Lineare Diskriminanzanalyse (LDA)

Multivariate Analyse

Mathematischer Hintergrund, durchgerechnetes Beispiel, Handhabung des Tools

Deutsch

Wolfgang Böhme

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

8. November 2026

Dokumentationen dieser Webseite wurden mit Unterstützung der KI Claude (Anthropic) entwickelt.

Inhaltsverzeichnis

1 Einleitung und Anwendungsgebiete **2**

2 Mathematischer Hintergrund **2**
2.1 Streumatrizen . . . 2
2.2 Diskriminanzfunktionen . . . 2
2.3 Diskriminanzscores . . . 3
2.4 Standardisierte Koeffizienten . . . 3
2.5 Kanonische Korrelation . . . 3
2.6 Wilks' Lambda . . . 4
2.7 Bartlett-Test auf Signifikanz . . . 4
2.8 Leave-One-Out-Kreuzvalidierung (LOO-CV) . . . 4

3 Durchgerechnetes Beispiel: Bilanzkennzahlen **4**
3.1 Datenbasis . . . 4
3.2 Schritt 1: Gruppenstatistiken . . . 5
3.3 Schritt 2: Diskriminanzfunktionen . . . 5
3.4 Schritt 3: Standardisierte Diskriminanzkoeffizienten . . . 5
3.5 Schritt 4: Gruppencentroide im Diskriminanzraum . . . 6
3.6 Schritt 5: Klassifikation und Kreuzvalidierung . . . 6

4 Handhabung des interaktiven LDA-Tools **7**
4.1 Dateneingabe . . . 7
4.2 Parameterauswahl . . . 7
4.3 Ergebnistabs . . . 7
4.4 Typischer Arbeitsablauf . . . 8

5 Formelübersicht **9**

6 Weiterführende Literatur **9**

Einleitung und Anwendungsgebiete

Die **Lineare Diskriminanzanalyse** (LDA) ist ein *ueberwachtes* Klassifikationsverfahren: Im Gegensatz zur Clusteranalyse sind die Gruppenzugehoerigkeiten der Beobachtungen bekannt. Die LDA findet **Diskriminanzfunktionen** – lineare Kombinationen der Variablen – die die Gruppen bestmoeglich voneinander trennen.

Duale Zielsetzung der LDA:

- **Deskriptiv:** Welche Variablen tragen zur Gruppentrennung bei?
- **Klassifikatorisch:** Wie wird eine neue Beobachtung einer Gruppe zugeordnet?

Typische Anwendungsgebiete:

- Kreditwuerdigkeitspruefung (Gut-/Schlecht-Schuldner)
- Medizinische Diagnostik (krank/gesund)
- Marktsegmentierung mit bekannten Zielgruppen
- Qualitaetskontrolle (Gut-/Ausschuss-Teile)
- Artbestimmung in Biologie und Archaeologie

Mathematischer Hintergrund

Streumatrizen

Sei $\mathbf{X} \in \mathbb{R}^{n \times p}$ mit g bekannten Gruppen, n_i Beobachtungen je Gruppe und Gruppenmittelwert $\bar{\mathbf{x}}_i$.

Within-Gruppen-Streumatrix $\mathbf{S}_W$:

$$\mathbf{S}_W = \sum_{i=1}^{g} \sum_{j \in C_i} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^\top \tag{1}$$

Between-Gruppen-Streumatrix $\mathbf{S}_B$:

$$\mathbf{S}_B = \sum_{i=1}^{g} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^\top \tag{2}$$

Diskriminanzfunktionen

Die LDA maximiert das Rayleigh-Quotient-Kriterium:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}} \tag{3}$$

Dies fuehrt auf das verallgemeinerte Eigenwertproblem:

$$\mathbf{S}_B\,\mathbf{w} = \lambda\,\mathbf{S}_W\,\mathbf{w} \quad\Longleftrightarrow\quad \mathbf{S}_W^{-1}\mathbf{S}_B\,\mathbf{w} = \lambda\,\mathbf{w} \tag{4}$$

Die Loesungen $\mathbf{w}_1, \mathbf{w}_2, \ldots$ sind die **Diskriminanzkoeffizienten** der Diskriminanzfunktionen DF1, DF2, ...

