7.3 Lineare Entmischung

a Für die Hauptkomponenten-Analyse haben wir eine Modell-Gleichung aufgeschrieben (7.1.i),

$$\boldsymbol{x}_c = \boldsymbol{S}\,\boldsymbol{C}^T + \boldsymbol{r}\ ,$$

deren Struktur sich für viele Anwendungen eignet, in denen eine Anzahl p von Komponenten sich in Bezug auf $m > p$ Messgrössen **linear überlagern.**

Beispielsweise liefern p chemische Substanzen (optische oder chromatographische) **Spektren** $\underline{c}_k = [c_k^{(1)}, ..., c_k^{(m)}]$, die sie charakterisieren. Die Spektren messen Intensitäten $c^{(j)}$ oder Mengen für m Wellenlängen oder Laufzeiten. Eine grafische Darstellung von 121 Near Infrared (NIR-) Spektren, die im Verlauf einer chemischen Reaktion gemessen wurden, ist in der Einleitung (1.2.g) zu finden. Bei einer Mischung ohne chemische Reaktion überlagern sich die Spektren meistens, mindestens genähert, linear (nach dem Gesetz von Lambert und Beer). Wenn eine chemische Reaktion erfolgt, mischt sich noch das Spektrum der neu entstehenden Substanz dazu.

Das Spektrum der $i$ten Mischung ist also bis auf Messfehler gleich $S_i^{(1)}\underline{c}_1 + S_i^{(2)}\underline{c}_2 + ... + S_i^{(p)}\underline{c}_p$, wobei die $S_i^{(k)}$ die Anteile der Substanz k am $i$ten Gemisch bedeuten. Die gemessenen Spektren $\boldsymbol{X}$ folgen also dem genannten Modell mit $\underline{\mu} = \underline{0}$. Die Spektren $\underline{c}_k$ der Substanzen bilden die Spalten der Matrix $\boldsymbol{C}$.

Da Spektren und Anteile nicht negativ sein können, muss $S_i^{(k)} \geq 0$ und $C_j^{(k)} \geq 0$ für alle i, j und k gelten. Das Modell mit diesen Nebenbedingungen nenen wir das **Modell der linearen Mischung**.

b In gleicher Weise überlagern sich

- Luftfremdstoffe, die aus p Quellen mit immer gleichem „Quellenprofil“ $\underline{c}_k$ stammen. Ihr Beitrag $S_i^{(k)}$ zu einer Schadstoffmessung $\underline{X}_i$ ist abhängig von ihrer Aktivität und dem Transport von der Quelle zum Messort, wie sie zum $i$ten Messzeitpunkt wirksam werden.
- chemische Elemente in Felsen, die aus mehreren Grundgesteinen zusammengesetzt sind,
- Spurenelemente in Quellwasser, das verschiedene Gesteinsschichten durchlaufen hat.

c **Modell.** Wir schreiben die Gleichung nochmals, aber nun für unzentrierte Zufallsvariable statt für zentrierte Daten:

$$\boldsymbol{X} = \boldsymbol{S}\,\boldsymbol{C}^T + \boldsymbol{E}\ .$$

Dass wir einen Fehlerterm $\boldsymbol{E}$ statt eines „Restterms“ $\boldsymbol{r}$ schreiben, deutet darauf hin, dass wir hier wie üblich normalverteilte Abweichungen annehmen werden, also $\underline{E}_i \sim \mathcal{N}\langle \underline{0}, \boldsymbol{\Sigma} \rangle$, unabhängig, für die Abweichungen $\underline{E}_i$ der $i$ten Beobachtung, wie in der multivariaten Regression.

