分析目标；

(2) 检查缺失值、重复值、异常值和类别比例；

(3) 根据方法需要进行中心化、标准化、变量变换或类别编码；

(4) 选择合适的距离、相似度、降维方法或概率模型；

(5) 拟合模型，并结合图形解释结果；

(6) 用误判率、方差贡献率、公共度、聚类稳定性或检验结果评价模型效果。

1.1 描述性统计、数据矩阵与可视化

描述性分析是所有建模之前的第一步。对多元数据，最基本的样本统计量是样本均值向量、样本协方差矩阵和样本相关矩阵：

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i, \qquad S=\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})', \qquad R=(r_{jk})_{p\times p}.$$

协方差度量两个变量的线性同变动方向和强度，但受量纲影响；相关系数消除了量纲影响，适合比较不同单位的变量。通过这些内容我认识到，多元数据分析不能只看单变量均值，还要借助箱线图、直方图、核密度估计、散点图、散点图矩阵、三维散点图、Andrews 曲线、Chernoff faces 和平行坐标图等工具，同时观察偏态、厚尾、异常值、类间差异和变量间相关结构。

一维核密度估计的基本形式为

$$\hat{f}_h(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right),$$

其中 K 为核函数，h 为窗宽。窗宽过小会产生很多局部波动，窗宽过大会过度平滑。

R 实现：常用 `summary()`、`boxplot()`、`hist()`、`density()`、`pairs()`、`plot()`、`cor()`、`cov()` 和 `scale()`；若使用扩展包，可以用 `ggplot2` 画箱线图、核密度和散点图矩阵。

注意事项：

- 分类变量不应机械标准化。
- 极端异常值会明显影响均值、协方差以及后续 PCA 或聚类结果。
- 量纲差异较大的连续变量通常需要先标准化。

1.2 矩阵代数、距离与几何观点

多元统计的模型多数可以写成矩阵形式，因此矩阵的逆、广义逆、秩、迹、行列式、二次型、谱分解和奇异值分解都是基础工具。若 S 为实对称非负定矩阵，则存在正交矩阵 P 和特征值对角矩阵 Λ，使得

$$S = P\Lambda P', \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p).$$

这一结论直接支撑主成分分析、因子分析和对应分析中的矩阵分解。二次型

$$Q(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})' A (\boldsymbol{x} - \boldsymbol{\mu})$$

决定了椭球等距曲面的形状；当 $A = \Sigma^{-1}$ 时得到马氏距离。

距离和相似度是聚类、判别和多维标度的核心。常见距离包括欧氏距离、标准化欧氏距离、曼哈顿距离和马氏距离：

$$d_M^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})' S^{-1} (\boldsymbol{x} - \boldsymbol{y}).$$

马氏距离考虑了变量间相关性和方差差异，适合协方差结构明显的数据。

R 实现：

- 矩阵乘法：`%*%`。
- 求逆：`solve()`。
- 特征分解：`eigen()`。
- 奇异值分解：`svd()`。
- 距离矩阵：`dist()`。

注意事项：

- 使用马氏距离或判别分析前，要检查协方差矩阵是否可逆。
- 变量高度共线或 p 接近 n 时，直接求逆可能不稳定。

1.3 多元随机变量、多元分布、估计与检验

多元随机变量 $\boldsymbol{X} = (X_1, \ldots, X_p)'$ 的分布由联合分布函数或联合密度函数刻画。我认为这里最重要的理论模型是多元正态分布：

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma), \qquad f(\boldsymbol{x}) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

其基本结论包括：任意线性变换仍为正态；边缘分布仍为正态；在多元正态条件下，不相关等价于独立；条件分布仍为正态。这些结论使得线性组合、椭球置信域、判别函数和 Hotelling T^2 统计量具有清楚的理论基础。

估计理论中，$\boldsymbol{\mu}$ 的自然估计为 $\bar{\boldsymbol{x}}$，协方差矩阵的无偏估计为 S，极大似然估计为

$$\hat{\Sigma}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'.$$

假设检验部分以似然比检验为一般框架，并给出多元均值检验和同时置信区间等结论。一个典型统计量是

$$T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'S^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0),$$

在多元正态总体下可转化为 F 分布进行检验。

R 实现： 可用 `cov()`、`cor()`、`lm()`、`anova()` 以及手工矩阵运算完成估计和检验；若需要模拟多元正态样本，可用 `MASS::mvrnorm()`。

注意事项：

- 这类检验依赖正态性、独立性和协方差矩阵可逆性。
- 高维小样本或强共线数据下，传统检验可能失真。

1.4 数据矩阵分解与主成分分析

主成分分析的目标是把 p 个相关变量转换为少数互不相关的综合变量，并尽可能保留总变异。若使用标准化后的变量 $\boldsymbol{Z}$ 和相关矩阵 R，令

$$R\boldsymbol{a}_j = \lambda_j \boldsymbol{a}_j, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p,$$

则第 j 个主成分为

$$Y_j = \boldsymbol{a}_j'\boldsymbol{Z}, \qquad \mathrm{Var}(Y_j) = \lambda_j, \qquad \mathrm{Cov}(Y_i, Y_j) = 0 \ (i \neq j).$$

基本结论是：第一个主成分解释最大方差，第二个主成分在与第一个主成分正交的条件下解释剩余最大方差；所有主成分方差之和等于总方差；在标准化 PCA 中，总方差为 p，第 j 个主成分贡献率为 $\lambda_j / \sum_{l=1}^{p}\lambda_l$。

模型拟合方法包括直接对协方差矩阵或相关矩阵做谱分解，也可以用奇异值分解从数据矩阵出发。R 中常用

```
prcomp(x, center = TRUE, scale. = TRUE)
```

或 `princomp(x, cor = TRUE, scores = TRUE)`。

实施步骤：

(1) 判断是否需要标准化，量纲不同或方差差异较大时优先使用相关矩阵。

(2) 计算主成分标准差、载荷和得分。

(3) 画碎石图，并结合特征值大于 1、累计贡献率和解释性确定主成分个数。

(4) 用变量投影图解释每个主成分的含义。

(5) 用观测得分图观察样本在低维空间中的结构。

注意事项：

- 主成分符号可整体反号，解释方向时不影响结论。
- PCA 是无监督降维方法，不保证对分类最优。
- 若变量量纲差异很大，应优先使用标准化后的相关矩阵。

1.5 因子分析

因子分析把多个可观测变量之间的相关结构解释为少数潜在公共因子的作用。正交因子模型写作

$$\boldsymbol{X} - \boldsymbol{\mu} = A\boldsymbol{F} + \varepsilon,$$

其中 $A = (a_{ij})$ 为因子载荷矩阵，$\boldsymbol{F}$ 为公共因子，ε 为特殊因子，并通常假设

$$E(\boldsymbol{F}) = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{F}) = I, \quad \mathrm{Cov}(\varepsilon) = \Psi, \quad \mathrm{Cov}(\boldsymbol{F}, \varepsilon) = 0.$$