Da $\mathbf{S}_W^{-1}\mathbf{S}_B$ nicht symmetrisch ist, loest das Tool das Problem ueber die Cholesky-Zerlegung $\mathbf{S}_W = \mathbf{L}\mathbf{L}^\top$: Die symmetrische Matrix $\mathbf{C} = \mathbf{L}^{-1}\mathbf{S}_B\mathbf{L}^{-\top}$ besitzt dieselben Eigenwerte λ_k; aus ihren Eigenvektoren $\mathbf{u}_k$ (Jacobi-Verfahren) folgt $\mathbf{w}_k = \sqrt{n-g}\,\mathbf{L}^{-\top}\mathbf{u}_k$. Durch diese Normierung haben die Diskriminanzscores innerhalb der Gruppen die (gepoolte) Varianz 1.

Die maximale Anzahl sinnvoller Diskriminanzfunktionen ist $\min(p,\; g-1)$.

Diskriminanzscores

$$d_{ik} = \mathbf{w}_k^\top \mathbf{x}_i = \sum_{j=1}^{p} w_{kj}\, x_{ij} \tag{5}$$

Klassifikationsregel: Beobachtung $\mathbf{x}_i$ wird der Gruppe g^* zugeordnet, deren Centroid $\bar{d}_g$ im Diskriminanzraum am naechsten liegt (wegen der Normierung entspricht der euklidische Abstand dem Mahalanobis-Abstand), korrigiert um die A-priori-Wahrscheinlichkeiten π_g:

$$g^* = \arg\min_{g} \Big[\sum_{k=1}^{K} (d_{ik} - \bar{d}_{gk})^2 - 2\ln \pi_g \Big] \tag{6}$$

Bei gleichverteilten Priors ($\pi_g = 1/g$) entfaellt der Korrekturterm; andernfalls gilt $\pi_g = n_g/n$.

Standardisierte Koeffizienten

$$w_{kj}^{\text{std}} = w_{kj} \cdot s_j^{(W)}, \qquad s_j^{(W)} = \sqrt{\frac{(\mathbf{S}_W)_{jj}}{n-g}} \tag{7}$$

Mit der gepoolten Innergruppen-Standardabweichung $s_j^{(W)}$ werden die Koeffizienten unabhaengig von der Messskala vergleichbar.

Kanonische Korrelation

$$r_k^* = \sqrt{\frac{\lambda_k}{1+\lambda_k}} \tag{8}$$

r_k^* ist die Korrelation zwischen den Diskriminanzscores DFk und den Gruppenindikatorvariablen – ein Mass fuer die Trennstaerke.

Wilks' Lambda

Wilks' Lambda Λ ist das zentrale Guetemass der LDA:

$$\Lambda = \prod_{k=1}^{K} \frac{1}{1+\lambda_k} = \frac{|\mathbf{S}_W|}{|\mathbf{S}_W + \mathbf{S}_B|} \tag{9}$$

$\Lambda = 0$: perfekte Trennung; $\Lambda = 1$: keine Trennung.

Bartlett-Test auf Signifikanz

$$\chi^2 \approx -\left(n - 1 - \frac{p+g}{2}\right) \ln \Lambda, \qquad \mathrm{df} = p(g-1) \tag{10}$$

Leave-One-Out-Kreuzvalidierung (LOO-CV)

Zur unverzerrten Schaetzung der Klassifikationsguete:

1. Beobachtung i aus dem Trainingsdatensatz entfernen
2. LDA auf den verbleibenden $n-1$ Beobachtungen schaetzen
3. Beobachtung i klassifizieren
4. Wiederholen fuer alle n Beobachtungen

Durchgerechnetes Beispiel: Bilanzkennzahlen

Datenbasis

Verwendet werden dieselben 24 synthetischen Unternehmen wie in Handbuch 1 (vollständige Datentabelle dort) mit sechs Kennzahlen: X_1 Eigenkapitalquote, X_2 Liquidität 2. Grades, X_3 Anlagendeckung II, X_4 Umsatzrendite, X_5 Gesamtkapitalrendite, X_6 Umsatzwachstum (alle in %). Anders als in den Handbüchern 1–3 wird hier die **bekannte Gruppenzugehörigkeit** genutzt: $g = 3$ Ratingklassen A, B, C mit je $n_g = 8$ Unternehmen (U01–U08: A, U09–U16: B, U17–U24: C). Berechnung mit z-standardisierten Variablen und gleichverteilten Priors (Werkseinstellungen); die Mittelwerte in Schritt 1 sind in Originaleinheiten angegeben, wie im Tool.