d Wenn die Quellenprofile (Spektren) $\underline{c}_k$ alle bekannt sind, können die Beiträge $S_i^{(k)}$ für jede Beobachtung i separat mit Hilfe einer multiplen Regressionsschätzung bestimmt werden. Es gilt ja $\underline{X}_i = \boldsymbol{C}\underline{s}_i + \underline{E}_i$, und das ist die Matrixform eines multiplen Regressionsmodells mit der „X-Matrix“ $\boldsymbol{C}$ und den Koeffizienten $s_i^{(k)}$, $k = 1, 2, ..., p$. Allerdings haben die Abweichungen $E_i^{(j)}$ kaum die gleichen Varianzen. Für eine gute Bestimmung braucht es Annahmen über diese Varianzen und dann gewichtete Regression (* oder eine nicht-lineare für eine transformierte Version der Zielgrösse, z.B. $\log\left\langle X_i^{(j)} \right\rangle = \log\langle \boldsymbol{C}\underline{s}_i \rangle + \widetilde{E}_i^{(j)}$).

e **Lineare Entmischung.** Interessanter ist der Fall, in dem aus einem Datensatz sowohl die **Quellenprofile** $\underline{c}_k$ als auch die Beiträge $\underline{S}_i$ **geschätzt werden** müssen. Wir sprechen von **linearer Entmischung**. Dies lässt sich mit einer Kombination von statistischen Methoden, anwendungsspezifischen Besonderheiten und Fachwissen oft gut erreichen.

Die Hauptschwierigkeit liegt dabei in der Tatsache, dass $\boldsymbol{S}$ und $\boldsymbol{C}$ im Modell nicht eindeutig – also **nicht identifizierbar** – sind: Für jede invertierbare Matrix $\boldsymbol{T}$ können $\boldsymbol{S}$ und $\boldsymbol{C}$ durch $\widetilde{\boldsymbol{S}} = \boldsymbol{S}\boldsymbol{T}$ und $\widetilde{\boldsymbol{C}} = \boldsymbol{C}(\boldsymbol{T}^T)^{-1}$ ersetzt werden, ohne dass sich bei gleichen Fehlern $\boldsymbol{E}$ die Daten $\boldsymbol{X}$ ändern.

f **Schätzung des Unterraums.** Mit statistischen Mitteln kann man deshalb zunächst nur die „fehlerkorrigierten Beobachtungen“ $\widetilde{\boldsymbol{X}} = \boldsymbol{S}\boldsymbol{C}^T = \widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{C}}^T$ schätzen.

Falls unabhängige $E_i^{(j)} \sim \mathcal{N}\langle 0, \sigma^2 \rangle$ mit gleichen Varianzen angenommen werden, liefert eine Version der **Hauptkomponenten-Analyse**, bei der die Daten nicht zentriert werden, die beste Schätzung von $\widetilde{\boldsymbol{X}}$. (S: `prcomp(..., center=FALSE)`.) Wenn man die Verhältnisse zwischen den Varianzen $\sigma_j^2 = \mathrm{var}\left\langle E_i^{(j)} \right\rangle$ zu kennen glaubt, kann man die Variablen $X^{(j)}$ je durch σ_j dividieren und dann die nicht-zentrierte Hauptkomponenten-Analyse verwenden. Wenn man über die Varianzen nichts weiss, liefert die Faktor-Analyse eine Schätzung (siehe 7.4.b).

g Die fehlerkorrigierten Beobachtungen liegen in einem Raum mit Dimension p. In diesem **Unterraum** sind die „**Achsen**“ also zunächst **unbestimmt**. In den Anwendungen führen verschiedene Überlegungen zu einer sinnvollen Festlegung der Achsen. Eine Unbestimmtheit der harmlosen Art besteht darin, dass man jedes Quellenprofil, also jede Spalte von $\boldsymbol{C}$, mit einem Faktor multiplizieren und die entsprechenden Spalten von $\boldsymbol{S}$ durch die gleiche Zahl dividieren kann, und man erhält wieder die gleichen fehlerkorrigierten Beobachtungen. Die Quellenprofile müssen also in diesem Sinne auf geeignete Art standardisiert werden, damit das Modell eindeutig wird.