于是协方差矩阵可分解为

$$\Sigma = AA' + \Psi.$$

第 i 个变量的公共度为 $h_i^2 = \sum_{j=1}^{m} a_{ij}^2$，表示该变量方差中由公共因子解释的比例；特殊方差为 ψ_i，表示未被公共因子解释的部分。

因子分析的拟合方法包括主成分法、主因子法和极大似然法。实际实现时，既可以从相关矩阵谱分解出发估计载荷，也可以直接使用 R 函数 `factanal()` 做极大似然因子分析。R 中可用 `factanal(x, factors = m, rotation = "varimax", scores = "regression")`。

实施步骤：

(1) 对连续变量进行标准化，并计算相关矩阵。

(2) 用特征值、碎石图、平行分析或累计解释率确定因子数。

(3) 估计因子载荷矩阵、公共度和特殊方差。

(4) 使用 varimax 等旋转方法增强解释性。

(5) 必要时计算因子得分，并作因子得分散点图。

注意事项：

- 旋转改变载荷表示，但不改变模型整体解释空间。
- 因子命名依赖高载荷变量的共同含义，不能只看单个变量。
- 若公共度过低，说明该变量不适合由所选公共因子解释。

1.6 聚类分析

聚类分析用于类别标签未知时的样本分组，核心是根据样本间距离或相似度形成组内相似、组间差异明显的类别。我主要掌握了系统聚类、动态聚类、自适应权重聚类和谱聚类。系统聚类从距离矩阵出发，按 single、complete、average、centroid、Ward 等连接准则逐步合并样本，并用树状图决定类别数。Ward 方法倾向于最小化类内平方和，常与标准化数据和欧氏距离一起使用。R 中实现流程是

```
d <- dist(scale(x)); hc <- hclust(d, method = "ward.D2"); group <- cutree(hc, k).
```

动态聚类以 K-means 为代表，其目标函数为

$$\min_{C_1,\dots,C_k} \sum_{r=1}^{k} \sum_{\boldsymbol{x}_i \in C_r} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{C_r}\|^2.$$

R 中可用 `kmeans(scale(x), centers = k, nstart = 25)`，并通过肘部法、总类内平方和、Gap statistic 或可解释性确定 k。谱聚类先构造相似矩阵 W，再构造图拉普拉斯矩阵 L，取其若干特征向量作为新坐标，最后在该坐标中使用 K-means。它适合非凸形状的数据，但对相似度尺度参数敏感。

实施步骤：

(1) 根据变量类型选择距离或相似度，并对连续变量标准化。

(2) 选择聚类方法，如系统聚类、K-means 或谱聚类。

(3) 确定类别数 k，并画树状图、碎石图或聚类散点图辅助判断。

(4) 解释各类中心、类内差异和类间差异。

(5) 若有真实标签，只在模型完成后用于评价聚类效果。

注意事项：

- 聚类是无监督方法，类别编号本身没有固定含义。
- 不同距离、连接方法、初始中心和类别数可能给出不同结果。
- 真实标签不能在无监督拟合中直接使用，否则会破坏聚类的含义。

1.7 判别分析与分类模型

判别分析处理类别已知的监督分类问题。若各类总体分布已知，最优分类规则是 Bayes 判别：把样本分到后验概率最大的类别。实际中常假设第 k 类服从多元正态分布 $N_p(\boldsymbol{\mu}_k, \Sigma_k)$，并用训练样本估计均值、协方差和先验概率。当各类协方差相等时得到线性判别分析（LDA），判别函数可写为

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}'\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\Sigma^{-1}\boldsymbol{\mu}_k + \log \pi_k.$$

当各类协方差不相等时得到二次判别分析（QDA），判别函数包含 Σ_k^{-1} 和 $\log|\Sigma_k|$，判别边界为二次曲面。Fisher 判别则从投影角度出发，寻找使类间差异相对类内差异最大的线性组合。

R 中 LDA 和 QDA 分别使用 `MASS::lda()` 与 `MASS::qda()`，用 `predict()` 得到类别、后验概率和判别函数得分；用 `table(actual, predicted)` 得到混淆矩阵。

实施步骤：

(1) 分层划分训练集和测试集。

(2) 在训练集上拟合 LDA、QDA 或 Logistic 回归模型。

(3) 先对训练集回判，再对测试集预测。

(4) 输出混淆矩阵，并计算误判率、准确率、灵敏度、特异度和精确率。

(5) 若类别比例不均衡，重点关注少数类召回率。

注意事项：

- LDA 假定各类协方差相同，QDA 允许协方差不同但对样本量要求更高。
- 高维变量或强共线会导致协方差矩阵估计不稳定。
- 分类阈值和先验概率会改变判别结果。

在二分类问题中，还可以使用 Logistic 回归。其模型为

$$\log \frac{P(Y=1 \mid \boldsymbol{x})}{1-P(Y=1 \mid \boldsymbol{x})} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

R 中用 `glm(Y ~ ., family = binomial, data = train)` 拟合，用 `predict(type = "response")` 得到违约概率，再按阈值转化为类别。

注意事项：

- Logistic 回归不要求解释变量联合正态，也不要求等协方差。
- 需要注意完全分离、异常值、共线性和类别不均衡。
- 默认阈值 0.5 不一定最合适，应结合误判代价调整。

1.8 对应分析

对应分析用于研究列联表中行变量和列变量之间的关联结构。其思想是以独立性模型为基准，把实际频数与独立性期望频数之间的偏离写成标准化残差矩阵，再通过类似奇异值分解的方式把行类别和列类别投影到低维平面。若 $N=(n_{ij})$ 为列联表，$P=N/n$，行边际为 $\boldsymbol{r}$，列边际为 $\boldsymbol{c}$，则对应分析围绕

$$D_r^{-1/2}(P-\boldsymbol{r}\boldsymbol{c}')D_c^{-1/2}$$

的分解展开。基本结论是：若行列变量独立，则标准化残差接近零；若不独立，则前几个维度可以解释主要的 χ^2 惯量。

R 实现与解释：

- 先用 `chisq.test()` 检验行变量和列变量是否独立。
- 再用 `MASS::corresp()` 或 `ca` 包进行对应分析并画双标图。
- 解释时关注行点和列点的相对位置、贡献率和业务含义。
- 不能简单把不同集合点之间的欧氏距离当成原始频数距离。

1.9 R 实施总步骤与注意事项

综合本学期所学和本次数据分析的实践，我把多元统计建模的 R 实施过程概括为以下步骤：

(1) 数据读入与整理：使用 `read.csv()`、`read.table()` 或 `data.table::fread()` 读入数据，检查变量类型、缺失值、重复值和异常值。

(2) 描述性分析：用 `summary()`、`boxplot()`、`hist()`、`density()`、`pairs()`、`cor()` 和 `cov()` 认识分布形态与相关结构。

(3) 预处理：根据方法需要进行中心化、标准化、对数变换、类别变量编码和训练集测试集划分。PCA、因子分析和聚类通常需要标准化；判别分析和 Logistic 回归还要注意类别不均衡。

(4) 模型拟合：PCA 用 `prcomp()` 或 `princomp()`；因子分析用 `factanal()`；系统聚类用 `dist()`、`hclust()` 和 `cutree()`；K-means 用 `kmeans()`；LDA/QDA 用 `MASS::lda()`、`MASS::qda()`；Logistic 回归用 `glm()`。

(5) 结果解释：PCA 解释载荷、得分和方差贡献率；因子分析解释载荷、公共度和特殊方差；聚类解释树状图、类中心和类间差异；判别和 Logistic 回归解释混淆矩阵、误判率、灵敏度、特异度和重要变量。

(6) 稳健性检查：改变标准化方式、距离度量、类别数、随机种子、阈值或训练测试划分，观察结论是否稳定。

总的注意事项：

- 多元统计模型既有数学假设，也有解释假设。

- 正态性、线性相关、等协方差、独立样本、低维表示和距离度量都不是自动成立的。
- 写报告时要同时交代：用了什么模型，为什么这样预处理，核心参数如何选择，结果是否可解释，模型有什么局限。

2 数据分析实战

台湾地区客户信贷违约案例数据集 Default of Credit Card Clients.csv 由 25 个变量、30000 条记录构成的一个真实数据集，数据集的前 6 行展示如下：

```
> library(data.table) # 数据读写包
> setwd("~/Desktop/2026_Multivariate Statistical Analysis/Exams/Fina Report") # 设置工作目录
> data = fread("Default of Credit Card Clients.csv", header = TRUE) # 读入数据集
> dim(data) # 数据集的规模
[1] 30000    25
> head(data) # 数据集的前6行
      ID     X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11   X12
   <int>  <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
1:     1  20000     2     2     1    24     2     2    -1    -1    -2    -2  3913
2:     2 120000     2     2     2    26    -1     2     0     0     0     2  2682
3:     3  90000     2     2     2    34     0     0     0     0     0     0 29239
4:     4  50000     2     2     1    37     0     0     0     0     0     0 46990
5:     5  50000     1     2     1    57    -1     0    -1     0     0     0  8617
6:     6  50000     1     1     2    37     0     0     0     0     0     0 64400
     X13   X14   X15   X16   X17   X18   X19   X20   X21   X22   X23     Y
   <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
1:  3102   689     0     0     0     0   689     0     0     0     0     1
2:  1725  2682  3272  3455  3261     0  1000  1000  1000     0  2000     1
3: 14027 13559 14331 14948 15549  1518  1500  1000  1000  1000  5000     0
4: 48233 49291 28314 28959 29547  2000  2019  1200  1100  1069  1000     0
5:  5670 35835 20940 19146 19131  2000 36681 10000  9000   689   679     0
6: 57069 57608 19394 19619 20024  2500  1815   657  1000  1000   800     0
```

从风险管理的角度出发，客户信用（是否违约）的预测与判定具有较高的实际参考价值。

数据集当中的第一个变量 ID 是数据编号，在进行数据分析时不使用它。X1 ～ X23 是解释变量，Y 是响应变量。各变量的含义如下：

- X1：授信额度 (单位：新台币)：包含个人消费信贷及家庭附属信贷。
- X2：性别，1 表示男性，2 表示女性。
- X3：受教育程度，1 表示研究生学历，2 表示本科学历，3 表示高中学历，4 表示其它学历。
- X4：婚姻状况，1 表示已婚，2 表示未婚，3 表示其它。
- X5：年龄 (单位：岁)。
- X6 ～ X11：过往还款记录。本数据集统计了 2005 年 4 月至 9 月的月度还款情况，其中 X6 代表 2005 年 9 月还款状态，X7 代表 2005 年 8 月还款状态，直至 X11 代表 2005 年 4 月还款状态。还款状态

赋值标准：−1 为按时还款；1 为逾期 1 个月；2 为逾期 2 个月，以此类推；8 为逾期 8 个月；9 为逾期 9 个月及以上。

- X12 ∼ X17：账单金额 (单位：新台币)。X12 为 2005 年 9 月账单金额，X13 为 2005 年 8 月账单金额，直至 X17 为 2005 年 4 月账单金额。
- X18 ∼ X23：往期还款金额 (新台币)。X18 为 2005 年 9 月还款额，X19 为 2005 年 8 月还款额，直至 X23 为 2005 年 4 月还款额。
- Y：还款违约情况，这是一个二分类的变量，违约记为 1，未违约记为 0。

2.1 数据的预处理与描述性分析 (满分：10 分)

2.1.1 数据的预处理

读入数据后，数据规模为 30000×25，其中 `ID` 为编号，建模时删除；$X_1 \sim X_{23}$ 为解释变量，Y 为响应变量。数据中没有缺失值。响应变量分布为：未违约 23364 人，占 77.88%；违约 6636 人，占 22.12%。因此数据存在明显的类别不均衡，若模型只预测“未违约”，也能得到约 77.88% 的表面准确率，所以后续评价不能只看 accuracy，还要看违约类的 sensitivity。

2.1.2 描述性分析

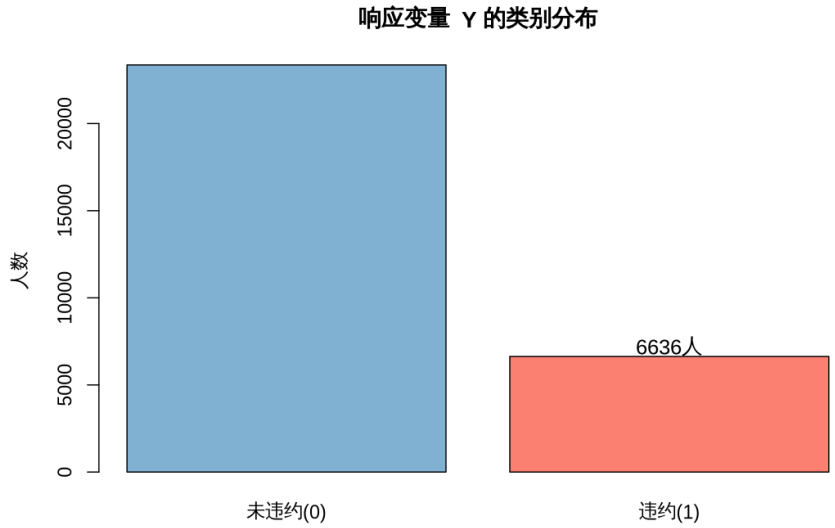

图 1: 响应变量 Y 的类别分布.