Ziel der Diskriminanzanalyse: Es soll ermittelt werden, welche Kennzahlen die Ratingklassen am besten trennen, und eine Regel gefunden werden, mit der sich neue Unternehmen einer Ratingklasse zuordnen lassen.

Schritt 1: Gruppenstatistiken

Tabelle 1: Gruppenmittelwerte (in %)

Gruppe	X_1 EK-Q.	X_2 Liqu.	X_3 AD II	X_4 UR	X_5 GKR	X_6 Wachst.
Rating A ($n = 8$)	45.14	127.63	125.13	5.26	7.26	5.20
Rating B ($n = 8$)	34.13	99.38	96.75	8.55	11.16	11.71
Rating C ($n = 8$)	28.00	84.38	88.88	1.88	4.60	−0.36
Gesamt ($n = 24$)	35.75	103.79	103.58	5.23	7.68	5.52

A-Unternehmen zeichnen sich durch hohe Stabilität aus, B-Unternehmen durch hohe Ertragskraft; C-Unternehmen liegen in allen Kennzahlen am niedrigsten.

Schritt 2: Diskriminanzfunktionen

Es gibt $\min(p,\, g-1) = \min(6,\, 2) = 2$ Diskriminanzfunktionen.

Tabelle 2: Eigenlösungen und erklärte Varianz

Funktion	Eigenwert λ	r_k^*	Varianz	Kumuliert
DF1	8.003	0.943	71.8 %	71.8 %
DF2	3.146	0.871	28.2 %	100.0 %

Beispiel: $r_1^* = \sqrt{8.003/9.003} = 0.943$; Varianzanteil DF1 $= 8.003/(8.003 + 3.146) = 71.8\,\%$.

Wilks' Lambda:

$$\Lambda = \frac{1}{1+8.003} \cdot \frac{1}{1+3.146} = 0.1111 \cdot 0.2412 = 0.0268 \quad \text{(sehr gute Trennung)},$$

Bartlett-Test: $\chi^2 = -\big(24 - 1 - \tfrac{6+3}{2}\big) \ln 0.0268 = 18.5 \cdot 3.620 = 66.96$, df $= 6 \cdot 2 = 12$, $p < 0.001$.

Schritt 3: Standardisierte Diskriminanzkoeffizienten

Tabelle 3: Standardisierte Koeffizienten

Variable	DF1	DF2
X_1 EK-Quote	−0.038	0.193
X_2 Liquidität	0.160	0.342
X_3 Anlagendeckung	0.069	**0.726**
X_4 Umsatzrendite	**0.748**	−0.184
X_5 GK-Rendite	**0.488**	−0.281
X_6 Umsatzwachstum	0.339	0.149

Interpretation:

- **DF1** (71.8 % der Trennung) ist eine **Ertragsfunktion**: Umsatzrendite, Gesamtkapitalrendite und Wachstum haben die größten Gewichte.
- **DF2** (28.2 %) ist eine **Stabilitätsfunktion**: Anlagendeckung und Liquidität dominieren.
- Die Eigenkapitalquote hat in beiden Funktionen nur kleine Gewichte, obwohl sie zwischen den Gruppen deutlich variiert. Grund: Sie ist eng mit Liquidität und Anlagendeckung korreliert; standardisierte Koeffizienten messen den Beitrag *zusätzlich* zu den übrigen Variablen.

Die Diskriminanzanalyse findet damit dieselben zwei Dimensionen wie PCA und Faktoranalyse – hier jedoch so gewichtet, dass die Ratingklassen maximal getrennt werden. Die Vorzeichen von Diskriminanzfunktionen sind willkürlich; eine Spiegelung ändert die Ergebnisse nicht.