h ▷ **Im Beispiel der NIR-Spektren** kann man in der Darstellung der ersten vier Hauptkomponenten (7.1.j) vier Phasen der Reaktion erkennen, in denen der Prozess sich jeweils in eine Richtung bewegt. Man kann die entsprechenden vier Richtungsvektoren als neue Achsen benützen und die entsprechenden Koordinaten oder Scores ausrechnen. Wenn man sie gegen die zeitliche Reihenfolge der Beobachtungen aufträgt, erhält man die Darstellung 1.2.g (ii), die in der Einleitung gezeigt wurde. ◁

i Will man in anderen Beispielen der linearen Entmischung geeignete Achsen finden, dann helfen die Ungleichungen $S_i^{(k)} \geq 0$ und $c_j^{(k)} \geq 0$ und vorhandenes Fachwissen über die Quellenprofile (Spektren der chemischen Substanzen) $\underline{c}_k$, also die Spalten der Matrix $\boldsymbol{C}$. Beides wollen wir noch kurz ausführen.

j **Nicht-Negativität.** Die Nicht-Negativität der $S_i^{(k)}$ und $c_j^{(k)}$ schränkt die Wahlfreiheit schon erheblich ein. Das Modell der linearen Mischung sagt ja, dass jeder Beobachtungsvektor $\underline{X}_i$ eine Linearkombination der Quellenprofile $\underline{c}_k$ ist, und zwar eine mit nicht-negativen Koeffizienten, eine so genannte konvexe Linearkombination.

Veranschaulichen wir uns diese Überlegung im dreidimensionalen Raum, also bei $m = 3$ Variablen und $p = 2$ Quellen (Abbildung 7.3.j). Die beiden Quellenprofile $\underline{c}_1$ und $\underline{c}_2$ liegen in einer Ebene. Die möglichen Mischungen $\underline{X}_i$ liegen, bis auf die Abweichungen $\underline{E}_i$, ebenfalls in dieser Ebene, aber nicht nur das, sie liegen auch im Sektor zwischen den beiden Quellenprofilen.

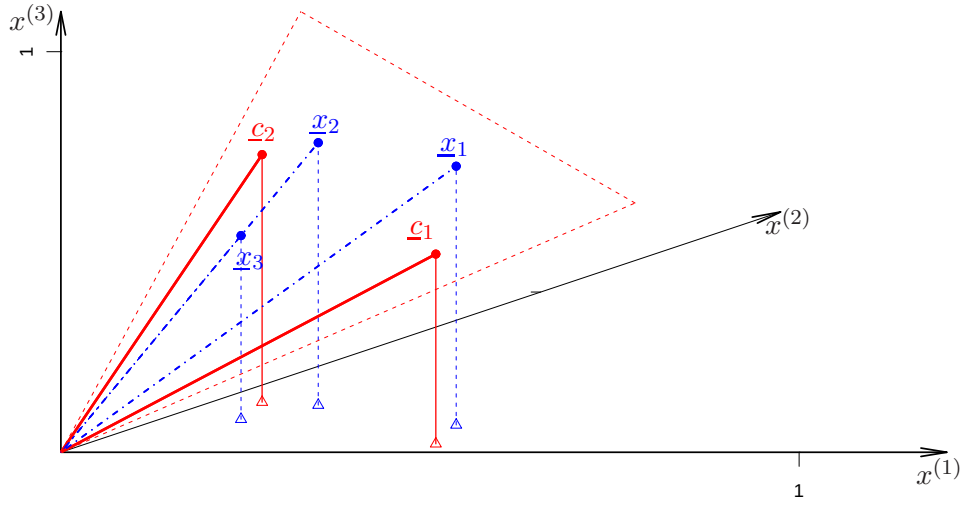

Abbildung 7.3.j: Veranschaulichung des Modells der linearen Mischung von 2 Quellenprofilen $\underline{c}_1$ und $\underline{c}_2$

Wenn wir nun annehmen, dass bei gewissen Beobachtungen nur jeweils eine einzige Quelle aktiv ist, dann sind diese Beobachtungsvektoren als extremste Vektoren unter allen beobachteten zu erkennen. Bei zwei Quellen bilden sie die Begrenzungen des Sektors, bei drei Quellen die Kanten einer dreieckigen Pyramide mit Spitze im Nullpunkt, usw.

Diese Eigenschaft kann man brauchen, um Quellenprofile grafisch zu bestimmen oder um mit einem numerischen Algorithmus die Kanten einer minimalen, die Beobachtungen umhüllenden (Hyper-) Pyramide zu bestimmen.

k **Compositional Data.** In gewissen Anwendungen bestehen die Daten aus Anteilen, beispielsweise von Gesteinsarten in Felsbrocken. Die Variablen ergänzen sich dann auf 100% oder eine andere feste Zahl.

Wenn dies von den Daten her nicht gegeben ist, kann es sinnvoll sein, auf diese Weise zu standardisieren, also jeden Eintrag $X_i^{(j)}$ durch die Zeilensumme $\sum_j X_i^{(j)}$ zu dividieren (und mal 100% zu rechnen). Die Quellenprofile müssen ja, wie oben erwähnt, auf eine Weise standardisiert werden, und das wird man dann auch so tun.

Solche Daten liegen bereits wegen der Nebenbedingung in einem Unterraum (mathematisch „Nebenraum“). Abbildung 7.3.k zeigt, wie das im Fall von 2 Quellen im dreidimensionalen Raum aussieht.

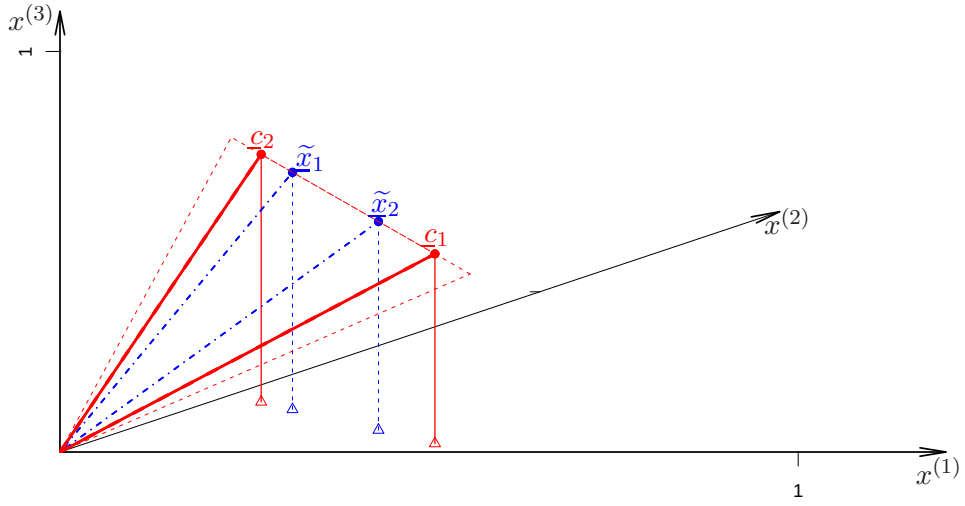

Abbildung 7.3.k: Veranschaulichung des Modells für standardisierte Daten

Durch Standardisierung wird der Raum, in dem man die geeigneten Achsen suchen muss, um eine Dimension kleiner, und man findet ihn mit gewöhnlicher Hauptkomponenten-Analyse, die mit zentrierten Daten arbeitet.

l **Chemische Reaktionen.** Bei chemischen Reaktionen ist für den Start-Zeitpunkt das Mischverhältnis der Ausgangs-Substanzen und deren Spektrum bekannt. Durch die Reaktion entsteht eine oder mehrere neue Substanzen, wobei gleichzeitig die Substanzen des Ausgangsgemisches in einem klaren, einfachen (stöchiometrischen) Verhältnis verschwinden. Das hilft, die unbekannten Spektren der entstehenden Substanzen diese zu finden.

L **Literatur:** Weitere Ausführung zur Methodik findet man in der Literatur unter den Namen *linear mixing model* und *mass balance*; als Startpunkt eignet sich Renner (1993). Grundlagen finden sich auch in den Büchern über Chemometrie.

7.4 Faktor-Analyse

a Das Modell 7.1.i ist in der Psychologie unter dem Namen **Faktor-Analyse** schon lange bekannt. In der typischen Anwendung ist $X_i^{(j)}$ die Punktezahl, die Proband i bei der $j$ten Test-Aufgabe erzielt. Sie wird aufgefasst als Ergebnis einer Überlagerung $S_i^{(1)}C_j^{(1)} + S_i^{(2)}C_j^{(2)} \ \big(+ \ldots + S_i^{(p)}C_j^{(p)}\big)$ von Faktoren, die oft als seine mathematische Intelligenz $S_i^{(1)}$ und seine sprachliche Intelligenz $S_i^{(2)}$ (und eventuell weiterer „Dimensionen“ der Intelligenz) interpretiert werden, bis auf eine zufällige Abweichung $E_i^{(j)}$. Die **„Faktoren“** $S^{(k)}$ sind allerdings nicht beobachtbar; sie können nur über die beobachteten Grössen $X_i^{(j)}$ erschlossen werden. Solche Zufallsvariable werden **latente Variable** genannt.

b **Modell.** Das Modell hat, wie sich aus dieser Überlegung ergibt, wieder die Form $\boldsymbol{X} = \boldsymbol{S}\,\boldsymbol{C}^T + \boldsymbol{E}$, wobei üblicherweise die Daten $\boldsymbol{X}$ als zentriert angenommen werden, im Gegensatz zum Modell der linearen Mischung. Von der Hauptkomponenten-Analyse unterscheidet sich die Faktor-Analyse in zweierlei Hinsicht:

- Für die Abweichungen $\underline{E}_i$ wird eine Normalverteilung vorausgesetzt, $\underline{E}_i \sim \mathcal{N}\langle \underline{0}, \boldsymbol{\Sigma} \rangle$. Speziell wird angenommen, dass die Abweichungen für die *Variablen* j, also die $E^{(j)}$, voneinander unabhängig sind (nicht nur die Beobachtungen). Die Kovarianzmatrix $\boldsymbol{\Sigma}$ wird also als diagonal vorausgesetzt, $\boldsymbol{\Sigma} = \mathrm{diag}\langle \sigma_1^2, \sigma_2^2, ..., \sigma_m^2 \rangle$. Das bedeutet, dass die Abhängigkeiten zwischen den Variablen $X^{(j)}$ vollständig durch den Term $\boldsymbol{S}\,\boldsymbol{C}^T$ „erklärt“ werden. Die Faktoren k mit den „Scores“ $\underline{S}^{(k)}$ werden werden deshalb genauer mit „common factors“ bezeichnet, während die unabhängigen $E^{(j)}$ „unique factors“ und deren Varianzen σ_j^2 „Uniquenesses“ heissen.
- Die Scores $S_i^{(k)}$ werden ebenfalls als Zufallsvariable aufgefasst und auch mit einer Normalverteilung $\mathcal{N}\langle \underline{0}, \boldsymbol{V} \rangle$ modelliert. Dadurch werden die $\underline{X}_i$ zu normalverteilten Beobachtungs-Vektoren mit

$$\underline{X}_i \sim \mathcal{N}\big\langle \underline{0},\ \boldsymbol{C}\,\boldsymbol{V}\,\boldsymbol{C}^T + \mathrm{diag}\langle \sigma_1^2, ..., \sigma_m^2 \rangle \big\rangle \ .$$

Damit ist ein Wahrscheinlichkeitsmodell entstanden, für das man die Parameter σ_j^2 und die Elemente von $\boldsymbol{C}\,\boldsymbol{V}\,\boldsymbol{C}^T$ schätzen kann. Auch hier bleibt aber die Aufspaltung von $\boldsymbol{C}\,\boldsymbol{V}\,\boldsymbol{C}^T$ in die Matrix $\boldsymbol{C}$ und die Kovarianzmatrix $\boldsymbol{V}$ der Scores nicht-eindeutig.

c **Anzahl Faktoren.** Im Gegensatz zur Hauptkomponenten-Analyse, bei der die Anzahl Faktoren, die zur Interpretation oder Darstellung der Daten benützt werden, willkürlich festgelegt wird, ist im Modell der Faktor-Analyse diese Zahl durch die eben erwähnte Forderung gegeben, dass die übrig bleibenden „unique factors“ $E^{(j)}$ unabhängig sind.

Nach dem allgemeinen Prinzip des Likelihood-Quotienten-Tests kann man überprüfen, ob p Faktoren genügen. Das ist der Fall, wenn die Likelihood für das Faktor-Modell nur wenig kleiner ist als die Likelihood für eine allgemeine multivariate Normalverteilung. Dieser Test ist in gängigen Programmen zur Anpassung der Faktoranalyse enthalten.

In der Anwendung kann es aber auch nützlich sein, ein Modell zu verwenden, das in diesem Sinne signifikant zu wenige Faktoren hat, vor allem, wenn die Anzahl Beobachtungen so gross ist, dass auch noch unbedeutende verbleibende Korrelationen der $E^{(j)}$ zur Signifikanz eines weiteren Faktors führen.

d **Hauptfaktoren.** Wie erwähnt, besteht auch in der Faktoranalyse das Problem, dass die Faktoren selbst, also $\boldsymbol{C}$, eigentlich nicht bestimmbar sind (vergleiche 7.3.e). Um sie eindeutig zu machen, sind verschiedene Kriterien vorgeschlagen worden.

Eine eher theoretisch motivierte Art, Eindeutigkeit zu erzielen, besteht darin, dass man als Faktoren die Hauptkomponenten der „Beobachtungen im Unterraum" $\underline{\widetilde{X}}_i = \underline{X}_i - \underline{E}_i$ verwendet. Dadurch erreicht man, dass die Scores $\underline{S}_i$ unabhängig sind und die Matrix $\boldsymbol{C}$ „pseudo-orthogonal" ist, $\boldsymbol{C}^T\boldsymbol{C} = \boldsymbol{I}$. Diese Variante der Faktor-Analyse heisst **Hauptfaktoren-Analyse** oder *principal factor analysis.*

e **Interpretierbarkeit der Faktoren.** Diese Lösung dient als Ausgangspunkt für die Bestimmung weiterer möglicher Festlegungen der Faktoren. Ziel einer solchen Festlegung ist, dass die Faktoren gut interpretierbar sein sollen. Dazu kann folgendes nützlich sein:

- Die Scores sollen standardisiert sein. Mittelwert null haben sie schon, weil die Daten vor der Faktor-Analyse jeweils zentriert werden. Varianz 1 führt dazu, dass beispielsweise ein Score von 1 einen um eine Standardabweichung höheren Wert als der Durchschnitt bedeutet, und man gemäss Normalverteilung erwarten kann, dass nur 1/6 der Grundgesamtheit noch höher abschneidet als das. – Diese Forderung ist einfach zu erfüllen, indem man irgendwie definierte Faktoren am Ende noch standardisiert.
- Die Scores sollen unabhängig sein, damit jeder Faktor einen von den anderen Faktoren unabhängigen Aspekt der Daten misst. Allerdings ist es wohl falsch, nach unabhängigen Faktoren zu suchen, die die mathematische und die sprachliche Intelligenz messen.
- Grundsätzlich ist ein Faktor dann gut interpretierbar, wenn er nur mit wenigen ursprünglichen Variablen $X^{(j)}$ eine hohe Korrelation hat, und wenn diese Variablen mit den anderen Faktoren eine kleine Korrelation haben.

Diese Korrelationen werden „**Ladungen**" *(loadings)* genannt,

$$\lambda_j^{(k)} = \operatorname{corr}\Big\langle X^{(j)}, S^{(k)} \Big\rangle \ .$$

Es gilt $\operatorname{cov}\langle \underline{X}, \underline{S} \rangle = \operatorname{cov}\langle \boldsymbol{C}\underline{S} + \underline{E}, \underline{S} \rangle = \boldsymbol{C} \operatorname{var}\langle \underline{S}, \underline{S} \rangle$. Wenn die ursprünglichen Variablen univariat standardisiert sind und die Faktoren multivariat standardisiert, gibt deshalb die Matrix $\boldsymbol{C}$ die Ladungen wieder. Achtung! Für die Hauptfaktoren-Analyse (ohne Standardisierung der Faktoren) gilt das nicht!