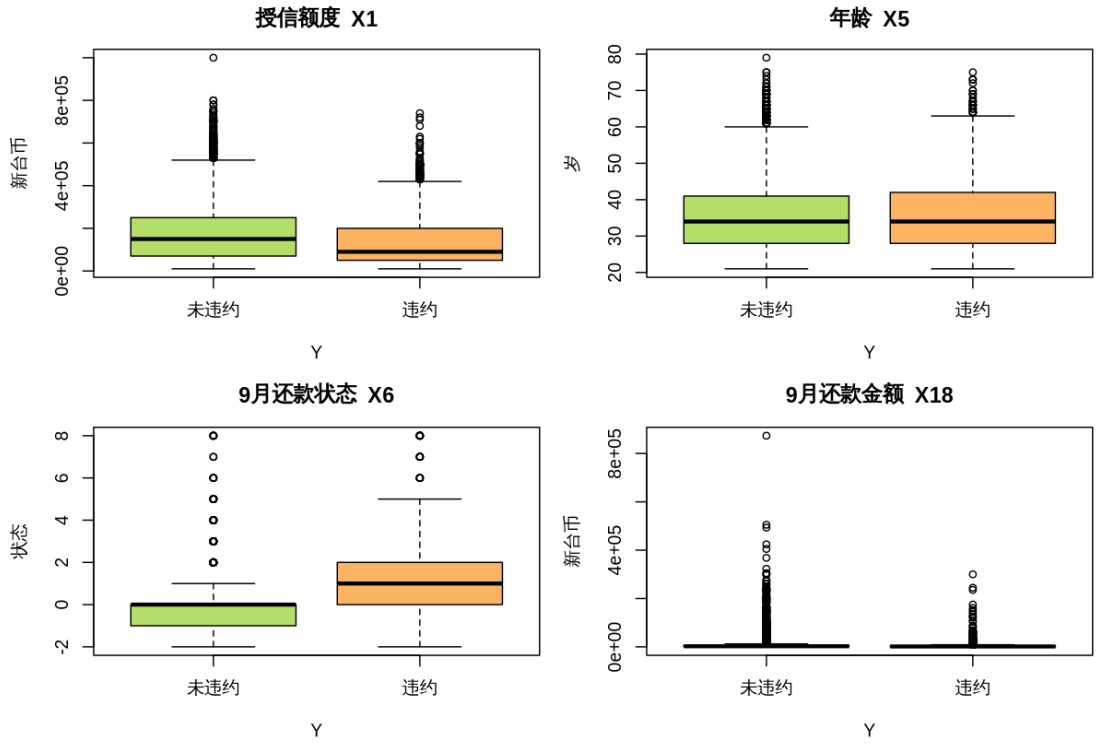

图 2: 若干关键变量在违约与未违约客户中的箱线图.

描述性统计显示，授信额度 X_1 的均值为 167484.32，标准差为 129747.66，说明额度差异较大；年龄 X_5 平均为 35.49 岁。账单金额 $X_{12} \sim X_{17}$ 和还款金额 $X_{18} \sim X_{23}$ 都呈现明显右偏，最大值远大于中位数。还款状态变量 $X_6 \sim X_{11}$ 的取值集中在 $-2, -1, 0, 1, 2, \ldots$，其中较大的正值代表更严重的逾期状态。从箱线图可以看出，违约客户在最近一期还款状态 X_6 上整体更偏向逾期，这也预示还款状态变量会是后续分类模型中的重要变量。

2.2 Logistic 回归 (满分：12 分)

变量 Y 是一个二元分类变量，给出了客户是否违约的状况。利用所给数据集，建立 Y 作为响应变量、X1 ∼ X23 作为解释变量的 Logistic 回归模型。

2.2.1 数据划分

使用分层抽样按 8 : 2 划分训练集和测试集。由于本机暂未安装 `caret`，我在 R 脚本里用基础 R 实现了与 `createDataPartition()` 功能相同的分层抽样；若在安装了 `caret` 的环境中运行，也可以直接使用题目要求的函数。以 Y 为响应变量、$X_1 \sim X_{23}$ 为解释变量拟合 Logistic 回归。测试集混淆矩阵如下：

2.2.2 模型拟合

在训练集上以 Y 为响应变量、$X_1 \sim X_{23}$ 为解释变量拟合 Logistic 回归模型，并使用测试集计算预测概率和预测类别。

2.2.3 模型分析

表 1: Logistic 回归测试集混淆矩阵

实际类别	预测未违约	预测违约
未违约	4539	134
违约	980	348

由此得到 accuracy = 0.8144，sensitivity = 0.2620，specificity = 0.9713，precision = 0.7220，AUC = 0.7253。模型总体准确率高于只预测未违约的基准水平，但对违约客户的召回率较低，说明默认阈值 0.5 下模型偏向识别未违约客户。若实际风控场景更重视发现违约客户，可以降低分类阈值或引入类别权重。

从回归系数的显著性看，最近一期还款状态 X_6 的 z 值最大，估计系数为正，说明最近还款逾期越严重，客户违约概率越高；账单金额、授信额度、教育程度、年龄、婚姻状况等变量也对违约概率有影响。这与信用风险管理的经验一致：最近还款记录通常比远期记录更能反映短期违约风险。

2.3 主成分分析 (满分：12 分)

2.3.1 作主成分分析

对 $X_1 \sim X_{23}$ 标准化后作主成分分析。前 10 个主成分的方差贡献如下：

表 2: 前 10 个主成分的方差贡献率

主成分	特征值	方差贡献率	累计贡献率
PC1	6.5431	0.2845	0.2845
PC2	4.0983	0.1782	0.4627
PC3	1.5510	0.0674	0.5301
PC4	1.4723	0.0640	0.5941
PC5	1.0252	0.0446	0.6387
PC6	0.9572	0.0416	0.6803
PC7	0.9076	0.0395	0.7198
PC8	0.8876	0.0386	0.7584
PC9	0.8712	0.0379	0.7962
PC10	0.7829	0.0340	0.8303

按特征值大于 1 的准则可取前 5 个主成分，累计解释约 63.87% 的标准化总方差；若按累计贡献率达到 80% 的准则，则需要取前 10 个主成分。考虑本数据变量较多且相关结构复杂，报告中解释主成分含义时重点看前 2 个主成分，建模降维时可选择前 10 个主成分。