Schritt 4: Gruppencentroide im Diskriminanzraum

Tabelle 4: Gruppencentroide

Gruppe	DF1-Centroid	DF2-Centroid
Rating A	0.22	2.34
Rating B	3.13	−1.29
Rating C	−3.35	−1.05

DF1 trennt die ertragsstarken B-Unternehmen von den C-Unternehmen (Abstand 6.5), DF2 hebt die A-Unternehmen mit ihrer hohen Stabilität von beiden anderen Gruppen ab. Im Gruppenplot (DF1 vs. DF2) bilden die drei Klassen deutlich getrennte Wolken.

Schritt 5: Klassifikation und Kreuzvalidierung

Tabelle 5: Konfusionsmatrix (Trainingsdaten)

	Vorhergesagt		
Ist	A	B	C
Rating A	**8**	0	0
Rating B	0	**8**	0
Rating C	0	0	**8**

Trefferquote (Resubstitution): 100 % (24 von 24 korrekt).

Diese Quote ist zu optimistisch, weil jedes Unternehmen mit einer Funktion klassifiziert wird, an deren Schätzung es selbst beteiligt war. Die **Leave-One-Out-Kreuzvalidierung** (im Tool per Checkbox aktivieren) ergibt **91.7 %** (22 von 24): Zwei A-Unternehmen werden falsch zugeordnet, wenn sie nicht zur Schätzung beitragen:

- **U03** → Rating C: unterdurchschnittliche Stabilität und Rendite (vgl. Clusteranalyse, Handbuch 3).

- **U07** → Rating B: sehr hohe Ertragskennzahlen (Umsatzrendite 8.8 %, GK-Rendite 9.5 %, Wachstum 10.1 %) – ein Ertragsprofil wie bei einem B-Unternehmen.

Die Differenz von 8.3 Prozentpunkten zwischen Resubstitution und LOO-CV ist bei $n = 24$ und $p = 6$ typisch und zeigt, warum die Kreuzvalidierung für kleine Stichproben empfohlen wird.

Handhabung des interaktiven LDA-Tools

Dateneingabe

Besonderheit des LDA-Tools: Die **GRP-Spalte** ganz links in der manuellen Eingabe enthaelt die Gruppenbezeichnungen (Text oder Zahlen, z. B. »Setosa«, »Gut«, »1«). Jede Zeile muss eine Gruppenbezeichnung enthalten, damit sie in die Analyse einfliesst.

Vorgeladen: Bilanzkennzahlen (3 Gruppen Rating A/B/C, $n = 24$, $p = 6$). Weitere Beispiele: Iris (3 Gruppen, $n = 15$, $p = 4$), Kreditwürdigkeit (2 Gruppen, $n = 21$, $p = 4$) und Marktsegmente (3 Gruppen, $n = 24$, $p = 4$). Zu jedem Beispiel erscheint über der Eingabetabelle ein Erläuterungskasten mit den Daten, dem Ziel der Analyse und dem erwarteten Ergebnis; er verschwindet bei »Leer« oder beim Laden einer CSV-Datei.

Parameterauswahl

Standardisierung
: Z-Transformation empfohlen. Bei gleichen Einheiten kann »Keine« gewaehlt werden – die unstandardisierten Koeffizienten sind dann direkt interpretierbar.

Anzahl Diskriminanzfunktionen
: *Automatisch:* $\min(p, g-1)$ Funktionen. Manuell: 1 bis 3 Funktionen (hoechstens $\min(p, g-1)$); Klassifikation und Grafiken verwenden dann nur diese Funktionen. Wilks' Λ wird stets ueber alle Funktionen berechnet.

Gruppenpriors
: Checkbox aktiviert: gleichverteilte Priors (alle Gruppen gleich wahrscheinlich). Standard und empfohlen, wenn keine A-priori-Wahrscheinlichkeiten bekannt sind. Checkbox deaktiviert: Priors proportional zu den Gruppengroessen (n_g/n).

LOO-Kreuzvalidierung
: Checkbox: Berechnet Leave-One-Out-Trefferquote (rechenintensiver). Empfohlen bei kleinen Stichproben ($n < 50$).