f **Rotationen.** Die Standardisierung der Scores der Hauptfaktoren-Analyse führt dazu, dass ihre Kovarianzmatrix die Einheitsmatrix ist, denn unkorreliert waren sie schon. Die standardisierten Hauptfaktoren sind also auch multivariat standardisiert. Nun kann man aber anschliessend diese standardisierten Faktoren mit jeder orthogonalen Matrix transformieren, und man behält damit multivariat standardisierte Scores. Die verschiedenen Vorschläge für eine solche Transformation mit dem Ziel, die Interpretierbarkeit

zu verbessern, werden deshalb **orthogonale Rotationen** genannt. Das wäre eigentlich ein Pleonasmus, wenn nicht diejenigen (linearen) Transformationen, die die Orthogonalität nicht einhalten, traditionellerweise „**schiefe Rotationen**"(!) *(oblique rotations)* genannt würden.

- Die Forderung, dass nur wenige Ladungen $\lambda_j^{(k)}$ gross sein sollen, kann man durch die Varianz der $(\lambda_j^{(k)})^2$ quantifizieren. Dieses Kriterium wird **varimax** genannt und üblicherweise zur Optimierung verwendet.
- Wenn „schiefe Rotationen" zugelassen sind, wird meist das Kriterium **oblimin** optimiert, siehe Literatur.

g ▷ **I**m **Beispiel der Abstimmungen** kann man zunächst die Anzahl Faktoren mit dem Likelihood-Quotienten-Test bestimmen. Die Antwort ist in diesem Fall klar: Der P-Wert wird für 2 Faktoren 0, für 3 Faktoren aber 0.54. Also werden drei Faktoren gebraucht.

Tabelle 7.4.g zeigt die Loadings für die varimax-Rotation. Hohe Ladungen erhält der erste Faktor für die Abstimmungen i, c und b, der zweite für f und e, der dritte für k, l und m. Die „Uniquenesses" sind für diese Variablen tendenziell kleiner als für die übrigen. Besonders hohe uniqueness zeigen die Abstimmungen n, h und g.

	Factor1	Factor2	Factor3	Uniqueness
a	0.518	0.576	-0.478	0.463
b	0.786	0.283	0.348	0.400
c	0.852	0.309	0.172	0.418
d	0.536	0.510	0.175	0.652
e	0.195	0.871	-0.216	0.709
f	0.346	0.922		0.626
g		0.635	-0.401	0.800
h	-0.751	0.555	-0.189	0.864
i	0.964			0.631
j	0.583		0.440	0.689
k	0.129		0.973	0.251
l	0.252	-0.123	0.957	0.071
m	0.252	-0.140	0.945	0.159
n	-0.131	0.721		0.963

Tabelle 7.4.g: Loadings und Uniquenesses im Beispiel der Abstimmungen mit 3 Faktoren und varimax-Rotation

Die Ergebnisse sind in Abbildung 7.4.g grafisch einsehbar. Der Gegensatz zwischen französischer und deutscher Schweiz zeigt sich weniger gut als in der Hauptkomponenten-Darstellung und der Stadt-Land-Unterschied noch schlechter. In diesem Beispiel bewährt sich die Faktoranalyse also schlecht. Das ist nicht erstaunlich, denn eine Faktor-Struktur ist für die Abstimmungs-Vorlagen nicht besonders plausibel. ◁