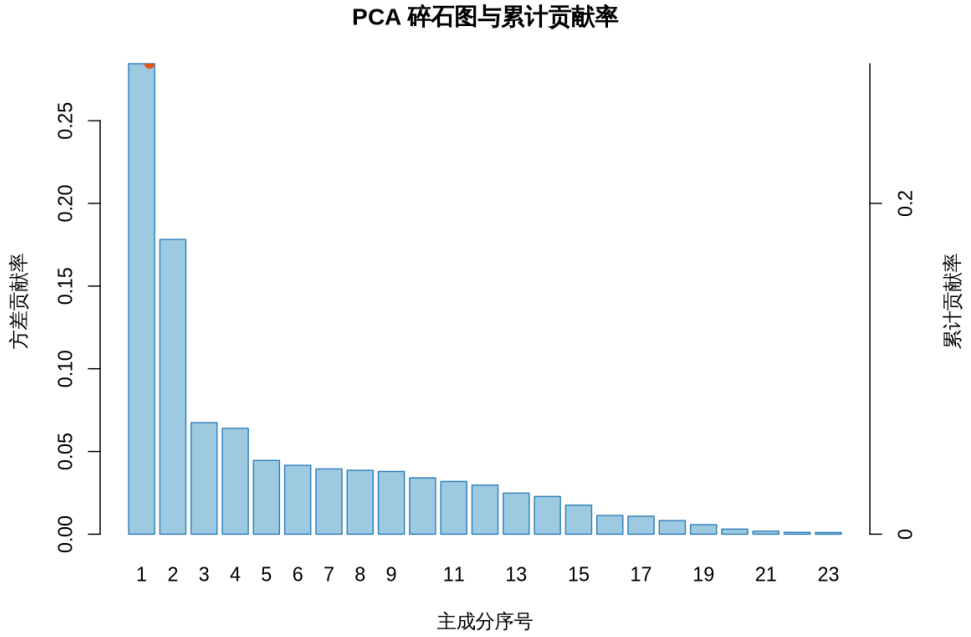

图 3: PCA 碎石图与累计贡献率.

2.3.2 主成分的解释

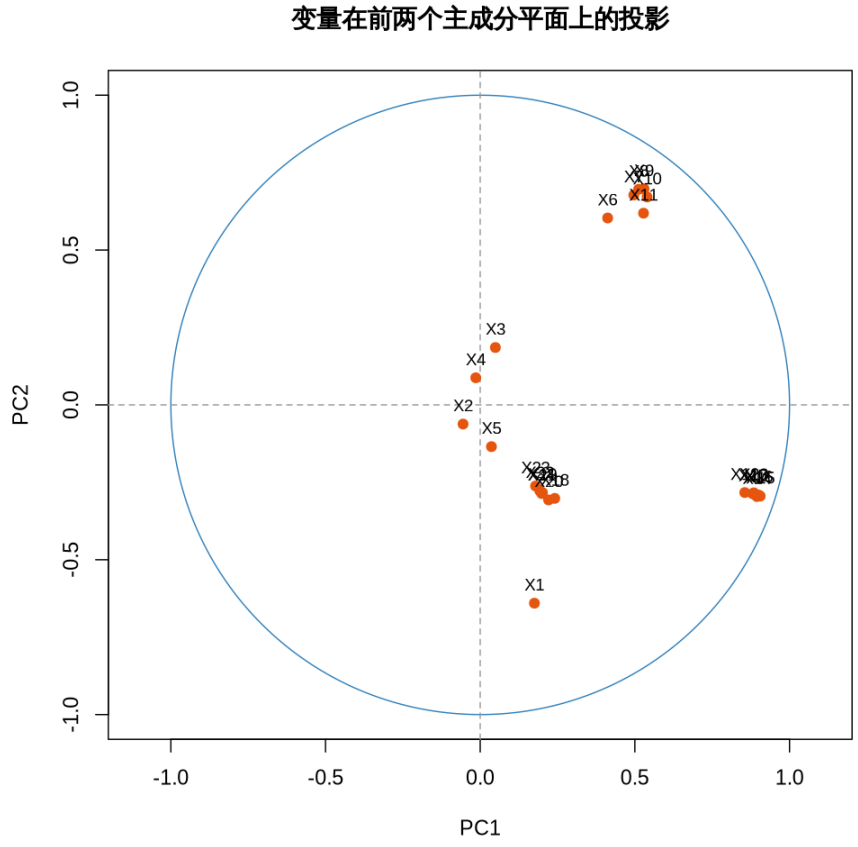

图 4: 变量在前两个主成分平面上的投影.

从变量投影看，PC1 与 $X_{12} \sim X_{17}$ 等账单金额变量关系最强，可解释为“账单规模/负债规模”因子；PC2 与 $X_7 \sim X_{11}$ 等还款状态变量关系更强，可解释为“历史逾期状态”因子。观测在 PC1-PC2 平面上的散点图显示，违约与未违约客户有一定重叠，单靠前两个主成分不能清晰分开两类客户，但违约客户在还款状态方向上有偏移趋势。

2.3.3 主成分对分辩客户违约与否的作用

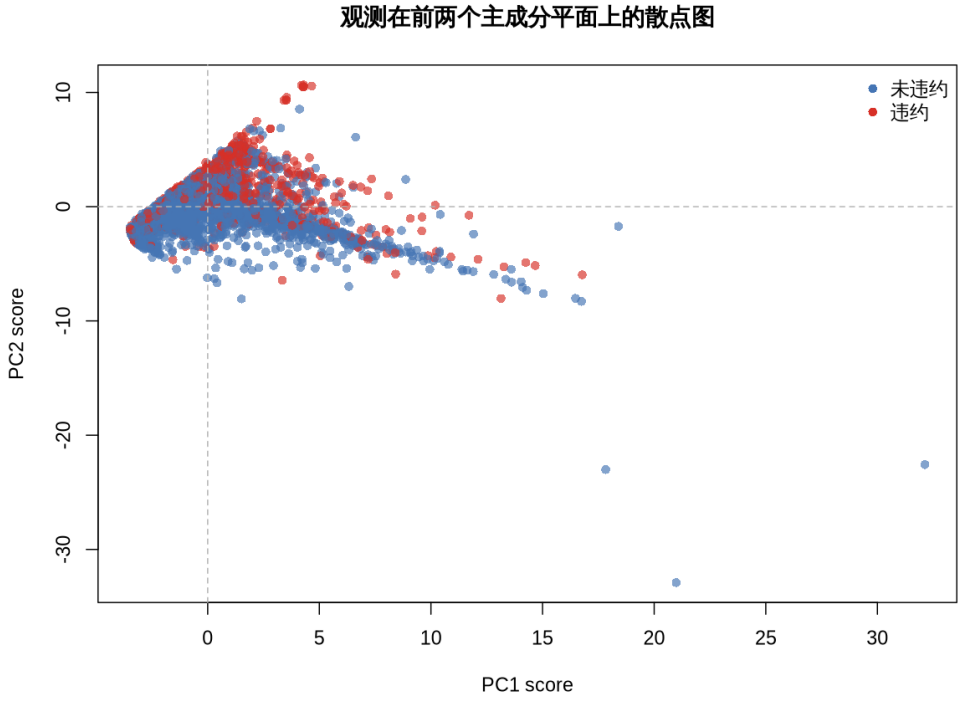

图 5: 观测在前两个主成分平面上的散点图.