Ergebnistabs

Uebersicht
: Kennzahlen (n, Gruppen, Diskriminanzfunktionen, Trefferquote) + Wilks-Lambda-Tabelle mit Signifikanzbewertung + DF-Eigenwert-Tabelle mit kanonischen Korrelationen + Tabelle der Gruppenmittelwerte (in Originaleinheiten).

Koeffizienten
: Standardisierte und unstandardisierte Diskriminanzkoeffizienten je Funktion. Gruppencentroide im Diskriminanzraum.

Klassifikation
Konfusionsmatrix mit Farbcodierung (Diagonale gruen, Fehler rot) + Gesamttrefferquote + Einzelklassifikationstabelle fuer jede Beobachtung.

Grafiken
Gruppenplot (DF1 vs. DF2): Streudiagramm aller Beobachtungen und Gruppencentroide (Rauten) im Diskriminanzraum. *Varianzbalken:* Erklaerte Varianz je Diskriminanzfunktion. *Koeffizientenplot:* Balkendiagramm der standardisierten Absolutkoeffizienten fuer DF1.

Interpretation
Automatischer Bewertungstext: Wilks-Lambda, Bartlett-Test, wichtigste Diskriminatoren, LOO-CV-Vergleich.

Export
CSV: Koeffizienten, Klassifikation, Diskriminanzscores. PNG: Gruppenplot, Koeffizientenbalken.

Typischer Arbeitsablauf

1. Daten mit GRP-Spalte laden (jede Gruppe ≥ 2 Beobachtungen)
2. Standardisierung auf Z-Transformation belassen
3. LOO-CV-Checkbox aktivieren (bei $n < 100$)
4. »**LDA berechnen**« klicken
5. Uebersicht-Tab: Wilks' Λ und Bartlett-p-Wert pruefen
6. Koeffizienten-Tab: Welche Variablen trennen am staerksten?
7. Klassifikation-Tab: Konfusionsmatrix analysieren – welche Gruppen werden verwechselt?
8. Grafiken-Tab: Gruppenplot – sind die Gruppen im DF1-DF2-Raum trennbar?
9. LOO-CV vs. Training: Grosse Differenz deutet auf Ueberanpassung

Wichtige Hinweise:
- Jede Gruppe muss mindestens 2 Beobachtungen enthalten.
- Faustformel: $n \geq 5 \cdot p$ je Gruppe fuer stabile LDA.
- Voraussetzungen: Multinormalverteilung und Homogenitaet der Kovarianzmatrizen (Homoskedastizitaet) je Gruppe.
- Bei verletzter Homoskedastizitaet: Quadratische Diskriminanzanalyse (QDA).
- Die im Tool ausgewiesene Trefferquote ist eine *Resubstitutions*-Schaetzung – sie ueberschaetzt die wahre Trefferquote; LOO-CV liefert einen unverzerrteren Schaetzer.
- Singulaere $\mathbf{S}_W$ bei $n < p$ oder perfekter Multikollinearitaet: mehr Beobachtungen oder weniger Variablen verwenden.

Formelübersicht

Tabelle 6: Formeln der Diskriminanzanalyse

Bezeichnung	Formel
Within-Streumatrix	$\mathbf{S}_W = \sum_i \sum_{j \in C_i} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^\top$
Between-Streumatrix	$\mathbf{S}_B = \sum_i n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^\top$
Eigenwertproblem	$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$
Std. Koeffizient	$w_{kj} \sqrt{(\mathbf{S}_W)_{jj}/(n-g)}$
Diskriminanzscore	$d_{ik} = \mathbf{w}_k^\top \mathbf{x}_i$
Kan. Korrelation	$r_k^* = \sqrt{\lambda_k/(1+\lambda_k)}$
Wilks' Lambda	$\Lambda = \prod_k 1/(1+\lambda_k)$
Bartlett χ^2	$-(n-1-(p+g)/2)\,\ln\Lambda$, $\mathrm{df} = p(g-1)$
Klassifikation	$g^* = \arg\min_g \left[\sum_k (d_{ik} - \bar{d}_{gk})^2 - 2\ln \pi_g \right]$

Weiterführende Literatur

- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
- Backhaus, K. et al. (2021). *Multivariate Analysemethoden*, 16. Aufl. Springer Gabler.
- McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley.