Course 2023-2024 in Financial Risk Management
Lecture 12. Credit Scoring Models

Thierry Roncalli*

*Amundi Asset Management[1]

*University of Paris-Saclay

September 2023

[1] The opinions expressed in this presentation are those of the authors and are not meant to represent the opinions or official positions of Amundi Asset Management.

General information

1. **Overview**
   The objective of this course is to understand the theoretical and practical aspects of risk management
2. **Prerequisites**
   M1 Finance or equivalent
3. **ECTS**
   4
4. **Keywords**
   Finance, Risk Management, Applied Mathematics, Statistics
5. **Hours**
   Lectures: 36h, Training sessions: 15h, HomeWork: 30h
6. **Evaluation**
   There will be a final three-hour exam, which is made up of questions and exercises
7. **Course website**
   `http://www.thierry-roncalli.com/RiskManagement.html`

Objective of the course

The objective of the course is twofold:

1. knowing and understanding the financial regulation (banking and others) and the international standards (especially the Basel Accords)
2. being proficient in risk measurement, including the mathematical tools and risk models

Class schedule

Course sessions

- September 15 (6 hours, AM+PM)
- September 22 (6 hours, AM+PM)
- September 19 (6 hours, AM+PM)
- October 6 (6 hours, AM+PM)
- October 13 (6 hours, AM+PM)
- October 27 (6 hours, AM+PM)

Tutorial sessions

- October 20 (3 hours, AM)
- October 20 (3 hours, PM)
- November 10 (3 hours, AM)
- November 10 (3 hours, PM)
- November 17 (3 hours, PM)

Class times: Fridays 9:00am-12:00pm, 1:00pm–4:00pm, University of Evry, Room 209 IDF

Agenda

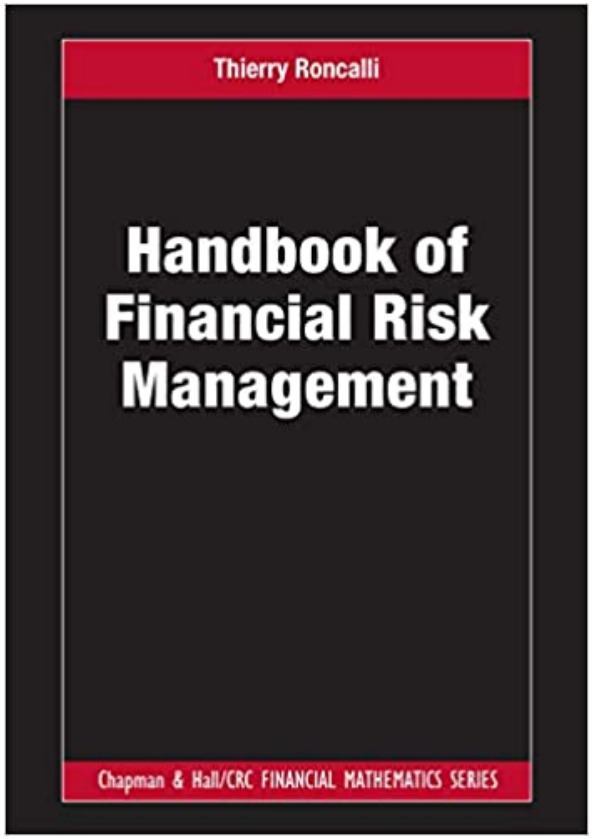
- Lecture 1: Introduction to Financial Risk Management
- Lecture 2: Market Risk
- Lecture 3: Credit Risk
- Lecture 4: Counterparty Credit Risk and Collateral Risk
- Lecture 5: Operational Risk
- Lecture 6: Liquidity Risk
- Lecture 7: Asset Liability Management Risk
- Lecture 8: Model Risk
- Lecture 9: Copulas and Extreme Value Theory
- Lecture 10: Monte Carlo Simulation Methods
- Lecture 11: Stress Testing and Scenario Analysis
- Lecture 12: Credit Scoring Models

Agenda

- Tutorial Session 1: Market Risk
- Tutorial Session 2: Credit Risk
- Tutorial Session 3: Counterparty Credit Risk and Collateral Risk
- Tutorial Session 4: Operational Risk & Asset Liability Management Risk
- Tutorial Session 5: Copulas, EVT & Stress Testing

Textbook

- Roncalli, T. (2020), *Handbook of Financial Risk Management*, Chapman & Hall/CRC Financial Mathematics Series.



Additional materials

- Slides, tutorial exercises and past exams can be downloaded at the following address:

  `http://www.thierry-roncalli.com/RiskManagement.html`

- Solutions of exercises can be found in the companion book, which can be downloaded at the following address:

  `http://www.thierry-roncalli.com/RiskManagementBook.html`

Agenda

- Lecture 1: Introduction to Financial Risk Management
- Lecture 2: Market Risk
- Lecture 3: Credit Risk
- Lecture 4: Counterparty Credit Risk and Collateral Risk
- Lecture 5: Operational Risk
- Lecture 6: Liquidity Risk
- Lecture 7: Asset Liability Management Risk
- Lecture 8: Model Risk
- Lecture 9: Copulas and Extreme Value Theory
- Lecture 10: Monte Carlo Simulation Methods
- Lecture 11: Stress Testing and Scenario Analysis
- **Lecture 12: Credit Scoring Models**

Credit scoring

- Credit scoring refers to statistical models to measure the creditworthiness of a person or a company
- Mortgage, credit card, personal loan, etc.
- Credit scoring first emerged in the United States
- The FICO score was introduced in 1989 by Fair Isaac Corporation

Judgmental credit systems versus credit scoring systems

- In 1941, Durand presented a statistical analysis of credit valuation
- He showed that credit analysts uses similar factors, and proposed a credit rating formula based on nine factors: (1) age, (2) sex, (3) stability of residence, (4) occupation, (5) industry, (6) stability of employment, (7) bank account, (8) real estate and (9) life insurance
- The score is additive and can take values between 0 and 3.46
- From an industrial point of view, a credit scoring system has two main advantages compared to a judgmental credit system:
  1. it is cost efficient, and can treat a huge number of applicants;
  2. decision-making process is rapid and consistent across customers.

Scoring models for corporate bankruptcy

Altman Z score model (1968)

- The score was equal to:

$$Z = 1.2 \cdot X_1 + 1.4 \cdot X_2 + 3.3 \cdot X_3 + 0.6 \cdot X_4 + 1.0 \cdot X_5$$

- The variables X_j represent the following financial ratios:

X_j	Ratio
X_1	Working capital / Total assets
X_2	Retained earnings / Total assets
X_3	Earnings before interest and tax / Total assets
X_4	Market value of equity / Total liabilities
X_5	Sales / Total assets

- If we note Z_i the score of the firm i, we can calculate the normalized score:

$$Z_i^\star = (Z_i - m_z) / \sigma_z$$

where m_z and σ_z are the mean and standard deviation of the observed scores

- A low value of $Z_i^\star$ (for instance $Z_i^\star < 2.5$) indicates that the firm has a high probability of default

New developments

- Default of corporate firms
- Consumer credit and retail debt management (credit cards, mortgages, etc.)
- Statistical methods: discriminant analysis, logistic regression, survival model, machine learning techniques

Choice of the risk factors

The five Cs:

1. **Capacity** measures the applicant's ability to meet the loan payments (e.g., debt-to-income, job stability, cash flow dynamics)
2. **Capital** is the size of assets that are held by the borrower (e.g. net wealth of the borrower)
3. **Character** measures the willingness to repay the loan (e.g. payment history of the applicant)
4. **Collateral** concerns additional forms of security that the borrower can provide to the lender
5. **Conditions** refer to the characteristics of the loan and the economic conditions that might affect the borrower (e.g. maturity, interests paid)

Choice of the risk factors

Table: An example of risk factors for consumer credit

Character	Age of applicant
	Marital status
	Number of children
	Educational background
	Time with bank
	Time at present address
Capacity	Annual income
	Current living expenses
	Current debts
	Time with employer
Capital	Purpose of the loan
	Home status
	Saving account
Condition	Maturity of the loan
	Paid interests

Choice of the risk factors

- Scores are developed by banks and financial institutions, but they can also be developed by consultancy companies
- This is the case of the FICO® scores, which are the most widely used credit scoring systems in the world

5 main categories

1. Payment history (35%)
2. Amount of debt (30%)
3. Length of credit history (15%)
4. New credit (10%)
5. Credit mix (10%)

Range

Generally from 300 to 850 (average score of US consumers is 695)

- Exceptional (800+)
- Very good (740-799)
- Good (670-739)
- Fair (580-669)
- Poor (580−)

Choice of the risk factors

Corporate credit scoring systems use financial ratios:

1. **Profitability**: gross profit margin, operating profit margin, return-on-equity (ROE), etc.
2. **Solvency**: debt-to-assets ratio, debt-to-equity ratio, interest coverage ratio, etc.
3. **Leverage**: liabilities-to-assets ratio (financial leverage ratio), long-term debt/assets, etc.
4. **Liquidity**: current assets/current liabilities (current ratio), quick assets/current liabilities (quick or cash ratio), total net working capital, assets with maturities of less than one year, etc.

Data preparation

- Check the data and remove outliers or fill missing values
- Variable transformation
- Slicing-and-dicing segmentation
- Potential interaction

Variable selection

- Many candidate variables $X = (X_1, \ldots, X_m)$ for explaining the variable Y
- The variable selection problem consists in finding the best set of optimal variables
- We assume the following statistical model:

$$Y = f(X) + u$$

where $u \sim \mathcal{N}(0, \sigma^2)$

- We denote the prediction by $\hat{Y} = \hat{f}(X)$. We have:

$$\begin{aligned}
\mathbb{E}\left[\left(Y - \hat{Y}\right)^2\right] &= \mathbb{E}\left[\left(f(X) + u - \hat{f}(X)\right)^2\right] \\
&= \left(\mathbb{E}\left[\hat{f}(X)\right] - f(X)\right)^2 + \mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\left[\hat{f}(X)\right]\right)^2\right] + \sigma^2 \\
&= \text{Bias}^2 + \text{Variance} + \text{Error}
\end{aligned}$$

Variable selection

- Best subset selection:

$$\mathrm{AIC}(\alpha) = -2\ell_{(k)}\left(\hat{\theta}\right) + \alpha \cdot \mathrm{df}_{(k)}^{(\mathrm{model})}$$

- Stepwise approach:

$$F = \frac{\mathrm{RSS}\left(\hat{\theta}_{(k)}\right) - \mathrm{RSS}\left(\hat{\theta}_{(k+1)}\right)}{\mathrm{RSS}\left(\hat{\theta}_{(k+1)}\right) / \mathrm{df}_{(k+1)}^{(\mathrm{residual})}}$$

- Lasso approach:

$$y_i = \sum_{k=1}^{K} \beta_k x_{i,k} + u_i \quad \text{s.t.} \quad \sum_{k=1}^{K} |\beta_k| \leq \tau$$

Score modeling, validation and follow-up

- Cross-validation approach (leave-p-out cross-validation or LpOCV, leave-one-out cross-validation or LOOCV, Press statistic)
- Score modeling
  - $S = f\left(X; \hat{\theta}\right)$ is the score
  - Decision rule:

$$\begin{cases} S < s \Longrightarrow Y = 0 \Longrightarrow \text{reject} \\ S \geq s \Longrightarrow Y = 1 \Longrightarrow \text{accept} \end{cases}$$

- Score follow-up
  - Stability
  - Rejected applicants (reject inference)
  - Backtesting

Statistical methods

- Unsupervised learning is a branch of statistical learning, where test data does not include a response variable
- It is opposed to supervised learning, whose goal is to predict the value of the response variable Y given a set of explanatory variables X
- In the case of unsupervised learning, we only know the X-values, because the Y-values do not exist or are not observed
- Supervised and unsupervised learning are also called '*learning with/without a teacher*' (Hastie *et al.*, 2009)