2.4 因子分析 (满分：12 分)

2.4.1 拟合因子分析模型

根据相关矩阵特征值大于 1 的准则，选择 5 个公共因子，并使用极大似然因子分析和 varimax 正交旋转。旋转后的因子载荷矩阵如下：

表 3: 旋转后的因子载荷矩阵、公共度和特殊方差

变量	Factor1	Factor2	Factor3	Factor4	Factor5	公共度	特殊方差
X1	0.293	-0.358	0.286	0.149	0.053	0.320	0.680
X2	-0.015	-0.069	-0.013	0.020	0.014	0.006	0.994
X3	0.011	0.127	-0.072	-0.079	-0.024	0.028	0.972
X4	-0.032	0.048	-0.004	-0.007	-0.008	0.003	0.997
X5	0.060	-0.071	0.037	0.006	0.003	0.010	0.990
X6	0.140	0.619	-0.163	-0.063	-0.043	0.435	0.565
X7	0.159	0.755	-0.137	-0.065	-0.039	0.621	0.379
X8	0.131	0.830	-0.061	-0.039	-0.021	0.711	0.289
X9	0.111	0.891	-0.016	0.004	0.069	0.812	0.188
X10	0.111	0.880	0.013	0.078	0.064	0.797	0.203

变量	Factor1	Factor2	Factor3	Factor4	Factor5	公共度	特殊方差
X11	0.130	0.806	-0.001	0.127	0.052	0.685	0.315
X12	0.915	0.127	0.199	-0.086	-0.119	0.914	0.086
X13	0.933	0.152	0.279	-0.081	-0.127	0.995	0.005
X14	0.934	0.145	0.210	-0.071	0.231	0.995	0.005
X15	0.912	0.157	0.166	0.172	0.115	0.927	0.073
X16	0.909	0.153	0.095	0.359	0.079	0.995	0.005
X17	0.871	0.154	0.110	0.317	0.076	0.901	0.099
X18	0.126	-0.013	0.621	0.116	0.028	0.416	0.584
X19	0.097	-0.053	0.384	0.031	0.668	0.607	0.393
X20	0.095	-0.050	0.371	0.390	0.009	0.302	0.698
X21	0.146	-0.075	0.204	0.407	0.013	0.234	0.766
X22	0.121	-0.060	0.262	0.044	0.103	0.099	0.900
X23	0.128	-0.052	0.278	0.062	0.068	0.105	0.895

2.4.2 公共因子的解释

Factor1 在 $X_{12} \sim X_{17}$ 上载荷最高，可解释为“账单金额/负债规模”因子；Factor2 在 $X_6 \sim X_{11}$ 上载荷最高，可解释为“历史还款状态/逾期程度”因子；Factor3 主要与 $X_{18} \sim X_{20}$ 等还款金额有关；Factor4 与 $X_{20}, X_{21}, X_{16}, X_{17}$ 有一定关系；Factor5 主要受 X_{19} 影响。前两个因子累计解释约 41.06% 的标准化总方差，前五个因子累计解释约 51.82%。

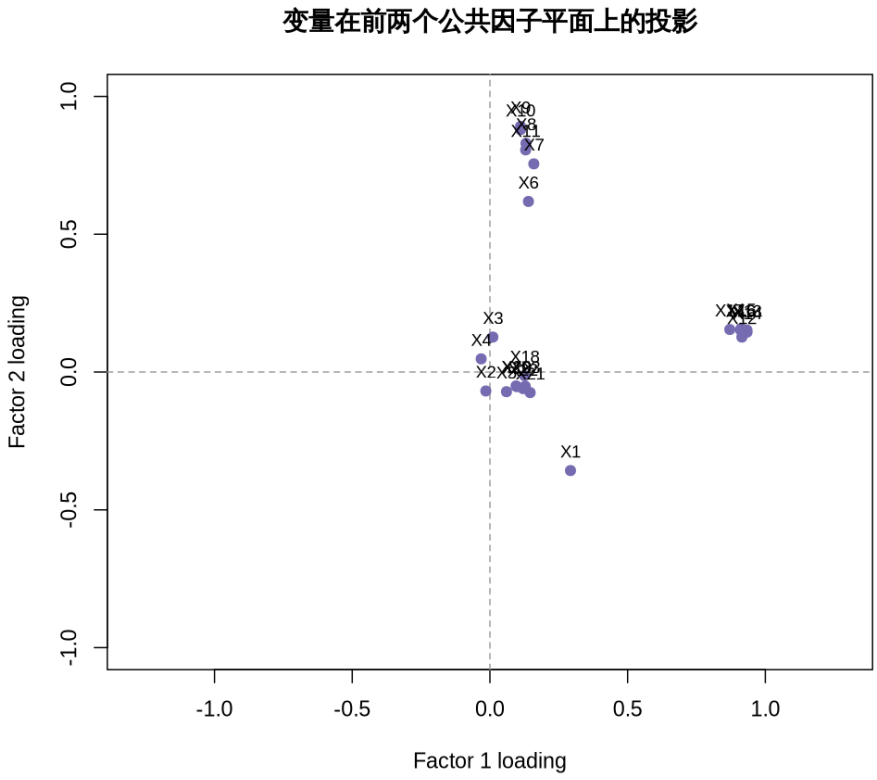

图 6: 变量在前两个公共因子平面上的投影.

2.4.3 公因子得分对分辩客户违约与否的作用

从因子得分散点图看，违约和未违约客户在前两个公共因子平面上仍有较多重叠。因子分析能解释变量结构、压缩信息，但单独使用前两个因子区分是否违约并不显著。

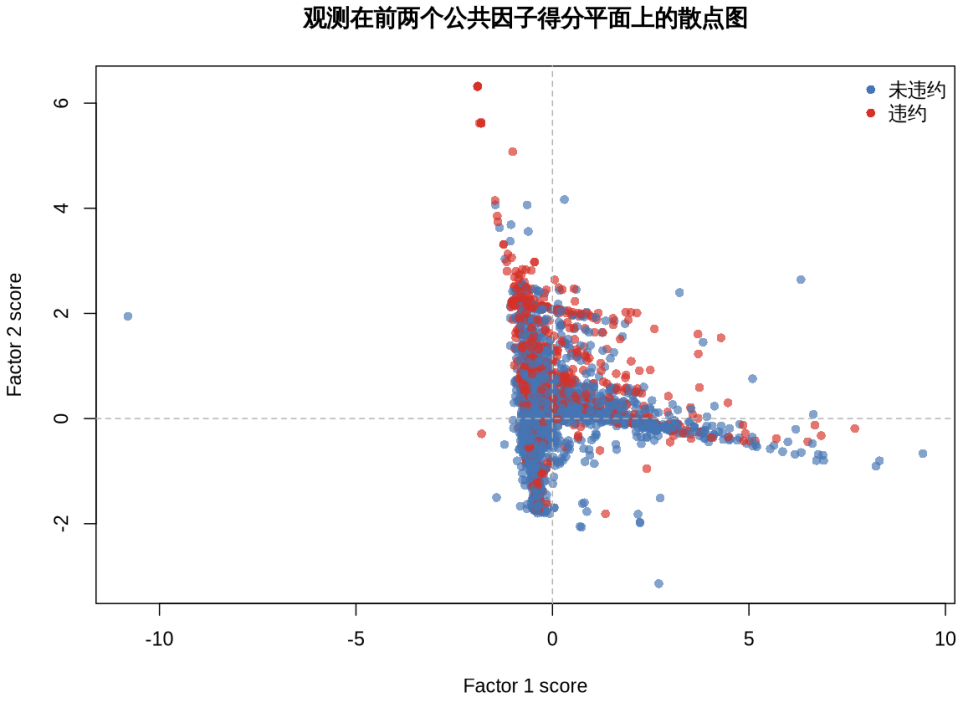

图 7: 观测在前两个公共因子得分平面上的散点图.

2.5 聚类分析 (满分：12 分)

系统聚类需要计算样本间距离矩阵，30000 个样本的全量距离矩阵规模过大，因此采用保持 Y 类别比例的分层抽样子集作系统聚类和谱聚类；动态聚类 `kmeans()` 可在全体样本上运行。聚类为无监督方法，聚类编号没有固定含义，因此与 Y 比较时对标签作最优匹配后计算指标。

2.5.1 系统聚类

表 4: 系统聚类不同连接方法的比较

方法	Accuracy	Sensitivity	Specificity	Precision	ErrorRate
single	0.7786	0.0000	0.9996	0.0000	0.2214
average	0.7786	0.0000	0.9996	0.0000	0.2214
complete	0.7786	0.0000	0.9996	0.0000	0.2214
ward.D2	0.6969	0.1554	0.8506	0.2279	0.3031

single、average、complete 的 accuracy 接近类别基准，但 sensitivity 为 0，说明几乎没有识别出违约类；ward.D2 虽然总体准确率较低，但能识别一部分违约客户。从“是否区分违约客户”的角度看，ward.D2 更有实际意义。

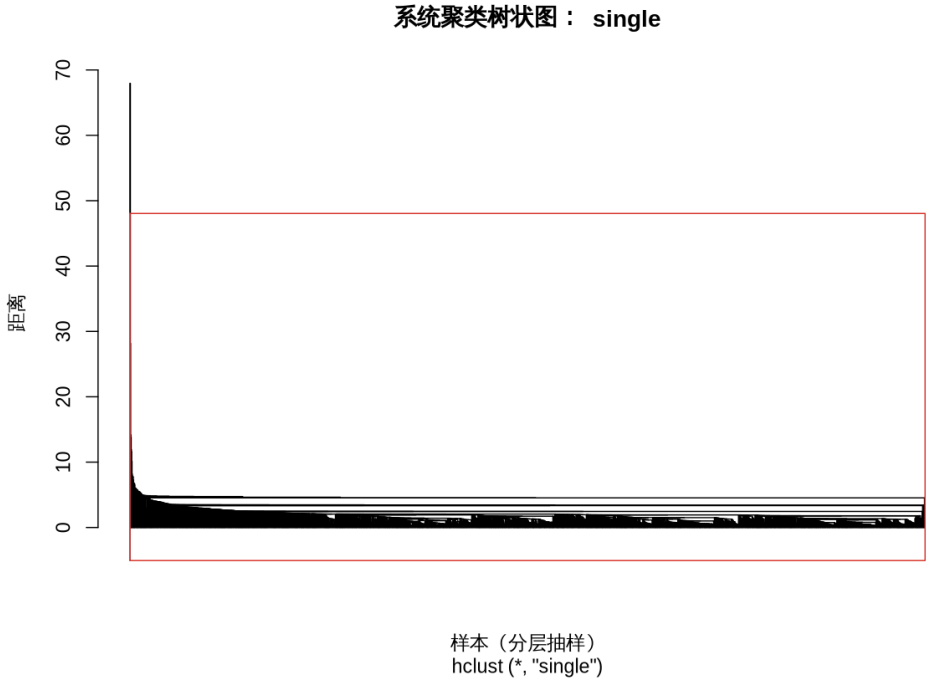

图 8: 系统聚类树状图.

2.5.2 动态聚类

表 5: 动态聚类不同算法的比较

算法	Accuracy	Sensitivity	Specificity	Precision	ErrorRate
Hartigan-Wong	0.6906	0.1492	0.8443	0.2140	0.3094
Lloyd	0.6903	0.1493	0.8439	0.2137	0.3097
Forgy	0.6903	0.1493	0.8439	0.2137	0.3097
MacQueen	0.6903	0.1493	0.8439	0.2137	0.3097

四种 kmeans 算法结果非常接近，Hartigan-Wong 的 accuracy 略高。与系统聚类相比，动态聚类能识别一部分违约客户，但 sensitivity 仍较低，说明在不使用 Y 的情况下，仅靠解释变量的自然分组难以完全对应违约状态。

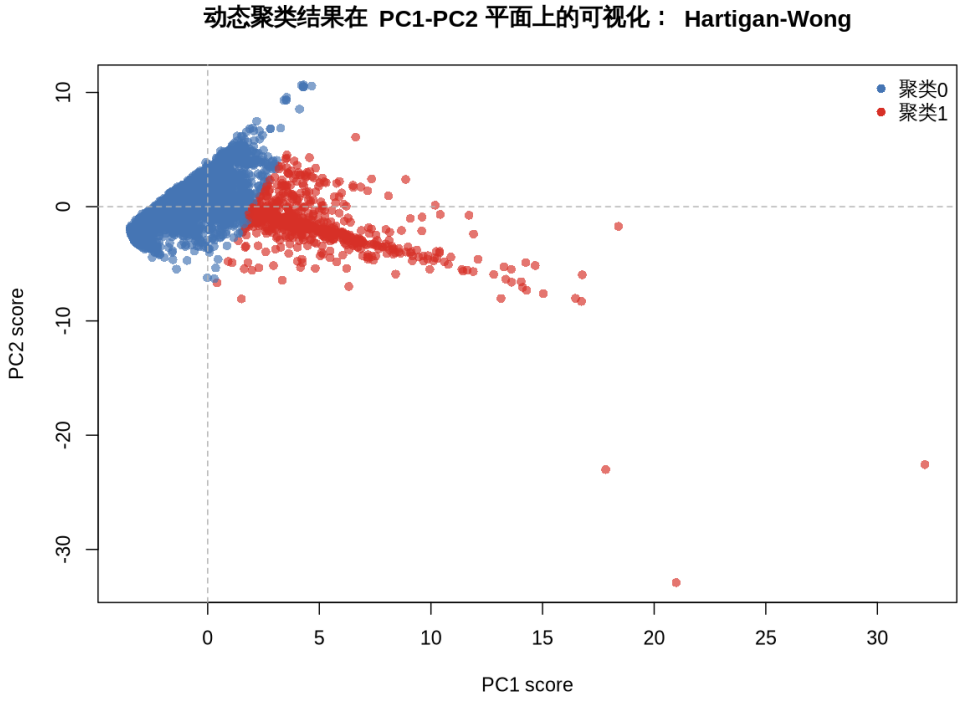

图 9: 动态聚类结果在 PC1-PC2 平面上的可视化.

2.5.3 谱聚类

由于本机未安装 `kernlab`，实际运行时使用与 `specc()` 思想一致的归一化谱聚类：构造 RBF 相似矩阵，计算归一化图拉普拉斯矩阵的特征向量，再在特征向量空间中作 kmeans。谱聚类结果 accuracy = 0.6389，sensitivity = 0.4868，specificity = 0.6820，precision = 0.3028。谱聚类总体准确率低于系统聚类和动态聚类，但对违约类的识别率更高，说明非线性相似性结构能捕捉部分违约客户特征，不过误报也更多。

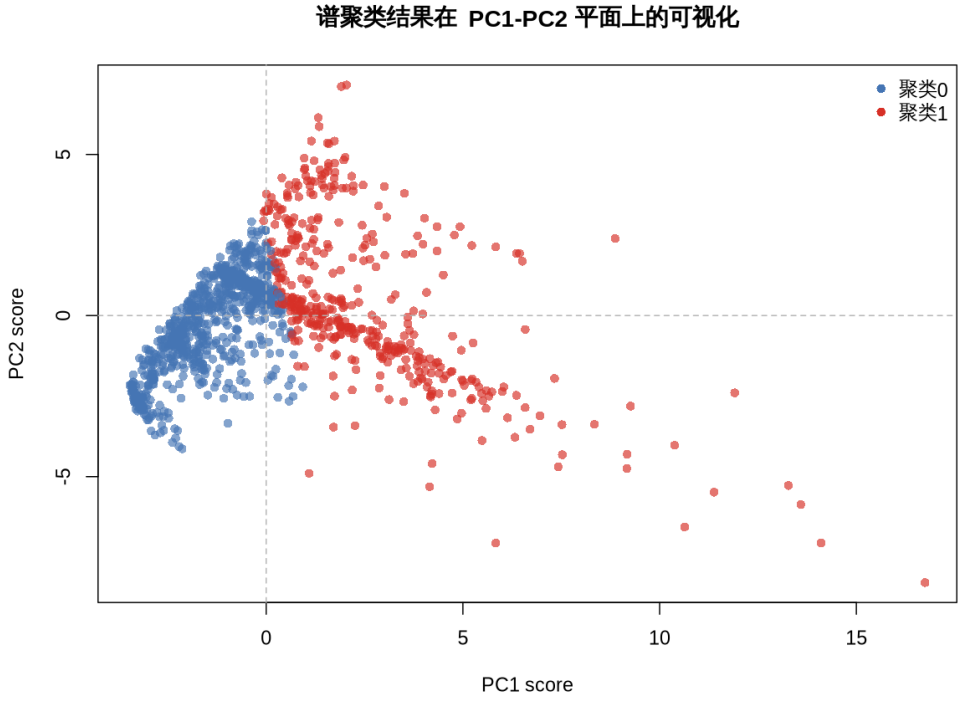

图 10: 谱聚类结果在 PC1-PC2 平面上的可视化.

2.6 判别分析 (满分：12 分)

2.6.1 数据划分

使用分层抽样按 7 : 3 划分训练集与测试集。

2.6.2 线性判别分析

LDA 训练集与测试集结果如下：

表 6: LDA 混淆矩阵

	训练集预测		测试集预测	
实际类别	未违约	违约	未违约	违约
未违约	15832	522	6770	240
违约	3407	1238	1448	543

LDA 测试集 accuracy = 0.8125，sensitivity = 0.2727，specificity = 0.9658，precision = 0.6935。其总体表现与 Logistic 回归非常接近，说明在该数据上，线性分类边界可以较好地区分未违约客户，但对违约客户的召回率仍偏低。

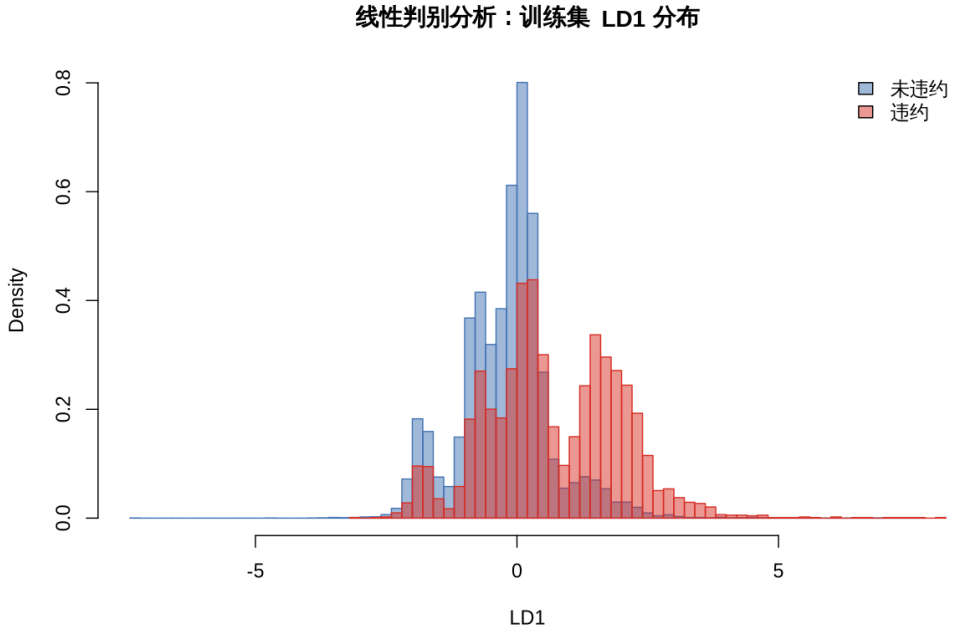

图 11: LDA 训练集在第一判别方向上的分布.

2.6.3 二次判别分析

QDA 训练集与测试集结果如下：

表 7: QDA 混淆矩阵

	训练集预测		测试集预测	
实际类别	未违约	违约	未违约	违约
未违约	6543	9811	2816	4194
违约	817	3828	308	1683

QDA 测试集 accuracy = 0.4998, sensitivity = 0.8453, specificity = 0.4017, precision = 0.2864。QDA 大幅提高了对违约客户的识别率，但把大量未违约客户误判为违约，导致总体准确率较低。综合比较，若目标是稳健预测总体类别，Logistic 回归和 LDA 更合适；若业务目标是尽可能发现潜在违约客户，QDA 或谱聚类这类高 sensitivity 方法可作为预警筛查，但需要后续人工复核或调整阈值以降低误报。

2.7 代码说明

建模过程中使用到的全部 R 代码已经单独整理为脚本文件

`mvsa-final-report-analysis.R`

并放入提交文件夹中，用于一次性复现全部数值结果和图形。