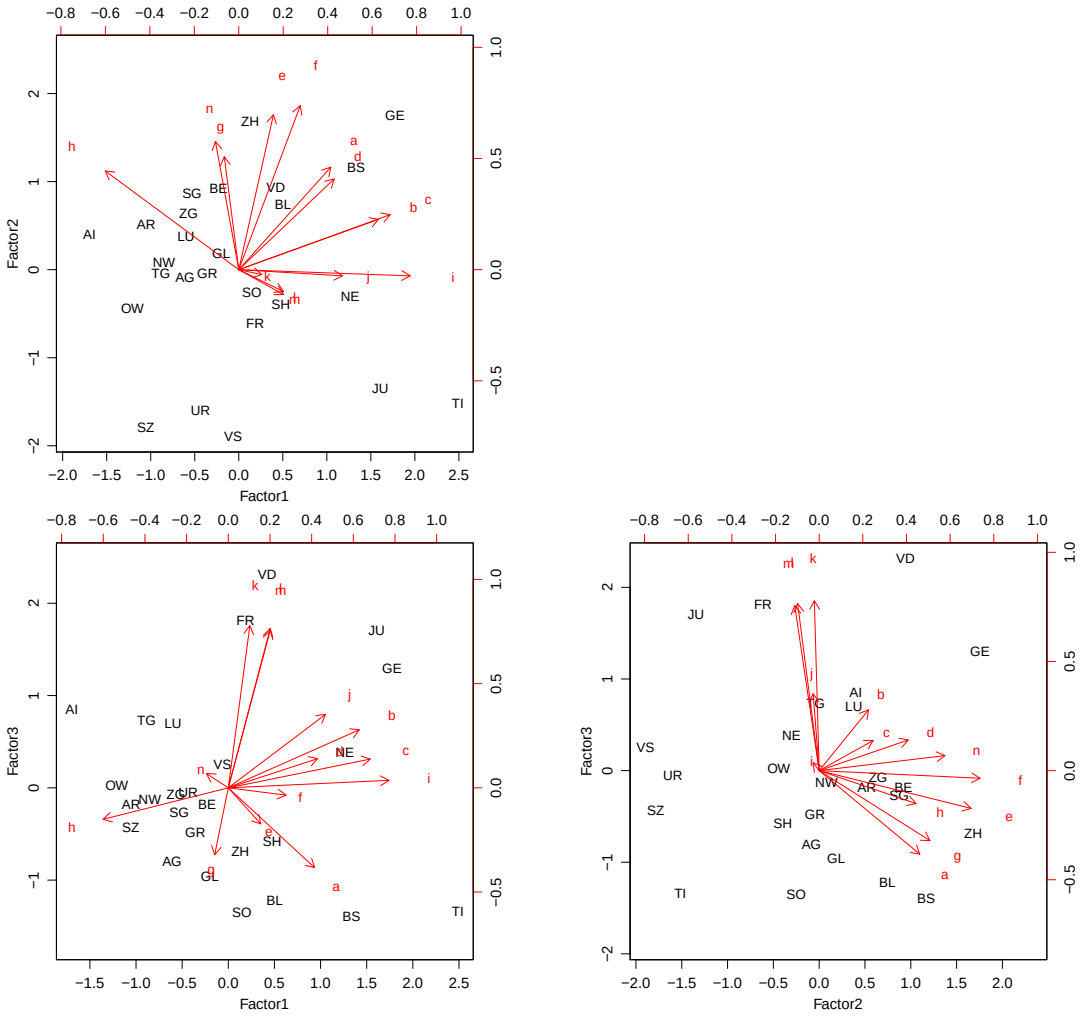

Abbildung 7.4.g: Biplot der Faktor-Analyse im Beispiel der Abstimmungen

h **Bedeutung der Faktor-Analyse.** Die Faktor-Analyse soll dann angewandt werden, wenn die Variablen so gewählt werden, dass sie bestimmte vermutete Faktoren – latente Variable, die man nicht direkt messen kann – wiedergeben. Wenn klare theoretische Vorstellungen über die Beziehung zwischen beobachtbaren Variablen und theoretisch vermuteten, nicht beobachtbaren vorliegen, dann kann man diese Zuordnungen auch direkt in ein Modell hineinstecken. Das führt zu den so genannten **Strukturgleichungs-Modellen**.

L **Literatur:** Eine Einführung in diese Methodik enthält Kapitel 14 von Bortz (1977). Bekannte Bücher sind Harman (1960, 1967) und Lawley and Maxwell (1971). Alles sehr alte Bücher! Ich suche noch nach neuerer Literatur.