Clustering

- K-means clustering
- Hierarchical clustering

Clustering

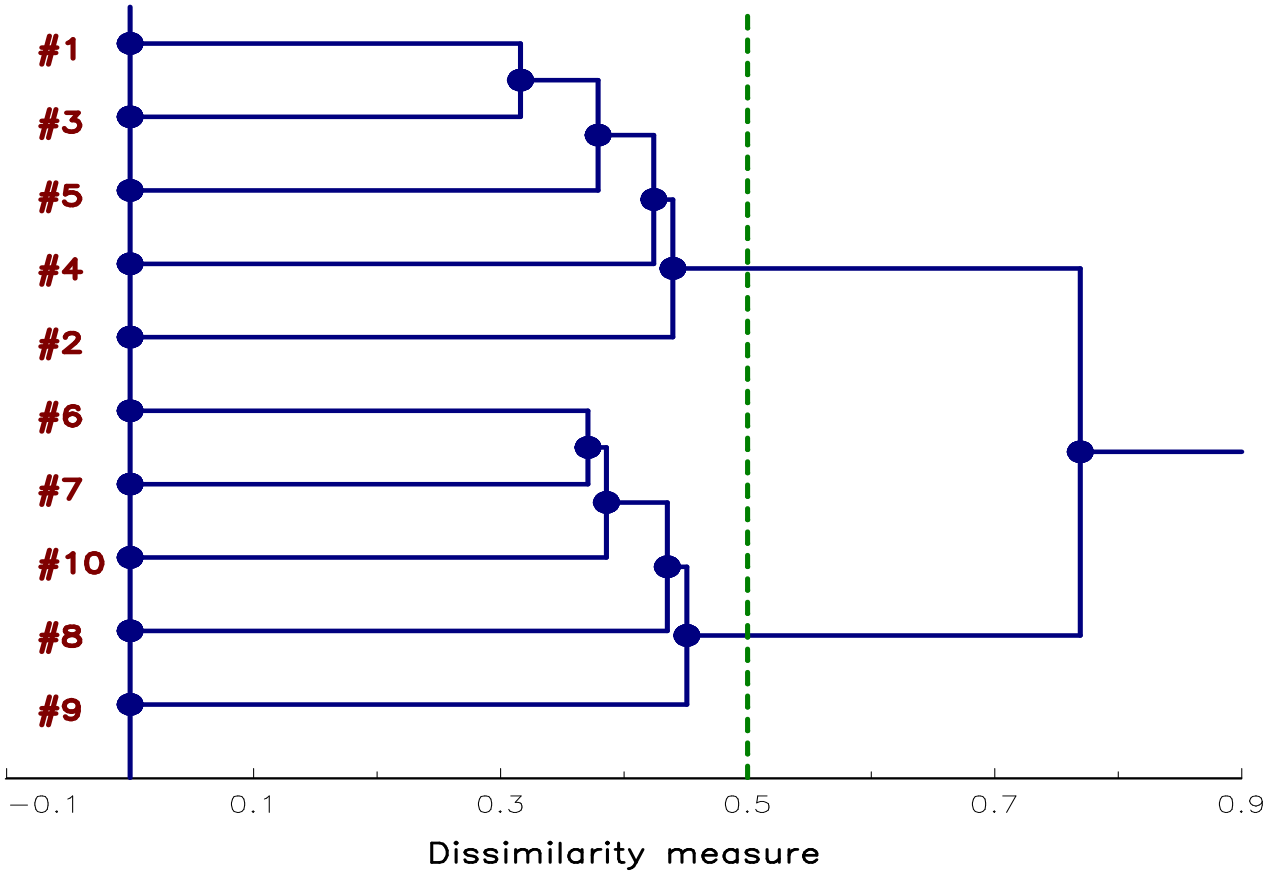


Figure: An example of dendrogram

Dimension reduction

- Principal component analysis
- Non-negative matrix factorization

Discriminant analysis

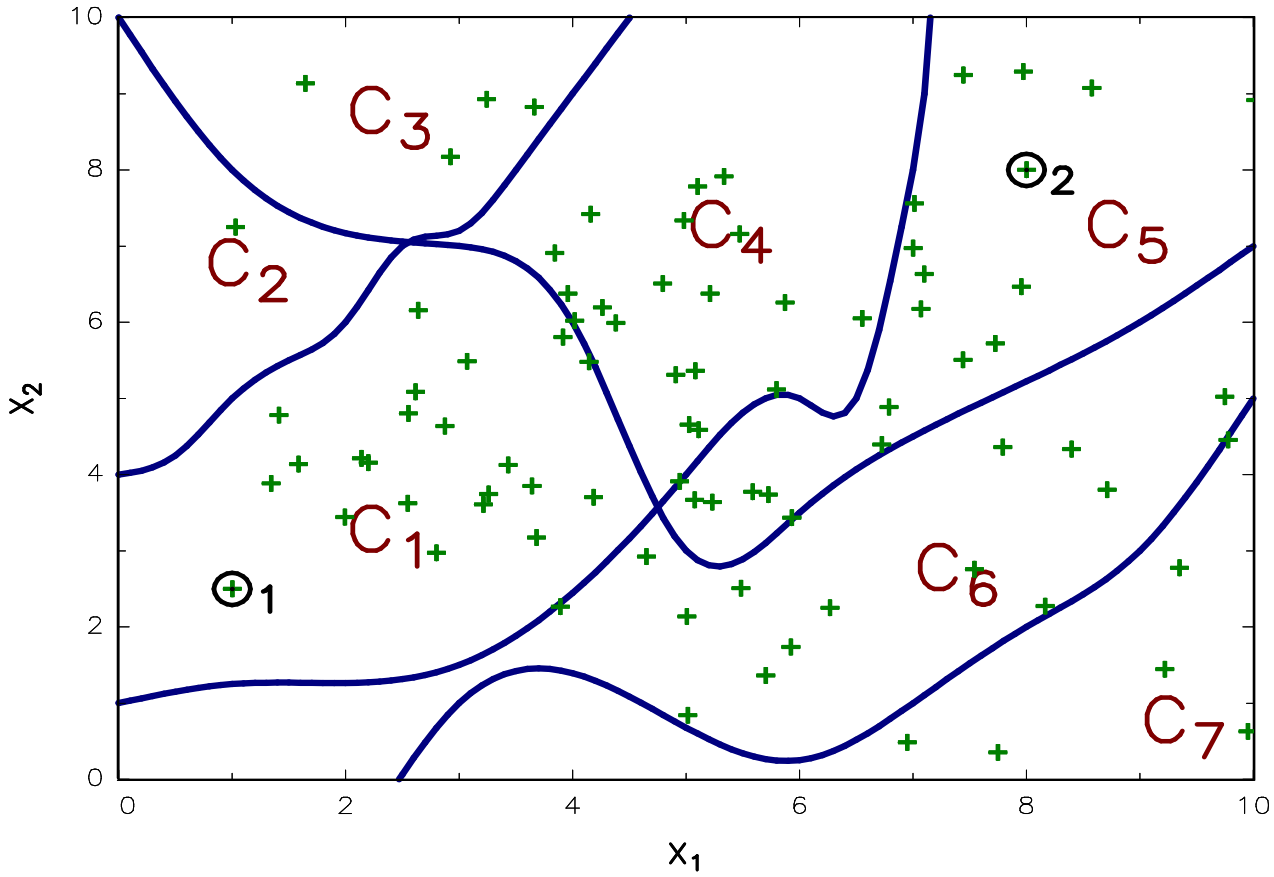


Figure: Classification statistical problem

Discriminant analysis

The two-dimensional case

- Using the Bayes theorem, we have:

$$\Pr\{A \cap B\} = \Pr\{A \mid B\} \cdot \Pr\{B\} = \Pr\{B \mid A\} \cdot \Pr\{A\}$$

- It follows that:

$$\Pr\{A \mid B\} = \Pr\{B \mid A\} \cdot \frac{\Pr\{A\}}{\Pr\{B\}}$$

- If we apply this result to the conditional probability $\Pr\{i \in \mathcal{C}_1 \mid X = x\}$, we obtain:

$$\Pr\{i \in \mathcal{C}_1 \mid X = x\} = \Pr\{X = x \mid i \in \mathcal{C}_1\} \cdot \frac{\Pr\{i \in \mathcal{C}_1\}}{\Pr\{X = x\}}$$

Discriminant analysis

The two-dimensional case

- The log-probability ratio is then equal to:

$$\begin{aligned}\ln \frac{\Pr\{i \in \mathcal{C}_1 \mid X = x\}}{\Pr\{i \in \mathcal{C}_2 \mid X = x\}} &= \ln \left(\frac{\Pr\{X = x \mid i \in \mathcal{C}_1\}}{\Pr\{X = x \mid i \in \mathcal{C}_2\}} \cdot \frac{\Pr\{i \in \mathcal{C}_1\}}{\Pr\{i \in \mathcal{C}_2\}} \right) \\ &= \ln \frac{f_1(x)}{f_2(x)} + \ln \frac{\pi_1}{\pi_2}\end{aligned}$$

  where $\pi_j = \Pr\{i \in \mathcal{C}_j\}$ is the probability of the j^{th} class and $f_j(x) = \Pr\{X = x \mid i \in \mathcal{C}_j\}$ is the conditional pdf of X

- By construction, the decision boundary is defined such that we are indifferent to an assignment rule ($i \in \mathcal{C}_1$ and $i \in \mathcal{C}_2$), implying that:

$$\Pr\{i \in \mathcal{C}_1 \mid X = x\} = \Pr\{i \in \mathcal{C}_2 \mid X = x\} = \frac{1}{2}$$

- Finally, we deduce that the decision boundary satisfies the following equation:

$$\ln \frac{f_1(x)}{f_2(x)} + \ln \frac{\pi_1}{\pi_2} = 0$$

Discriminant analysis

Quadratic discriminant analysis (QDA)

- If we model each class density as a multivariate normal distribution:

$$X \mid i \in \mathcal{C}_j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

we have:

$$f_j(x) = \frac{1}{(2\pi)^{K/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_j)^\top \Sigma_j^{-1}(x-\mu_j)\right)$$

- We deduce that:

$$\ln \frac{f_1(x)}{f_2(x)} = \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^\top \Sigma_2^{-1}(x-\mu_2)$$

- The decision boundary is then given by:

$$\frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^\top \Sigma_2^{-1}(x-\mu_2) + \ln\frac{\pi_1}{\pi_2} = 0$$

Discriminant analysis

Linear discriminant analysis (LDA)

- If we assume that $\Sigma_1 = \Sigma_2 = \Sigma$, we obtain:

$$\frac{1}{2}(x-\mu_2)^\top \Sigma^{-1}(x-\mu_2) - \frac{1}{2}(x-\mu_1)^\top \Sigma^{-1}(x-\mu_1) + \ln\frac{\pi_1}{\pi_2} = 0$$

- We deduce that:

$$(\mu_2-\mu_1)^\top \Sigma^{-1} x = \frac{1}{2}\left(\mu_2^\top \Sigma^{-1}\mu_2 - \mu_1^\top \Sigma^{-1}\mu_1\right) + \ln\frac{\pi_2}{\pi_1}$$

- The decision boundary is then linear in x (and not quadratic)

Discriminant analysis

Example #1

We consider two classes and two explanatory variables $X = (X_1, X_2)$ where $\pi_1 = 50\%$, $\pi_2 = 1 - \pi_1 = 50\%$, $\mu_1 = (1, 3)$, $\mu_2 = (4, 1)$, $\Sigma_1 = I_2$ and $\Sigma_2 = \gamma I_2$ where $\gamma = 1.5$.

Discriminant analysis

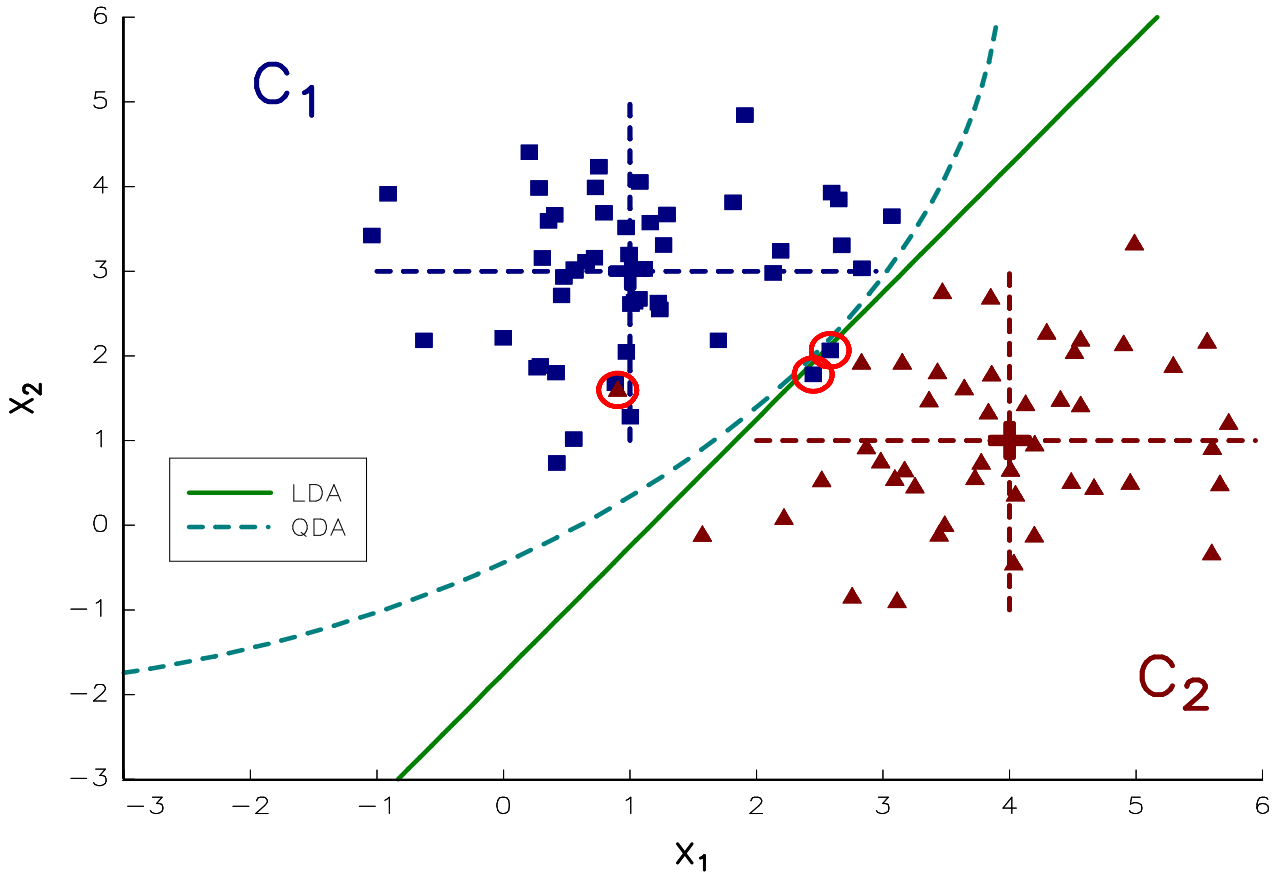


Figure: Boundary decision of discriminant analysis

Discriminant analysis

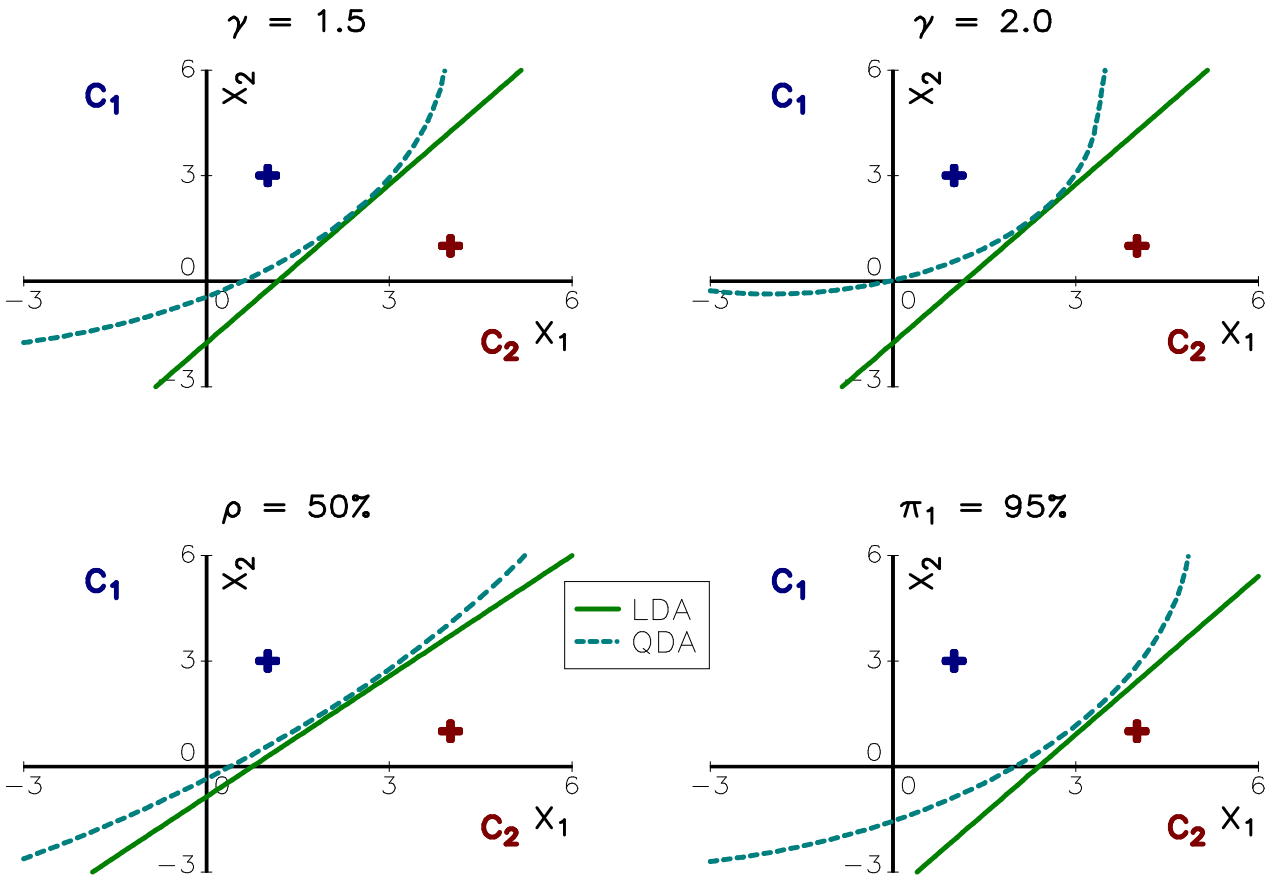


Figure: Impact of the parameters on LDA/QDA boundary decisions

Discriminant analysis

The general case

- We can generalize the previous analysis to J classes
- The Bayes formula gives:

$$\begin{aligned}\Pr\{i \in \mathcal{C}_j \mid X = x\} &= \Pr\{X = x \mid i \in \mathcal{C}_j\} \cdot \frac{\Pr\{i \in \mathcal{C}_j\}}{\Pr\{X = x\}} \\ &= c \cdot f_j(x) \cdot \pi_j\end{aligned}$$

  where $c = 1/\Pr\{X = x\}$ is a normalization constant that does not depend on j
- We note $S_j(x) = \ln \Pr\{i \in \mathcal{C}_j \mid X = x\}$ the discriminant score function for the j^{th} class
- We have:

$$S_j(x) = \ln c + \ln f_j(x) + \ln \pi_j$$

Discriminant analysis

The general case

- If we again assume that $X \mid i \in \mathcal{C}_j \sim \mathcal{N}(\mu_j, \Sigma_j)$, the QDA score function is:

$$\begin{aligned} S_j(x) &= \ln c' + \ln \pi_j - \frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (x - \mu_j)^\top \Sigma_j^{-1} (x - \mu_j) \\ &\propto \ln \pi_j - \frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (x - \mu_j)^\top \Sigma_j^{-1} (x - \mu_j) \end{aligned}$$

  where $\ln c' = \ln c - \dfrac{K}{2} \ln 2\pi$

- Given an input x, we calculate the scores $S_j(x)$ for $j = 1, \ldots, J$ and we choose the label $j^\star$ with the highest score value

Discriminant analysis

The general case

- If we assume an homoscedastic model ($\Sigma_j = \Sigma$), the LDA score function becomes:

$$\begin{aligned} S_j(x) &= \ln c'' + \ln \pi_j - \frac{1}{2}(x-\mu_j)^\top \Sigma_j^{-1}(x-\mu_j) \\ &\propto \ln \pi_j + \mu_j^\top \Sigma^{-1} x - \frac{1}{2}\mu_j^\top \Sigma^{-1}\mu_j \end{aligned}$$

where $\ln c'' = \ln c' - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}x^\top \Sigma^{-1} x$

Remark

In practice, the parameters π_j, μ_j and Σ_j are unknown. We replace them by the corresponding estimates $\hat{\pi}_j$, $\hat{\mu}_j$ and $\hat{\Sigma}_j$. For the linear discriminant analysis, $\hat{\Sigma}$ is estimated by pooling all the classes.

Discriminant analysis

The general case

Example #2

We consider the classification problem of 33 observations with two explanatory variables X_1 and X_2, and three classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$:

i	$\mathcal{C}_j$	X_1	X_2	i	$\mathcal{C}_j$	X_1	X_2	i	$\mathcal{C}_j$	X_1	X_2
1	1	1.03	2.85	12	2	3.70	5.08	23	3	3.55	0.58
2	1	0.20	3.30	13	2	2.81	1.99	24	3	3.86	1.83
3	1	1.69	3.73	14	2	3.66	2.61	25	3	5.39	0.47
4	1	0.98	3.52	15	2	5.63	4.19	26	3	3.15	−0.18
5	1	0.98	5.15	16	2	3.35	3.64	27	3	4.93	1.91
6	1	3.47	6.56	17	2	2.97	3.55	28	3	3.87	2.61
7	1	3.94	4.68	18	2	3.16	2.92	29	3	4.09	1.43
8	1	1.55	5.99	19	3	3.00	0.98	30	3	3.80	2.11
9	1	1.15	3.60	20	3	3.09	1.99	31	3	2.79	2.10
10	2	1.20	2.27	21	3	5.45	0.60	32	3	4.49	2.71
11	2	3.66	5.49	22	3	3.59	−0.46	33	3	3.51	1.82

Discriminant analysis

The general case

Table: Parameter estimation of the discriminant analysis

Class	$\mathcal{C}_1$		$\mathcal{C}_2$		$\mathcal{C}_3$	
$\hat{\pi}_j$	0.273		0.273		0.455	
$\hat{\mu}_j$	1.666	4.376	3.349	3.527	3.904	1.367
$\hat{\Sigma}_j$	1.525	0.929	1.326	0.752	0.694	−0.031
	0.929	1.663	0.752	1.484	−0.031	0.960

For the LDA method, we have:

$$\hat{\Sigma} = \begin{pmatrix} 1.91355 & -0.71720 \\ -0.71720 & 3.01577 \end{pmatrix}$$

Discriminant analysis

The general case

Table: Computation of the discriminant scores $S_j(x)$

i	QDA			LDA			LDA^2		
	$S_1(x)$	$S_2(x)$	$S_3(x)$	$S_1(x)$	$S_2(x)$	$S_3(x)$	$S_1(x)$	$S_2(x)$	$S_3(x)$
1	−2.28	−3.69	−7.49	0.21	−0.96	−0.79	6.93	5.60	5.76
2	−2.28	−6.36	−12.10	−0.26	−2.17	−2.34	1.38	−2.13	−1.89
3	−1.76	−3.13	−6.79	2.84	2.16	1.71	12.13	12.01	11.38
4	−1.80	−4.43	−8.88	1.35	0.09	−0.22	7.73	6.20	5.93
5	−2.36	−7.75	−13.70	4.32	2.93	1.45	8.12	5.54	4.76
6	−3.16	−5.63	−14.68	10.75	11.36	8.95	14.82	13.99	12.96
7	−3.79	−1.92	−6.32	8.06	9.22	8.15	17.36	19.03	17.89
8	−2.85	−8.43	−15.23	6.73	5.76	3.70	10.47	8.09	7.15
9	−1.74	−4.12	−8.37	1.76	0.64	0.27	8.94	7.77	7.39
10	−3.14	−3.21	−6.17	−0.58	−1.56	−0.98	6.59	5.55	6.15
11	−2.87	−3.01	−9.45	9.10	9.96	8.31	16.89	17.65	16.42
12	−3.04	−2.38	−7.77	8.42	9.34	7.98	17.28	18.50	17.28
13	−6.32	−2.29	−1.62	1.41	1.82	2.64	12.48	13.94	14.46
14	−6.91	−2.07	−1.42	3.86	4.94	5.34	15.15	17.41	17.34
15	−9.79	−3.62	−7.12	9.79	12.43	11.75	12.58	14.01	13.50
16	−3.90	−1.47	−3.44	5.25	5.99	5.65	16.84	18.82	18.03
17	−3.31	−1.55	−3.61	4.50	4.92	4.63	16.25	17.95	17.21
18	−4.84	−1.60	−2.19	3.65	4.28	4.45	15.51	17.48	17.14
19	−10.21	−4.12	−1.27	−0.13	0.52	2.06	8.98	9.99	11.70
20	−7.05	−2.41	−1.24	1.85	2.50	3.32	12.99	14.72	15.22
21	−23.11	−11.16	−2.56	2.98	5.75	7.61	3.79	4.57	7.26
22	−19.22	−9.53	−2.42	−1.84	−0.57	2.01	1.81	1.53	5.51
23	−13.86	−5.92	−1.01	−0.01	1.15	2.98	7.65	8.67	10.95
24	−10.01	−3.43	−0.70	2.75	4.07	5.02	12.84	14.95	15.65
25	−23.48	−11.44	−2.54	2.65	5.38	7.33	3.40	4.09	6.95
26	−15.87	−7.59	−2.30	−2.01	−1.14	1.23	3.19	3.02	6.50
27	−14.09	−5.40	−1.52	4.56	6.78	7.70	11.17	13.24	14.08
28	−7.55	−2.27	−1.39	4.18	5.45	5.85	15.10	17.44	17.40
29	−12.40	−4.67	−0.61	2.38	3.92	5.17	11.21	13.14	14.33
30	−8.85	−2.87	−0.88	3.17	4.41	5.17	13.77	15.97	16.37
31	−5.97	−2.17	−1.72	1.58	1.97	2.70	12.78	14.26	14.67
32	−9.40	−2.97	−1.81	5.33	7.11	7.46	14.55	16.95	16.93
33	−8.84	−3.01	−0.80	2.19	3.21	4.16	12.82	14.77	15.45

Discriminant analysis

The general case

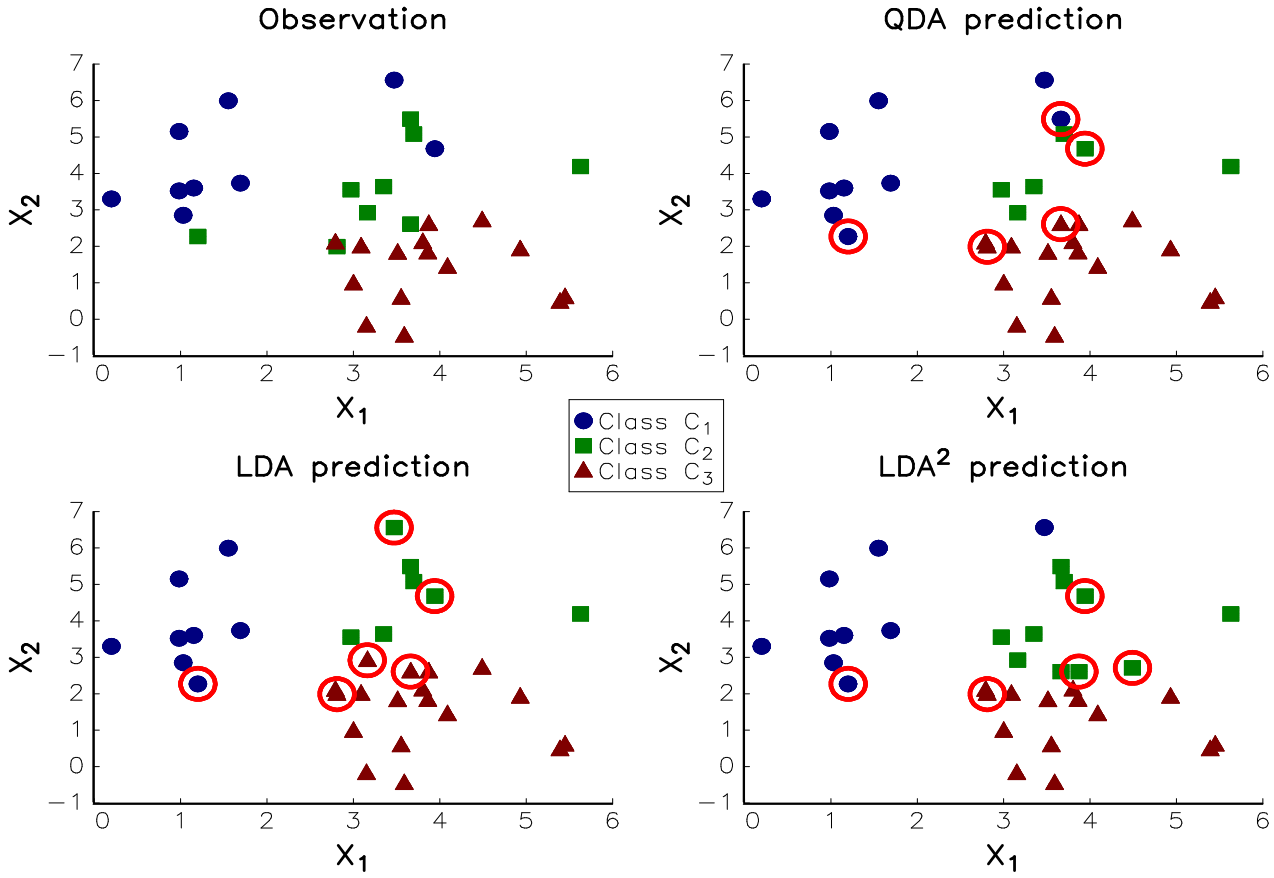


Figure: Comparing QDA, LDA and LDA^2 predictions

Discriminant analysis

The general case

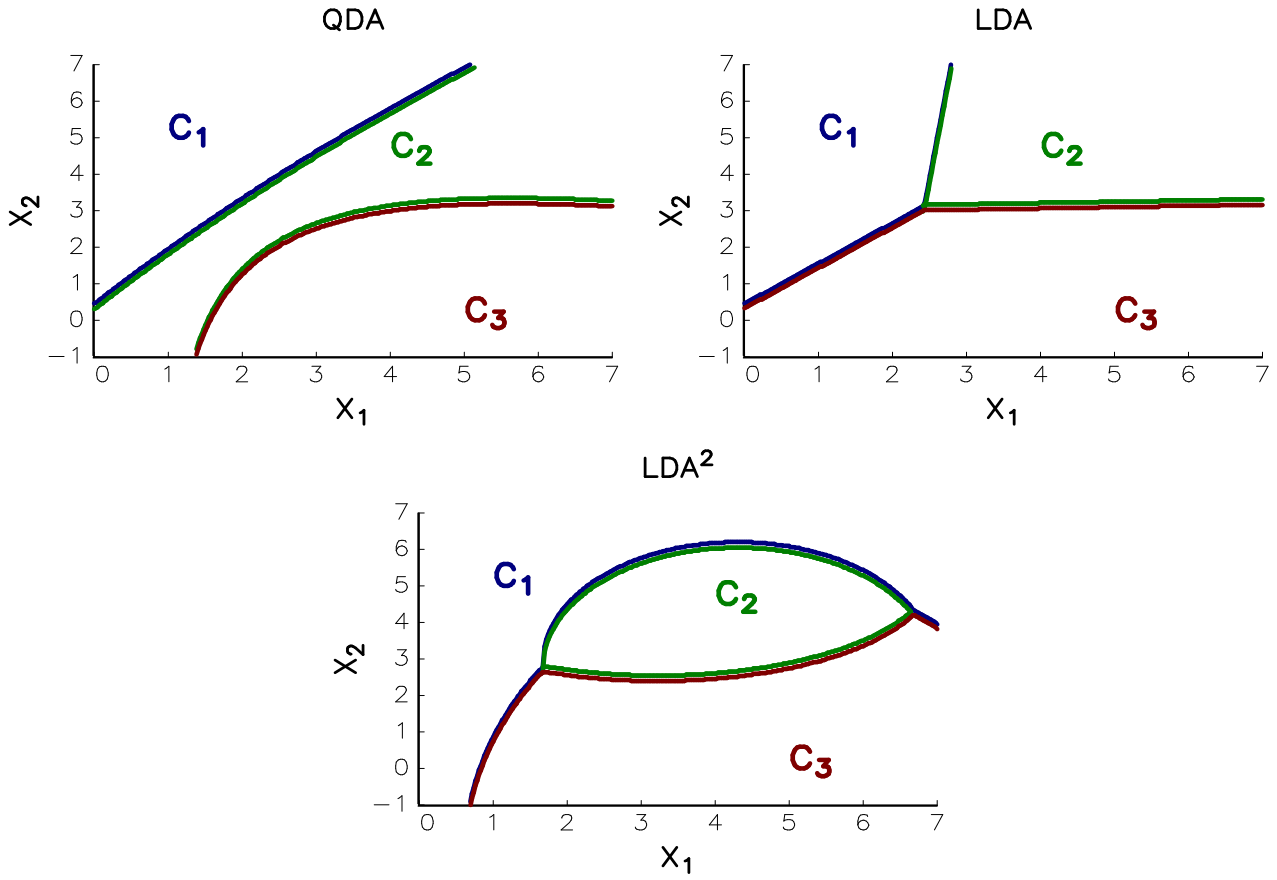


Figure: QDA, LDA and LDA^2 decision regions

Discriminant analysis

Class separation maximization

- We note $x_i = (x_{i,1}, \ldots, x_{i,K})$ the $K \times 1$ vector of exogenous variables X for the i^{th} observation
- The mean vector and the variance (or scatter) matrix of Class $\mathcal{C}_j$ is equal to $\hat{\mu}_j = \frac{1}{n_j} \sum_{i \in \mathcal{C}_j} x_i$ and $\mathbf{S}_j = n\hat{\Sigma}_j = \sum_{i \in \mathcal{C}_j} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^\top$ where n_j is the number of observations in the j^{th} class
- If consider the total population, we also have $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\mathbf{S} = n\hat{\Sigma} = \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^\top$

Discriminant analysis

Class separation maximization

- We notice that:

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^{J} n_j \hat{\mu}_j$$

- We define the between-class variance matrix as:

$$\mathbf{S}_B = \sum_{j=1}^{J} n_j \left(\hat{\mu}_j - \hat{\mu}\right) \left(\hat{\mu}_j - \hat{\mu}\right)^\top$$

and the within-class variance matrix as:

$$\mathbf{S}_W = \sum_{j=1}^{J} \mathbf{S}_j$$

- We can show that the total variance matrix can be decomposed into the sum of the within-class and between-class variance matrices:

$$\mathbf{S} = \mathbf{S}_W + \mathbf{S}_B$$

Discriminant analysis

Class separation maximization

- The discriminant analysis consists in finding the discriminant linear combination $\beta^\top X$ that has the maximum between-class variance relative to the within-class variance:

$$\beta^\star = \arg\max J(\beta)$$

where $J(\beta)$ is the Fisher criterion:

$$J(\beta) = \frac{\beta^\top \mathbf{S}_B \beta}{\beta^\top \mathbf{S}_W \beta}$$

- Since the objective function is invariant if we rescale the vector β – $J(\beta') = J(\beta)$ if $\beta' = c\beta$, we can impose that $\beta^\top \mathbf{S}_W \beta = 1$. It follows that:

$$\begin{aligned} \hat{\beta} &= \arg\max \beta^\top \mathbf{S}_B \beta \\ \text{s.t.} &\quad \beta^\top \mathbf{S}_W \beta = 1 \end{aligned}$$

Discriminant analysis

Class separation maximization

- The Lagrange function is:

$$\mathcal{L}(\beta;\lambda) = \beta^\top \mathbf{S}_B \beta - \lambda\left(\beta^\top \mathbf{S}_W \beta - 1\right)$$

- We deduce that the first-order condition is equal to:

$$\frac{\partial \mathcal{L}(\beta;\lambda)}{\partial \beta^\top} = 2\mathbf{S}_B\beta - 2\lambda\mathbf{S}_W\beta = \mathbf{0}$$

- It is remarkable that we obtain a generalized eigenvalue $\mathbf{S}_B\beta = \lambda\mathbf{S}_W\beta$ or equivalently:

$$\mathbf{S}_W^{-1}\mathbf{S}_B\beta = \lambda\beta$$

- Even if $\mathbf{S}_W$ and $\mathbf{S}_B$ are two symmetric matrices, it is not necessarily the case for the product $\mathbf{S}_W^{-1}\mathbf{S}_B$
- Using the eigendecomposition $\mathbf{S}_B = V\Lambda V^\top$, we have $\mathbf{S}_B^{1/2} = V\Lambda^{1/2}V^\top$

Discriminant analysis

Class separation maximization

- With the parametrization $\alpha = \mathbf{S}_B^{1/2}\beta$, the first-order condition becomes:

$$\mathbf{S}_B^{1/2}\mathbf{S}_W^{-1}\mathbf{S}_B^{1/2}\alpha = \lambda\alpha$$

  because $\beta = \mathbf{S}_B^{-1/2}\alpha$
- We have a right regular eigenvalue problem
- Let λ_k and v_k be the k^{th} eigenvalue and eigenvector of the symmetric matrix $\mathbf{S}_B^{1/2}\mathbf{S}_W^{-1}\mathbf{S}_B^{1/2}$
- It is obvious that the optimal solution $\alpha^\star$ is the first eigenvector v_1 corresponding to the largest eigenvalue λ_1
- We conclude that the estimator is $\hat{\beta} = \mathbf{S}_B^{-1/2}v_1$ and the discriminant linear relationship is $Y^c = v_1^\top \mathbf{S}_B^{-1/2}X$
- Moreover, we have:

$$\lambda_1 = J\left(\hat{\beta}\right) = \frac{\hat{\beta}^\top \mathbf{S}_B \hat{\beta}}{\hat{\beta}^\top \mathbf{S}_W \hat{\beta}}$$

Discriminant analysis

Class separation maximization

Example #3

We consider a problem with two classes $\mathcal{C}_1$ and $\mathcal{C}_2$, and two explanatory variables (X_1, X_2). Class $\mathcal{C}_1$ is composed of 7 observations: $(1, 2)$, $(1, 4)$, $(3, 6)$, $(3, 3)$, $(4, 2)$, $(5, 6)$, $(5, 5)$, whereas class $\mathcal{C}_2$ is composed of 6 observations: $(1, 0)$, $(2, 1)$, $(4, 1)$, $(3, 2)$, $(6, 4)$ and $(6, 5)$.

Discriminant analysis

Class separation maximization

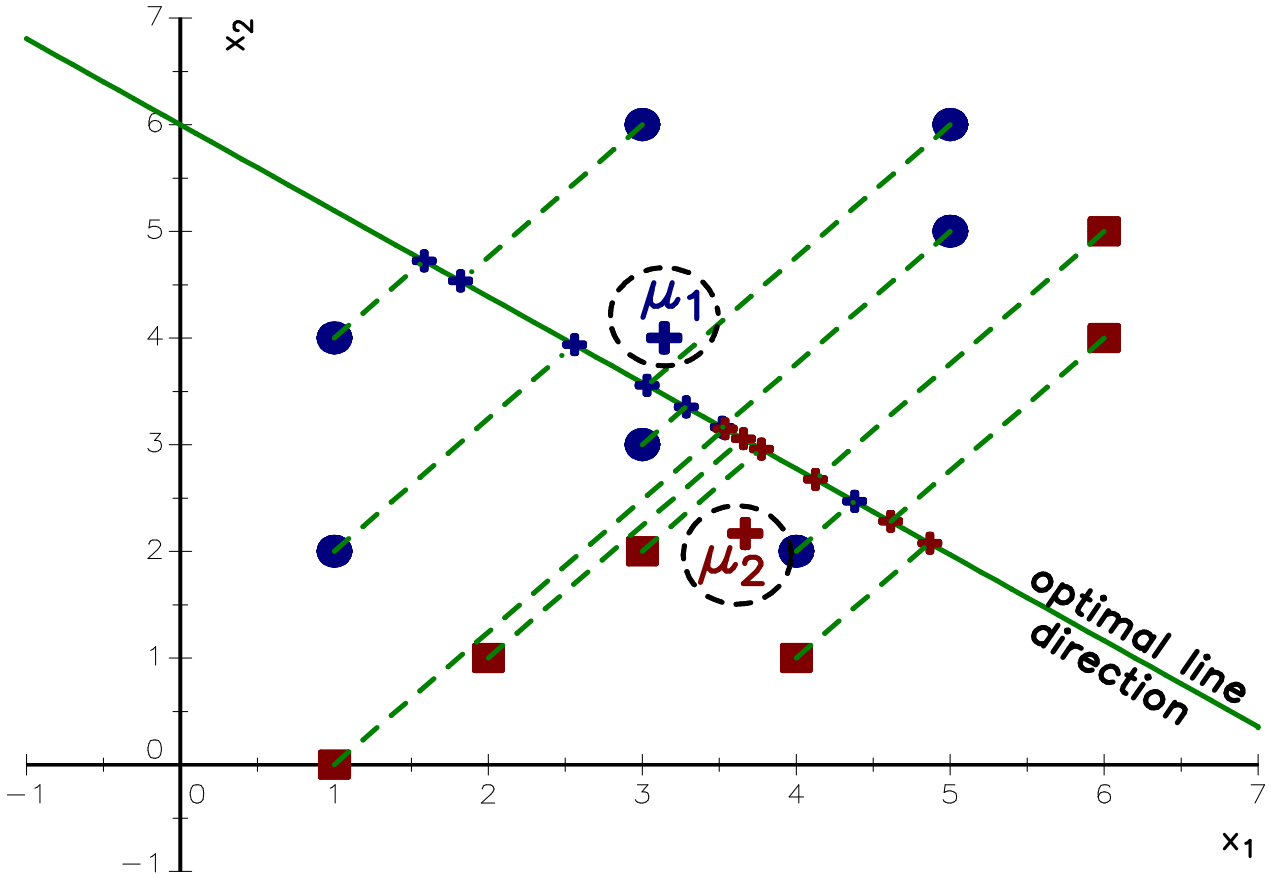


Figure: Linear projection and the Fisher solution

Discriminant analysis

Class separation maximization

Concerning the assignment decision, we can consider the midpoint rule:

$$\begin{cases} s_i < \bar{\mu} \Rightarrow i \in \mathcal{C}_1 \\ s_i > \bar{\mu} \Rightarrow i \in \mathcal{C}_2 \end{cases}$$

where $\bar{\mu} = (\bar{\mu}_1 + \bar{\mu}_2)/2$, $\bar{\mu}_1 = \beta^\top \hat{\mu}_1$ and $\bar{\mu}_2 = \beta^\top \hat{\mu}_2$

This rule is not optimal because it does not depend on the variance $\bar{s}_1^2$ and $\bar{s}_2^2$ of each class

Discriminant analysis

Class separation maximization

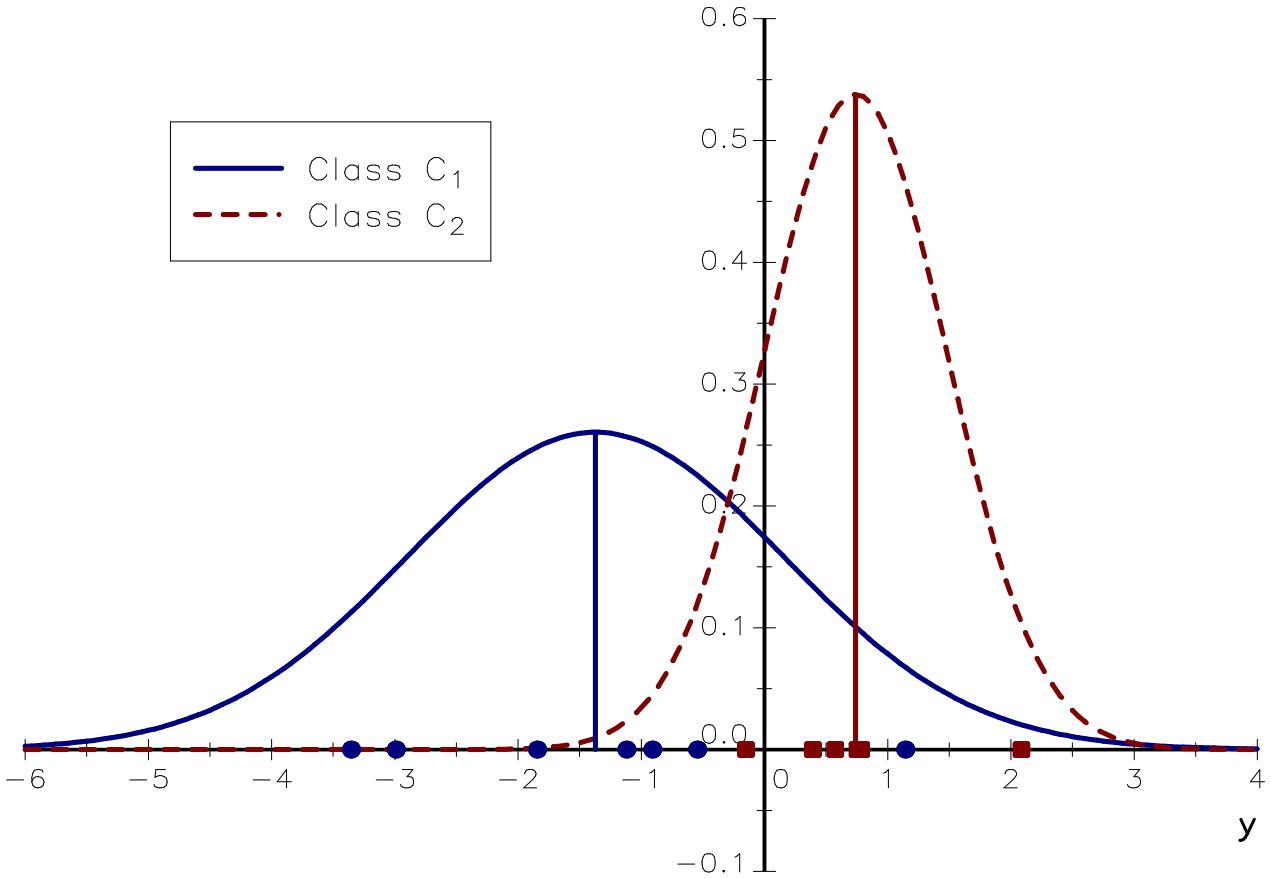


Figure: Class separation and the cut-off criterion

Binary choice models

General framework

- We assume that Y can take two values 0 and 1
- We consider models that link the outcome to a set of factors X:

$$\Pr\{Y = 1 \mid X = x\} = \mathbf{F}\left(x^\top \beta\right)$$

- $\mathbf{F}$ must be a cumulative distribution function in order to ensure that $\mathbf{F}(z) \in [0, 1]$
- We also assume that the model is symmetric, implying that $\mathbf{F}(z) + \mathbf{F}(-z) = 1$
- Given a sample $\{(x_i, y_i), i = 1, \ldots, n\}$, the log-likelihood function is equal to:

$$\ell(\theta) = \sum_{i=1}^{n} \ln \Pr\{Y_i = y_i\}$$

where y_i takes the values 0 or 1

Binary choice models

General framework

- We have:

$$\Pr\{Y_i = y_i\} = p_i^{y_i} \cdot (1 - p_i)^{1-y_i}$$

  where $p_i = \Pr\{Y_i = 1 \mid X_i = x_i\}$

- We deduce that:

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln (1 - p_i) \\ &= \sum_{i=1}^{n} y_i \ln \mathbf{F}\left(x_i^\top \beta\right) + (1 - y_i) \ln \left(1 - \mathbf{F}\left(x_i^\top \beta\right)\right) \end{aligned}$$

- We notice that the vector θ includes only the parameters β

Binary choice models

General framework

- By noting $f(z)$ the probability density function, it follows that the associated score vector of the log-likelihood function is:

$$\begin{aligned}\mathcal{S}(\beta) &= \frac{\partial \boldsymbol{\ell}(\beta)}{\partial \beta} \\ &= \sum_{i=1}^{n} \frac{f\left(x_i^\top \beta\right)}{\mathbf{F}\left(x_i^\top \beta\right)\mathbf{F}\left(-x_i^\top \beta\right)}\left(y_i - \mathbf{F}\left(x_i^\top \beta\right)\right) x_i\end{aligned}$$

Binary choice models

General framework

- The Hessian matrix is:

$$H(\beta) = \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^\top} = -\sum_{i=1}^{n} H_i \cdot \left(x_i x_i^\top\right)$$

where:

$$H_i = \frac{f\left(x_i^\top \beta\right)^2}{\mathbf{F}\left(x_i^\top \beta\right)\mathbf{F}\left(-x_i^\top \beta\right)} - \left(y_i - \mathbf{F}\left(x_i^\top \beta\right)\right) \cdot \left(\frac{f'\left(x_i^\top \beta\right)}{\mathbf{F}\left(x_i^\top \beta\right)\mathbf{F}\left(-x_i^\top \beta\right)} - \frac{f\left(x_i^\top \beta\right)^2\left(1 - 2\mathbf{F}\left(x_i^\top \beta\right)\right)}{\mathbf{F}\left(x_i^\top \beta\right)^2\mathbf{F}\left(-x_i^\top \beta\right)^2}\right)$$

Binary choice models

General framework

- Once $\hat{\beta}$ is estimated by the method of maximum likelihood, we can calculated the predicted probability for the i^{th} observation:

$$\hat{p}_i = \mathbf{F}\left(x_i^\top \hat{\beta}\right)$$

- Like a linear regression model, we can define the residual as the difference between the observation y_i and the predicted value $\hat{p}_i$
- We can also exploit the property that the conditional distribution of Y_i is a Bernoulli distribution $\mathcal{B}(p_i)$
- It is better to use the standardized (or Pearson) residuals:

$$\hat{u}_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i\left(1-\hat{p}_i\right)}}$$

- These residuals are related to the Pearson's chi-squared statistic:

$$\chi^2_{\text{Pearson}} = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{p}_i\right)^2}{\hat{p}_i\left(1-\hat{p}_i\right)}$$

Binary choice models

General framework

- This statistic may used to measure the goodness-of-fit of the model
- Under the assumption $\mathcal{H}_0$ that there is no lack-of-fit, we have $\chi^2_{\text{Pearson}} \sim \chi^2_{n-K}$ where K is the number of exogenous variables
- Another goodness-of-fit statistic is the likelihood ratio. For the '*saturated*' model, the estimated probability $\hat{p}_i$ is exactly equal to y_i
- We deduce that the likelihood ratio is equal to:

$$-2 \ln \Lambda = 2 \sum_{i=1}^{n} y_i \ln \left(\frac{y_i}{\hat{p}_i} \right) + (1 - y_i) \ln \left(\frac{1 - y_i}{1 - \hat{p}_i} \right)$$

Binary choice models

General framework

- In binomial choice models, $D = -2 \ln \Lambda$ is also called the deviance and we have $D \sim \chi^2_{n-K}$
- In a perfect fit $\hat{p}_i = y_i$, the likelihood ratio is exactly equal to zero
- The forecasting procedure consists of estimating the probability $\hat{p} = \mathbf{F}\left(x^\top \hat{\beta}\right)$ for a given set of variables x and to use the following decision criterion:

$$Y = 1 \Leftrightarrow \hat{p} \geq \frac{1}{2}$$

Binary choice models

Logistic regression

- The logit model uses the following cumulative distribution function:

$$\mathbf{F}(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{e^z+1}$$

- The probability density function is then equal to:

$$f(z) = \frac{e^{-z}}{\left(1+e^{-z}\right)^2}$$

- The log-likelihood function is equal to:

$$\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{n} (1-y_i) \ln\left(1-\mathbf{F}\left(x_i^\top \beta\right)\right) + y_i \ln \mathbf{F}\left(x_i^\top \beta\right) \\
&= \sum_{i=1}^{n} (1-y_i) \ln\left(\frac{e^{-x_i^\top \beta}}{1+e^{-x_i^\top \beta}}\right) - y_i \ln\left(1+e^{-x_i^\top \beta}\right) \\
&= -\sum_{i=1}^{n} \ln\left(1+e^{-x_i^\top \beta}\right) + (1-y_i)\left(x_i^\top \beta\right)
\end{aligned}$$

Binary choice models

Logistic regression

- We also have:

$$\mathcal{S}(\beta) = \sum_{i=1}^{n} \left(y_i - \mathbf{F}\left(x_i^\top \beta\right)\right) x_i$$

and:

$$H(\beta) = -\sum_{i=1}^{n} f\left(x_i^\top \beta\right) \cdot \left(x_i x_i^\top\right)$$

Binary choice models

Probit analysis

- The probit model assumes that $\mathbf{F}(z)$ is the Gaussian distribution
- The log-likelihood function is then:

$$\ell(\beta) = \sum_{i=1}^{n} (1 - y_i) \ln \left(1 - \Phi\left(x_i^\top \beta\right)\right) + y_i \ln \Phi\left(x_i^\top \beta\right)$$

- The probit model can be seen as a latent variable model
- Let us consider the linear model $Y^\star = \beta^\top X + U$ where $U \sim \mathcal{N}\left(0, \sigma^2\right)$
- We assume that we do not observe $Y^\star$ but $Y = g\left(Y^\star\right)$
- For example, if $g(z) = \mathbb{1}\{z > 0\}$, we obtain:

$$\Pr\{Y = 1 \mid X = x\} = \Pr\left\{\beta^\top X + U > 0 \mid X = x\right\} = \Phi\left(\frac{\beta^\top x}{\sigma}\right)$$

- We notice that only the ratio β/σ is identifiable
- Since we can set $\sigma = 1$, we obtain the probit model

Binary choice models

Regularization

- The regularized log-likelihood function is equal to:

$$\ell(\theta; \lambda) = \ell(\theta) - \frac{\lambda}{p} \|\theta\|_p^p$$

- The case $p = 1$ is equivalent to consider a lasso penalization
- The case $p = 2$ corresponds to the ridge regularization

Binary choice models

Extension to multinomial logistic regression

- We assume that Y can take J labels $(\mathfrak{L}_1, \ldots, \mathfrak{L}_J)$ or belongs to J disjoint classes $(\mathcal{C}_1, \ldots, \mathcal{C}_J)$
- We define the conditional probability as follows:

$$p_j(x) = \Pr\{Y = \mathfrak{L}_j \mid X = x\} = \Pr\{Y \in \mathcal{C}_j \mid X = x\} = \frac{e^{\beta_j^\top x}}{1 + \sum_{j=1}^{J-1} e^{\beta_j^\top x}}$$

- The probability of the last label is then equal to:

$$p_J(x) = 1 - \sum_{j=1}^{J-1} p_j(x) = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\beta_j^\top x}}$$

- The log-likelihood function becomes:

$$\ell(\theta) = \sum_{i=1}^{n} \ln \left(\prod_{j=1}^{J} p_j(x_i)^{i \in \mathcal{C}_j} \right)$$

where θ is the vector of parameters $(\beta_1, \ldots, \beta_{J-1})$

Non-parametric supervised methods

- k-nearest neighbor classifier (k-NN)
- Neural networks (NN)
- Support vector machines (SVM)
- Model averaging (bagging or bootstrap aggregation, random forests, boosting)

Definition and properties

- The entropy is a measure of unpredictability or uncertainty of a random variable
- Let (X, Y) be a random vector where $p_{i,j} = \Pr\{X = x_i, Y = y_j\}$, $p_i = \Pr\{X = x_i\}$ and $p_j = \Pr\{Y = y_j\}$
- The Shannon entropy of the discrete random variable X is given by:

$$H(X) = -\sum_{i=1}^{n} p_i \ln p_i$$

- We have the property $0 \leq H(X) \leq \ln n$.
- The Shannon entropy is a measure of the average information of the system
- The lower the Shannon entropy, the more informative the system

Definition and properties

- For a random vector (X, Y), we have:

$$H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{n} p_{i,j} \ln p_{i,j}$$

- We deduce that the conditional information of Y given X is equal to:

$$\begin{aligned} H(Y \mid X) &= \mathbb{E}_X [H(Y \mid X = x)] \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{n} p_{i,j} \ln \frac{p_{i,j}}{p_i} \\ &= H(X, Y) - H(X) \end{aligned}$$

Definition and properties

We have the following properties:

- if X and Y are independent, we have $H(Y \mid X) = H(Y)$ and $H(X, Y) = H(Y) + H(X)$;
- if X and Y are perfectly dependent, we have $H(Y \mid X) = 0$ and $H(X, Y) = H(X)$.

The amount of information obtained about one random variable, through the other random variable is measured by the mutual information:

$$\begin{aligned} I(X, Y) &= H(Y) + H(X) - H(X, Y) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} p_{i,j} \ln \frac{p_{i,j}}{p_i p_j} \end{aligned}$$

Definition and properties

$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$

$$H(X) = H(Y) = 1.792$$
$$H(X, Y) = 3.584$$
$$I(X, Y) = 0$$

$1/6$					
	$1/6$				
		$1/6$			
			$1/6$		
				$1/6$	
					$1/6$

$$H(X) = H(Y) = 1.792$$
$$H(X, Y) = 1.792$$
$$I(X, Y) = 1.792$$

Figure: Examples of Shannon entropy calculation

Definition and properties

$1/24$	$1/24$				
$1/24$	$1/24$	$1/24$	$1/48$		
	$1/24$	$1/6$	$1/24$	$1/48$	
	$1/48$	$1/24$	$1/6$	$1/24$	
		$1/48$	$1/24$	$1/24$	$1/24$
				$1/24$	$1/24$

$$H(X) = H(Y) = 1.683$$
$$H(X, Y) = 2.774$$
$$I(X, Y) = 0.593$$

					$1/12$
$1/8$			$1/8$		
	$1/24$				
$5/24$		$1/24$			
$3/24$				$1/24$	
$3/24$	$1/24$	$1/24$			

$$H(X) = 1.658$$
$$H(Y) = 1.328$$
$$I(X, Y) = 0.750$$

Figure: Examples of Shannon entropy calculation

Application to scoring

- Let S and Y be the score and the control variable
- For instance, Y is a binary random variable that may indicate a bad credit ($Y = 0$) or a good credit ($Y = 1$)
- We consider the following decision rule:

$$\begin{cases} S \leq 0 \Rightarrow S^\star = 0 \\ S > 0 \Rightarrow S^\star = 1 \end{cases}$$

Application to scoring

- We note $n_{i,j}$ the number of observations such that $S^\star = i$ and $Y = j$. We obtain the following system $(S^\star, Y)$:

	$Y = 0$	$Y = 1$
$S^\star = 0$	$n_{0,0}$	$n_{0,1}$
$S^\star = 1$	$n_{1,0}$	$n_{1,1}$

  where $n = n_{0,0} + n_{0,1} + n_{1,0} + n_{1,1}$ is the total number of observations
- The hit rate is the ratio of good bets:

$$H = \frac{n_{0,0} + n_{1,1}}{n}$$

- This statistic can be viewed as an information measure of the system (S, Y)
- When there are more states, we can consider the Shannon entropy

Application to scoring

	y_1	y_2	y_3	y_4	y_5
s_1	10	9			
s_2	7	9			
s_3	3		7	2	
s_4		2	10	4	5
s_5				10	2
s_6			3	4	13

$H(S_1) = 1.767$
$H(Y) = 1.609$
$H(S_1, Y) = 2.614$
$I(S_1, Y) = 0.763$

	y_1	y_2	y_3	y_4	y_5
s_1	7	10			
s_2	10	8			
s_3			5	4	3
s_4	3		10	6	4
s_5	2			5	8
s_6			5	5	5

$H(S_1) = 1.771$
$H(Y) = 1.609$
$H(S_1, Y) = 2.745$
$I(S_1, Y) = 0.636$

Figure: Scorecards S_1 and S_2

Graphical methods

- We assume that the control variable Y can takes two values
  - $Y = 0$ corresponds to a bad risk (or bad signal)
  - $Y = 1$ corresponds to a good risk (or good signal)

Graphical methods

- We assume that the probability $\Pr\{Y = 1 \mid S \geq s\}$ is increasing with respect to the level $s \in [0, 1]$, which corresponds to the rate of acceptance.
- We deduce that the decision rule is the following:
  - if the score of the observation is above the threshold s, the observation is selected;
  - if the score of the observation is below the threshold s, the observation is not selected.
- If s is equal to one, we select no observation
- If s is equal to zero, we select all the observations

Performance curve

- The performance curve is the parametric function $y = \mathcal{P}(x)$ defined by:

$$\begin{cases} x(s) = \Pr\{S \geq s\} \\ y(s) = \dfrac{\Pr\{Y = 0 \mid S \geq s\}}{\Pr\{Y = 0\}} \end{cases}$$

  where $x(s)$ corresponds to the proportion of selected observations and $y(s)$ corresponds to the ratio between the proportion of selected bad risks and the proportion of bad risks in the population
- The score is efficient if the ratio is below one
- If $y(s) > 1$, the score selects more bad risks than those we can find in the population
- If $y(s) = 1$, the score is random and the performance is equal to zero. In this case, the selected population is representative of the total population

Selection curve

- The selection curve is the parametric curve $y = \mathcal{S}(x)$ defined by:

$$\begin{cases} x(s) = \Pr\{S \geq s\} \\ y(s) = \Pr\{S \geq s \mid Y = 0\} \end{cases}$$

  where $y(s)$ corresponds to the ratio of observations that are wrongly selected
- By construction, we would like that the curve $y = \mathcal{S}(x)$ is located below the bisecting line $y = x$ in order to verify that $\Pr\{S \geq s \mid Y = 0\} < \Pr\{S \geq s\}$

Performance and selection curves

- We have:

$$\begin{aligned}
\Pr\{S \geq s \mid Y = 0\} &= \frac{\Pr\{S \geq s, Y = 0\}}{\Pr\{Y = 0\}} \\
&= \Pr\{S \geq s\} \cdot \frac{\Pr\{S \geq s, Y = 0\}}{\Pr\{S \geq s\} \Pr\{Y = 0\}} \\
&= \Pr\{S \geq s\} \cdot \frac{\Pr\{Y = 0 \mid S \geq s\}}{\Pr\{Y = 0\}}
\end{aligned}$$

- The performance and selection curves are related as follows:

$$\mathcal{S}(x) = x\mathcal{P}(x)$$

Discriminant curve

- The discriminant curve is the parametric curve $y = \mathcal{D}(x)$ defined by:

$$\mathcal{D}(x) = g_1\left(g_0^{-1}(x)\right)$$

where:

$$g_y(s) = \Pr\{S \geq s \mid Y = y\}$$

- It represents the proportion of good risks in the selected population with respect to the proportion of bad risks in the selected population
- The score is said to be discriminant if the curve $y = \mathcal{D}(x)$ is located above the bisecting line $y = x$

Some properties

1. the performance curve (respectively, the selection curve) is located below the line $y = 1$ (respectively, the bisecting line $y = x$) if and only if $\operatorname{cov}(f(Y), g(S)) \geq 0$ for any increasing functions f and g
2. the performance curve is increasing if and only if:

$$\operatorname{cov}(f(Y), g(S) \mid S \geq s) \geq 0$$

   for any increasing functions f and g, and any threshold level s
3. the selection curve is convex if and only if $\mathbb{E}[f(Y) \mid S = s]$ is increasing with respect to the threshold level s for any increasing function f
4. We can show that (3) $\Rightarrow$ (2) $\Rightarrow$ (1)

Some properties

- A score is perfect or optimal if there is a threshold level $s^\star$ such that $\Pr\{Y = 1 \mid S \geq s^\star\} = 1$ and $\Pr\{Y = 0 \mid S < s^\star\} = 1$
- It separates the population between good and bad risks
- Graphically, the selection curve of a perfect score is equal to:

$$y = \mathbb{1}\{x > \Pr\{Y = 1\}\} \cdot \left(1 + \frac{x - 1}{\Pr\{Y = 0\}}\right)$$

- Using the relationship $\mathcal{S}(x) = x\mathcal{P}(x)$, we deduce that the performance curve of a perfect score is given by:

$$y = \mathbb{1}\{x > \Pr\{Y = 1\}\} \cdot \left(\frac{x - \Pr\{Y = 1\}}{x \cdot \Pr\{Y = 0\}}\right)$$

- For the discriminant curve, a perfect score satisfies $\mathcal{D}(x) = 1$
- When the score is random, we have $\mathcal{S}(x) = \mathcal{D}(x) = x$ and $\mathcal{P}(x) = 1$

Some properties

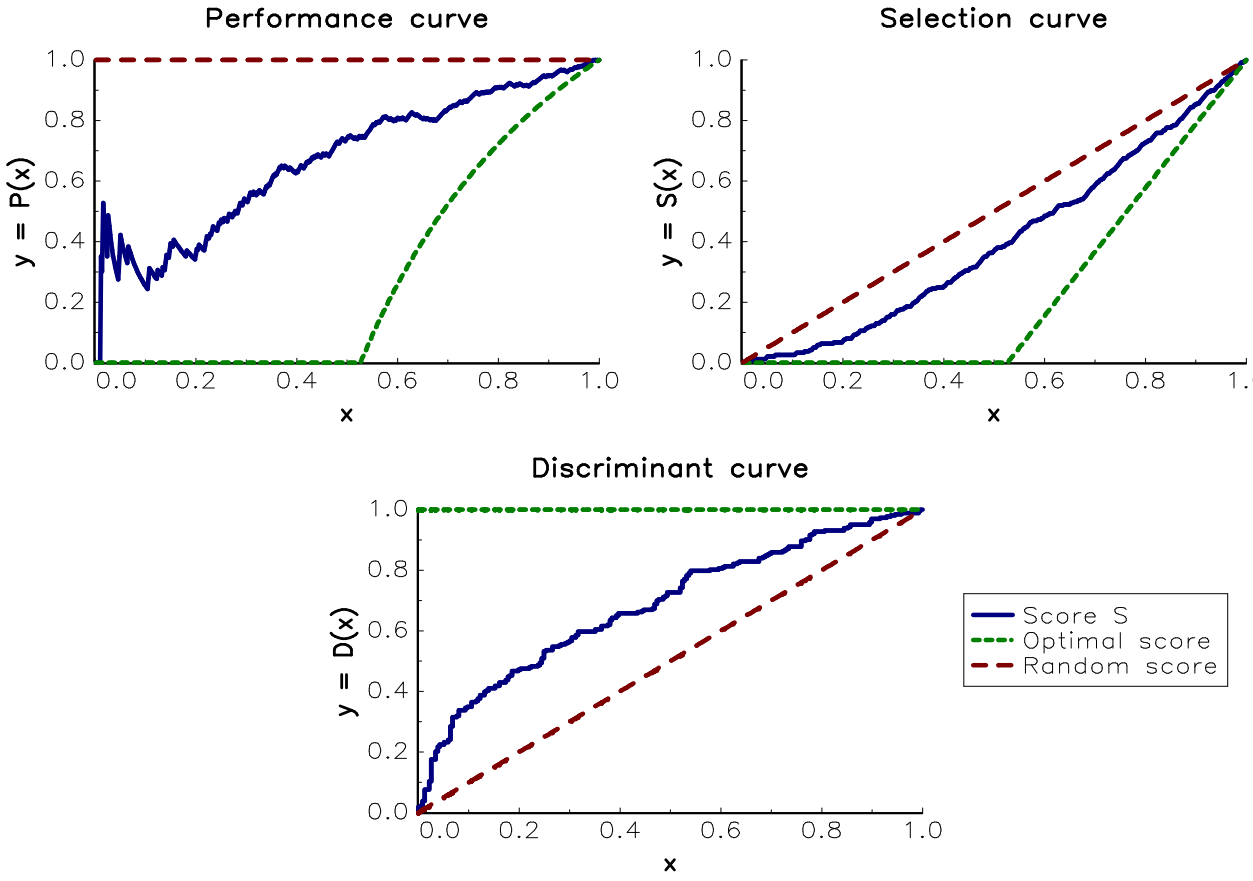


Figure: Performance, selection and discriminant curves

Some properties

- The score S_1 is more performing on the population P_1 than the score S_2 on the population P_2 if and only if the performance (or selection) curve of (S_1, P_1) is below the performance (or selection) curve of (S_2, P_2)
- The score S_1 is more discriminatory on the population P_1 than the score S_2 on the population P_2 if and only if the discriminant curve of (S_1, P_1) is above the discriminant curve of (S_2, P_2)

Some properties

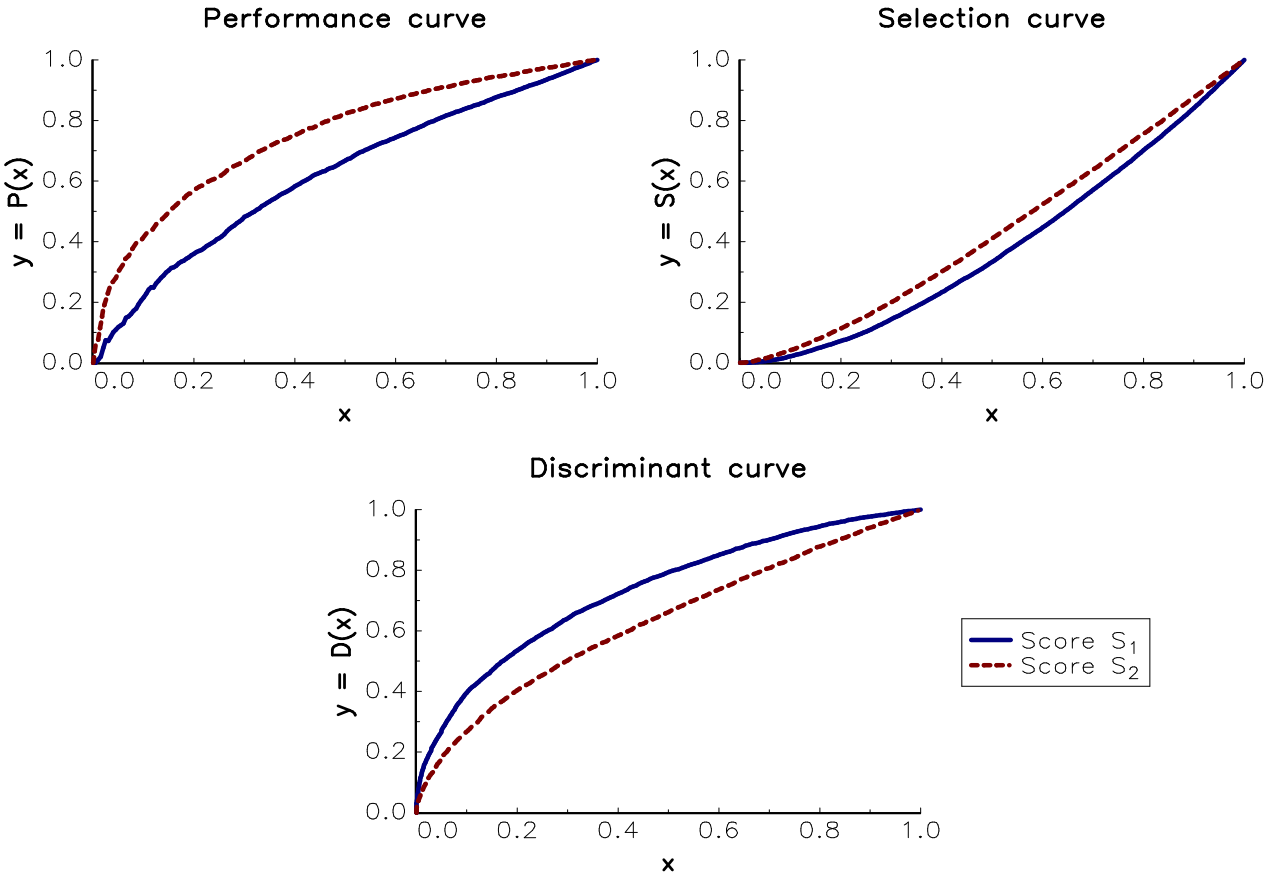


Figure: The score S_1 is better than the score S_2

Some properties

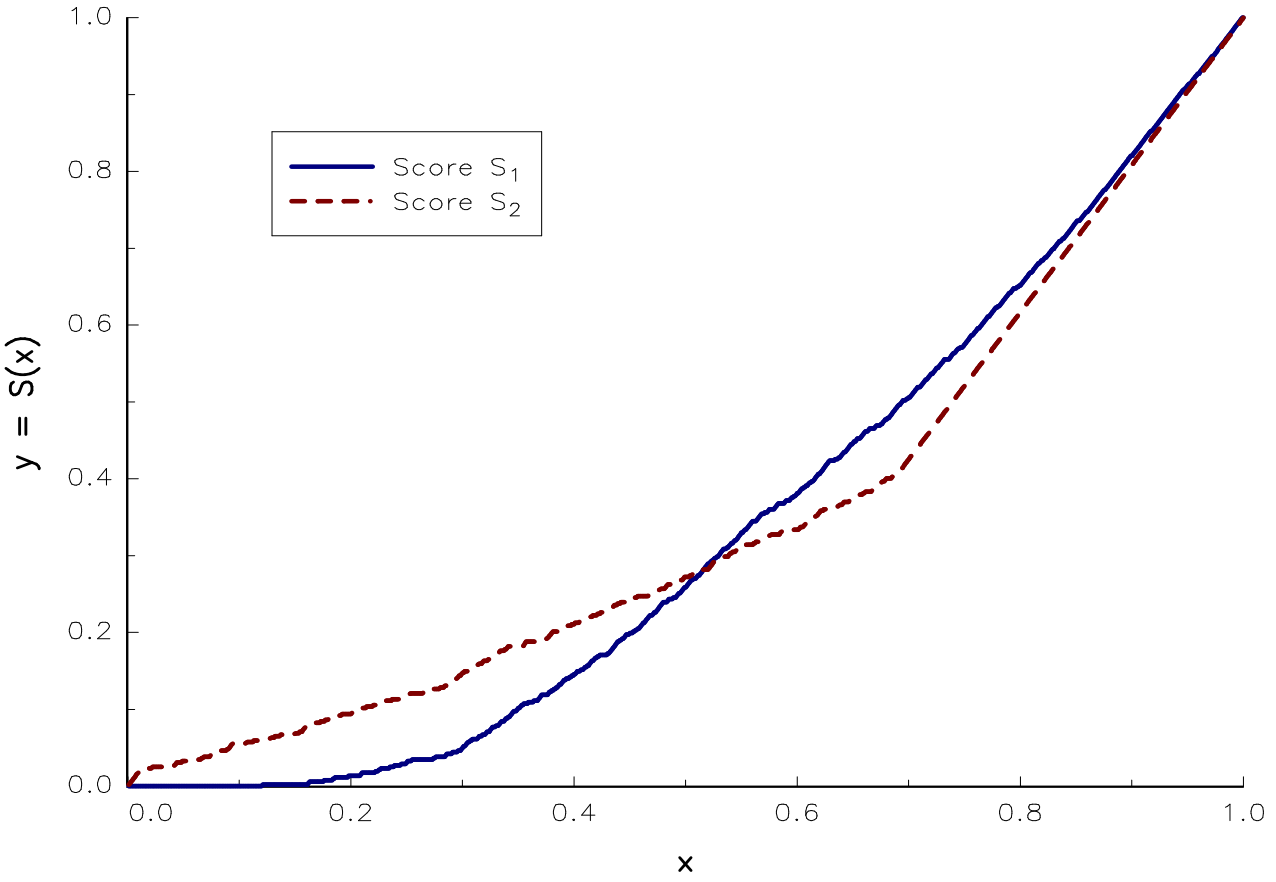


Figure: Illustration of the partial ordering between two scores

Kolmogorov-Smirnov test

- We consider the cumulative distribution functions:

$$\mathbf{F}_0(s) = \Pr\{S \leq s \mid Y = 0\}$$

and:

$$\mathbf{F}_1(s) = \Pr\{S \leq s \mid Y = 1\}$$

- The score S is relevant if we have the stochastic dominance order $\mathbf{F}_0 \succ \mathbf{F}_1$
- In this case, the score quality is measured by the Kolmogorov-Smirnov statistic:

$$\mathrm{KS} = \max_s |\mathbf{F}_0(s) - \mathbf{F}_1(s)|$$

It takes the value 1 if the score is perfect

- The KS statistic may be used to verify that the score is not random ($\mathcal{H}_0 : \mathrm{KS} = 0$)

Kolmogorov-Smirnov test

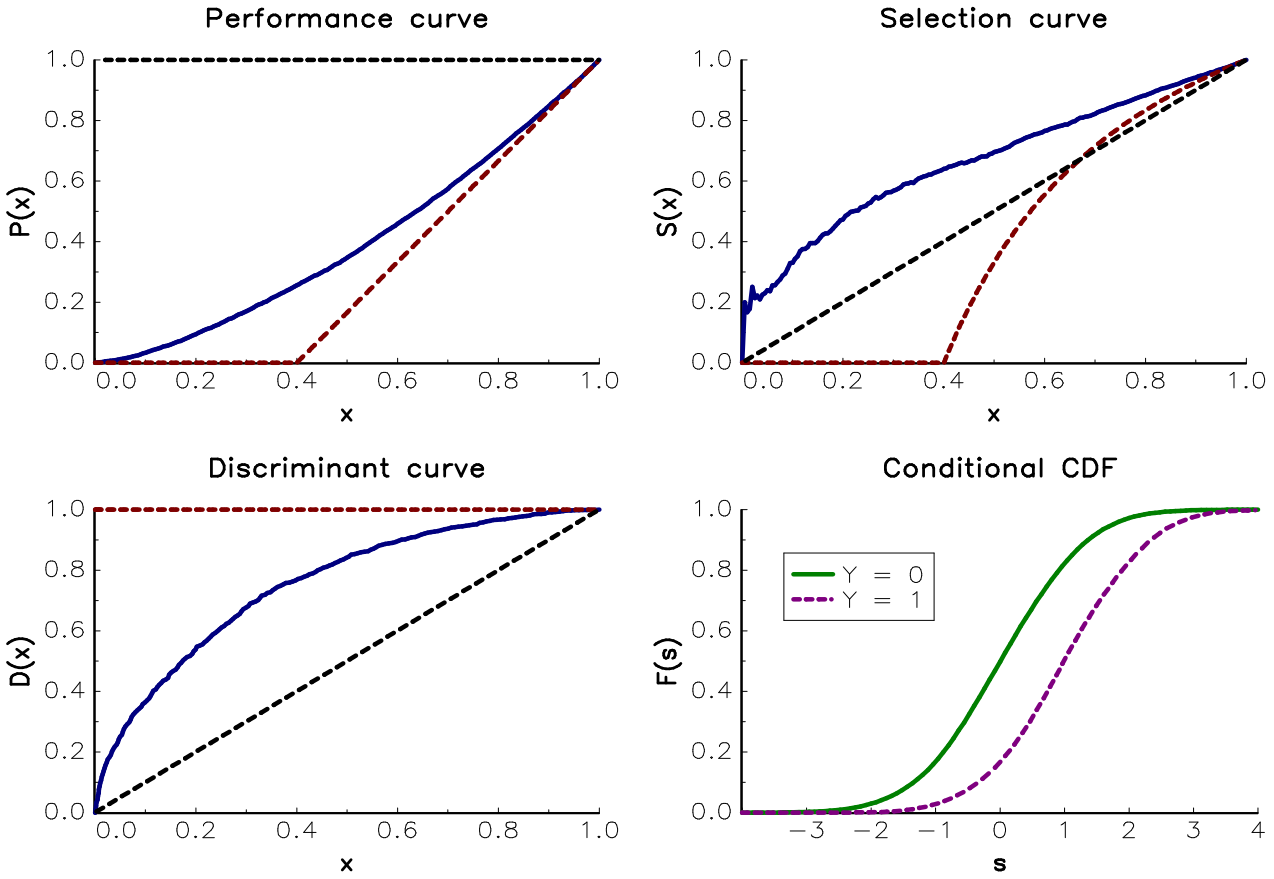


Figure: Comparison of the distributions $\mathbf{F}_0(s)$ and $\mathbf{F}_1(s)$

Gini coefficient

The Lorenz curve

- Let X and Y be two random variables
- The Lorenz curve $y = \mathcal{L}(x)$ is the parametric curve defined by:

$$\begin{cases} x = \Pr\{X \leq x\} \\ y = \Pr\{Y \leq y \mid X \leq x\} \end{cases}$$

- In economics, x represents the proportion of individuals that are ranked by income while y represents the proportion of income
- In this case, the Lorenz curve is a graphical representation of the distribution of income and is used for illustrating inequality of the wealth distribution between individuals

Gini coefficient

The Lorenz curve

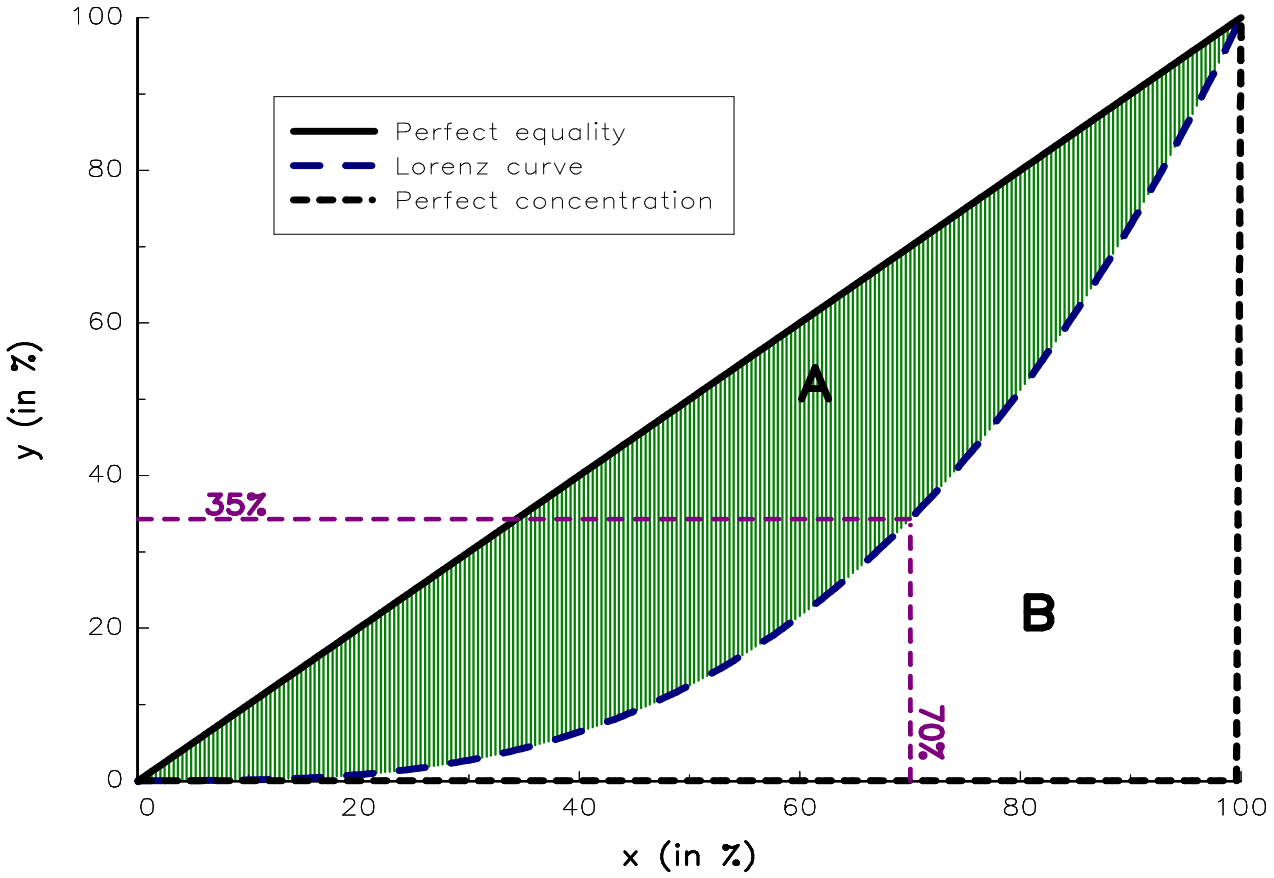


Figure: An example of Lorenz curve

Gini coefficient

Definition

- We define the Gini coefficient by:

$$\mathcal{G}ini\left(\mathcal{L}\right) = \frac{A}{A+B}$$

  where A is the area between the Lorenz curve and the curve of perfect equality, and B is the area between the curve of perfect concentration and the Lorenz curve
- By construction, we have $0 \leq \mathcal{G}ini\left(\mathcal{L}\right) \leq 1$
- The Gini coefficient is equal to zero in the case of perfect equality and one in the case of perfect concentration
- We have:

$$\mathcal{G}ini\left(\mathcal{L}\right) = 1 - 2\int_0^1 \mathcal{L}\left(x\right)\,\mathrm{d}x$$

Gini coefficient

Application to credit scoring

- We can interpret the selection curve as a Lorenz curve
- We recall that $\mathbf{F}(s) = \Pr\{S \leq s\}$, $\mathbf{F}_0(s) = \Pr\{S \leq s \mid Y = 0\}$ and $\mathbf{F}_1(s) = \Pr\{S \leq s \mid Y = 1\}$
- The selection curve is defined by the following parametric coordinates:

$$\begin{cases} x(s) = 1 - \mathbf{F}(s) \\ y(s) = 1 - \mathbf{F}_0(s) \end{cases}$$

- The selection curve measures the capacity of the score for not selecting bad risks
- We could also build the Lorenz curve that measures the capacity of the score for selecting good risks:

$$\begin{cases} x(s) = \Pr\{S \geq s\} = 1 - \mathbf{F}(s) \\ y(s) = \Pr\{S \geq s \mid Y = 1\} = 1 - \mathbf{F}_1(s) \end{cases}$$

It is called the precision curve

Gini coefficient

Application to credit scoring

- Another popular graphical tool is the receiver operating characteristic (or ROC curve), which is defined by:

$$\begin{cases} x(s) = \Pr\{S \geq s \mid Y = 0\} = 1 - \mathbf{F}_0(s) \\ y(s) = \Pr\{S \geq s \mid Y = 1\} = 1 - \mathbf{F}_1(s) \end{cases}$$

- The Gini coefficient associated to the Lorenz curve $\mathcal{L}$ becomes:

$$\mathcal{G}ini(\mathcal{L}) = 2\int_0^1 \mathcal{L}(x)\,\mathrm{d}x - 1$$

- The Gini coefficient of the score S is then computed as follows:

$$\mathcal{G}ini^{\star}(S) = \frac{\mathcal{G}ini(\mathcal{L})}{\mathcal{G}ini(\mathcal{L}^{\star})}$$

where $\mathcal{L}^{\star}$ is the Lorenz curve associated to the perfect score

- An alternative to the Gini coefficient is the AUC measure, which corresponds to the area under the ROC curve:

$$\mathcal{G}ini(\mathrm{ROC}) = 2 \times \mathrm{AUC}(\mathrm{ROC}) - 1$$

Gini coefficient

Application to credit scoring

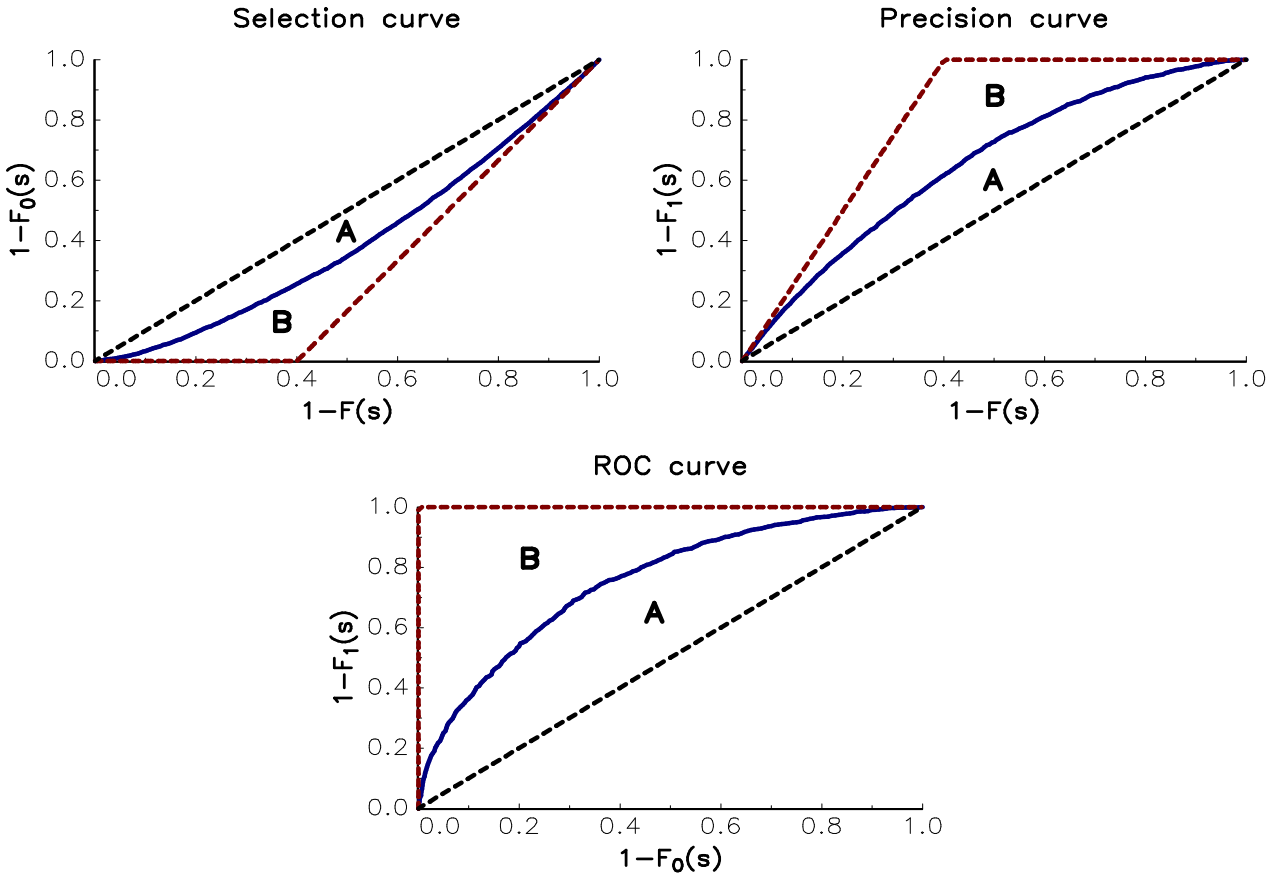


Figure: Selection, precision and ROC curves

Choice of the optimal cut-off

Confusion matrix

- A confusion matrix is a special case of contingency matrix
- Each row of the matrix represents the frequency in a predicted class while each column represents the frequency in an actual class
- Using the test set, it takes the following form:

	$Y = 0$	$Y = 1$
$S < s$	$n_{0,0}$	$n_{0,1}$
$S \geq s$	$n_{1,0}$	$n_{1,1}$
	$n_0 = n_{0,0} + n_{1,0}$	$n_1 = n_{0,1} + n_{1,1}$

where $n_{i,j}$ represents the number of observations of the cell (i, j)

Choice of the optimal cut-off

Confusion matrix

- We notice that each cell of this table can be interpreted as follows:

	$Y = 0$	$Y = 1$
$S < s$	It is rejected and it is a bad risk (true negative)	It is rejected, but it is a good risk (false negative)
$S \geq s$	It is accepted, but it is a bad risk (false positive)	It is accepted and it is a good risk (true positive)
	(negative)	(positive)

- The cells $(S < s, Y = 0)$ and $(S \geq s, Y = 1)$ correspond to observations that are well-classified: true negative (TN) and true positive (TP)
- The cells $(S \geq s, Y = 0)$ and $(S < s, Y = 1)$ correspond to two types of errors:
  1. a false positive (FP) can induce a future loss, because it may default: this is a type I error
  2. a false negative (FN) potentially corresponds to a loss of a future P&L: this is a type II error

Choice of the optimal cut-off

Classification ratios

- We have

$$\begin{aligned}
\text{True Positive Rate} \quad & \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\
\text{False Negative Rate} \quad & \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} = 1 - \mathrm{TPR} \\
\text{True Negative Rate} \quad & \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\
\text{False Positive Rate} \quad & \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}
\end{aligned}$$

- The true positive rate (TPR) is also known as the sensitivity or the recall
- It measures the proportion of real good risks that are correctly predicted good risk

Choice of the optimal cut-off

Classification ratios

- The precision or the positive predictive value (PPV) is

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

  It measures the proportion of predicted good risks that are correctly real good risk

- The accuracy considers the classification of both negatives and positives:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

- The F_1 score is the harmonic mean of precision and sensitivity:

$$F_1 = \frac{2}{1/\mathrm{precision} + 1/\mathrm{sensitivity}} = \frac{2 \cdot \mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}}$$

Choice of the optimal cut-off

Classification ratios

Table: Confusion matrix of three scoring systems and three cut-off values s

Score	$s = 100$		$s = 200$		$s = 500$	
S_1	386	616	698	1 304	1 330	3 672
	1 614	7 384	1 302	6 696	670	4 328
S_2	372	632	700	1 304	1 386	3 616
	1 628	7 368	1 300	6 696	614	4 384
S_3	382	616	656	1 344	1 378	3 624
	1 618	7 384	1 344	6 656	622	4 376
Perfect	1 000	0	2 000	0	2 000	3 000
	1 000	8 000	0	8 000	0	5 000

Choice of the optimal cut-off

Classification ratios

Table: Binary classification ratios (in %) of the three scoring systems

Score	s	TPR	FNR	TNR	FPR	PPV	ACC	F_1
S_1	100	92.3	7.7	19.3	80.7	82.1	77.7	86.9
	200	83.7	16.3	34.9	65.1	83.7	73.9	83.7
	500	54.1	45.9	66.5	33.5	86.6	56.6	66.6
S_2	100	92.1	7.9	18.6	81.4	81.9	77.4	86.7
	200	83.7	16.3	35.0	65.0	83.7	74.0	83.7
	500	54.8	45.2	69.3	30.7	87.7	57.7	67.5
S_3	100	92.3	7.7	19.1	80.9	82.0	77.7	86.9
	200	83.2	16.8	32.8	67.2	83.2	73.1	83.2
	500	54.7	45.3	68.9	31.1	87.6	57.5	67.3
Perfect	100	100.0	0.0	50.0	50.0	88.9	90.0	94.1
	200	100.0	0.0	100.0	0.0	100.0	100.0	100.0
	500	62.5	37.5	100.0	0.0	100.0	70.0	76.9

Choice of the optimal cut-off

Classification ratios

Table: Best scoring system

Cut-off	TPR	FNR	TNR	FPR	PPV	ACC	F_1
100	S_1/S_3	S_1/S_3	S_1	S_1	S_1	S_1	S_1
200	S_1/S_2	S_1/S_2	S_2	S_2	S_2	S_2	S_2
500	S_2	S_2	S_2	S_2	S_2	S_2	S_2

Exercises

- Exercise 15.4.5 – Two-class separation maximization
- Exercise 15.4.6 – Maximum likelihood estimation of the probit model

References

Gouriéroux, C., and Jasiak, J. (2007)
The Econometrics of Individual Risk: Credit, Insurance, and Marketing, Princeton University Press.

Roncalli, T. (2020)
Handbook of Financial Risk Management, Chapman and Hall/CRC Financial Mathematics Series, Chapter 15.