Modifikationen der Hauptkomponentenanalyse

Independent Component Analysis (**ICA**)

- Berücksichtigung nicht-linearer Zusammenhänge („***Independent*** Components“)
- Linearkombination der beobachteten Variablen
- Orthogonale Faktoren
- Verteilungsfrei

Non-linear Principal Component Analysis (**nlPCA**)

- Berücksichtigung auch nicht-linearer Zusammenhänge („Independent Components“)
- Keine Linearkombination der beobachteten Variablen
- Modifikation des Mehrschicht-Perzeptrons
- Orthogonale Faktoren
- Verteilungsfrei

Hintergrund

Problem der „Blind Source Separation“

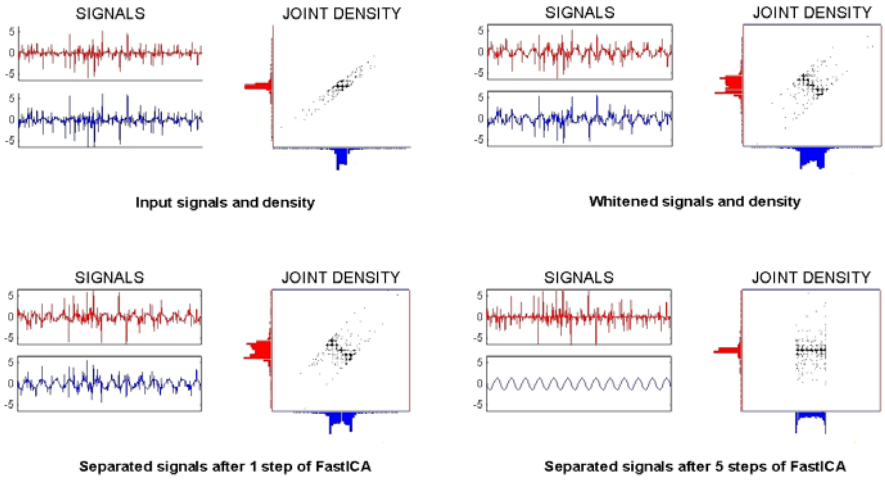
- s. „Cocktail-Party-Problem“: verschiedene Mikrofone nehmen Gespräche und Musik während einer Party auf
- Ziel: Separierung der einzelnen Quellen (Einzelpersonen, einzelne Instrumente)
- Verallgemeinerung: Verschiedene Sensoren empfangen zu unterschiedlichen Anteilen die Signale verschiedener „Sender“

Ansatz: Zentraler Grenzwertsatz

Die Summe einer großen Zahl von unabhängigen, identisch verteilten Zufallsvariablen ($S_n = X_1 + X_2 + ... + X_n$) ist annähernd normalverteilt für $n \rightarrow \infty$.

=> Eine Linearkombination von Zufallsvariablen ist eher normalverteilt als die ursprünglichen Variablen.

ICA-Prinzip

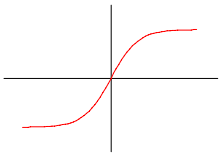


(http://www.cis.hut.fi/projects/ica/icademo/)

Whitening (Sphering)

= Ent-korrelieren, z.B. durch

- Lineare Hauptkomponentenanalyse
- lineare Transformation **y** = $\mathbf{C}^{-1/2}$ **Vx,** wobei **C** = Korrelationsmatrix der Daten
- Unabhängigkeit: Auch jegliche Funktionen der Datensätze sind nicht korreliert, z.B. *f(x) = tanh(x)*



Nicht-Normalität

Verschiedene Maße:

- Momente der Verteilung (-> sensitiv auf einzelne Ausreißer)
- Differentielle Entropie: $H(y) = -\int p(y) \cdot \log p(y)\, dy$

  ist maximal für Normalverteilung
- ...

Bestimmung der Anzahl der Komponenten

- vorab festzulegen
- ist maximal gleich der Anzahl der Variablen
- Mit zunehmender Anzahl der ICs zunehmende Aufspaltung einzelner Komponenten
- Kriterien zur Bestimmung der Anzahl der ICs:
  - Anzahl der (linearen) Hauptkomponenten
  - Stabilität der ICs
  - ...

Beispiel: Vergleich PCA vs. ICA

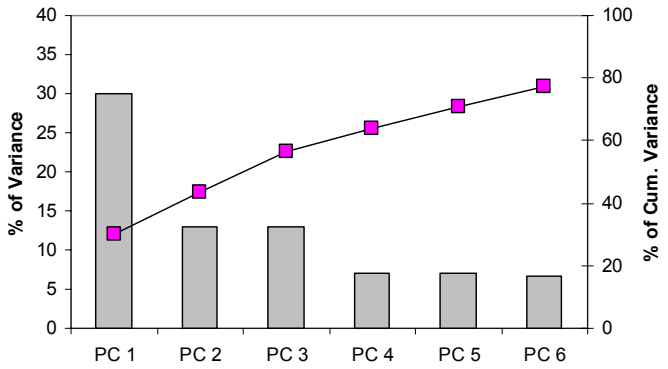
Data: K. Küsel (BITÖK)

→ 1160 Wasserproben, 11 Parameter:

- 3 Moorstandorte
- 11 Sammelperioden à 14 Tage
- 1-40 cm Tiefe (Δz = 1 cm)
- pH, NH_4, NO_3, SO_4, DOC,
  Succinat, Laktat, Formiat, Azetat, Propionat, Butyrat

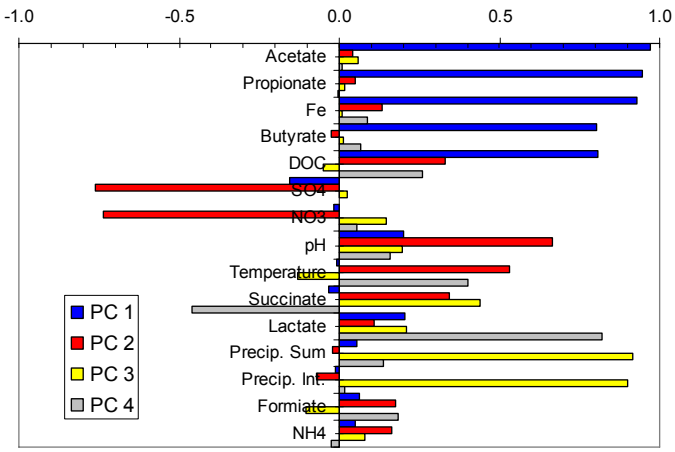
Hauptkomponentenanalyse

Erklärte Varianz

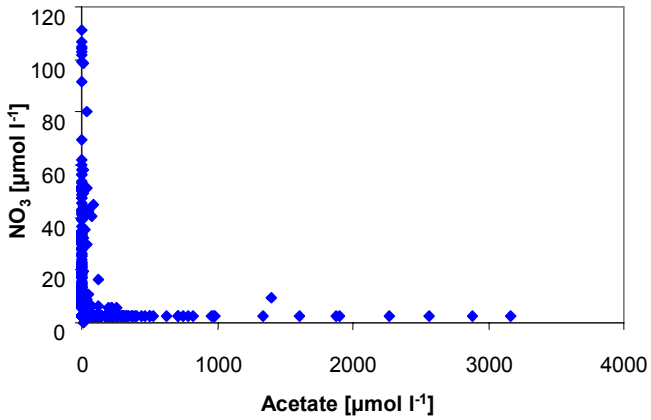


Hauptkomponentenanalyse

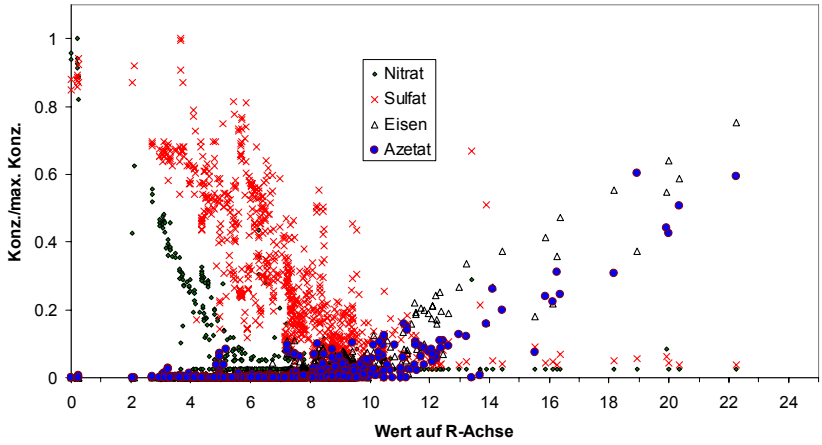
Ladungen (Varimax-Rotation)



NO_3 vs. Azetat

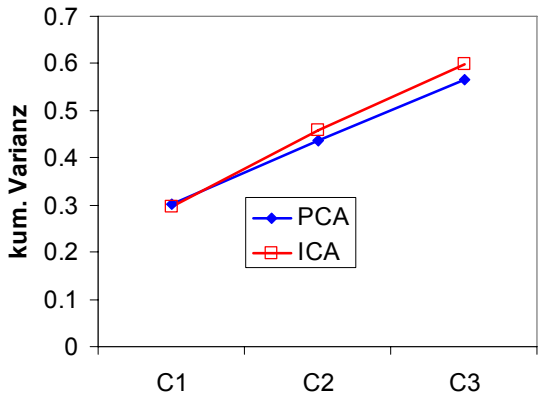


Redox-Sequenz



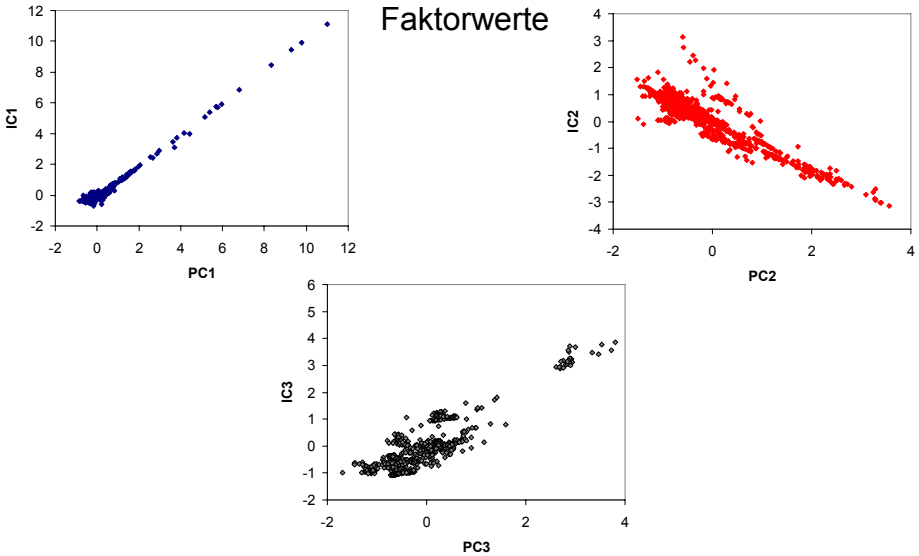
Vergleich PCA – ICA

Erklärte Varianz



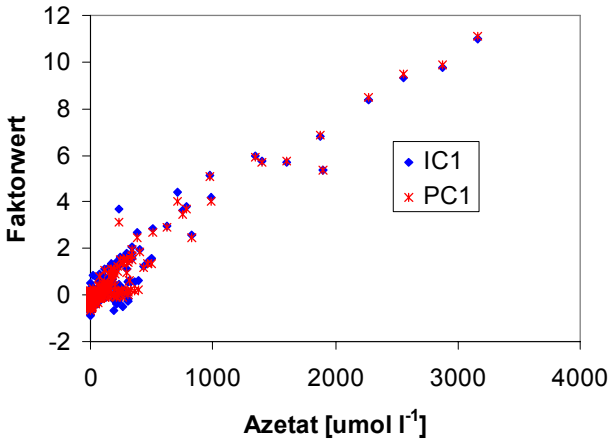
Vergleich PCA – ICA

Faktorwerte



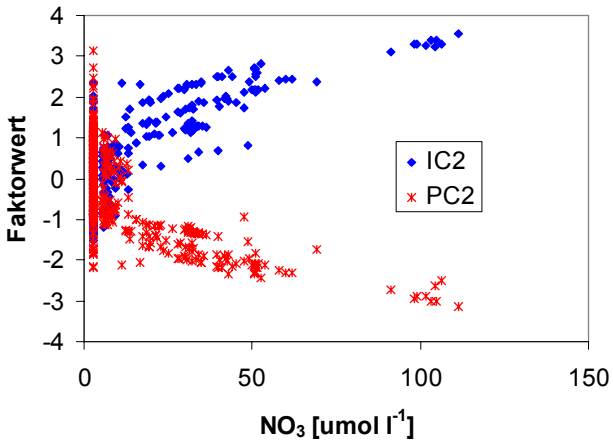
Vergleich PCA – ICA

"Ladungen": Azetat



Vergleich PCA – ICA

"Ladungen": NO_3



Schrittweise Erweiterung der ICs

Ladungen

	1_IC_1	2_IC_1	2_IC_2	3_IC_1	3_IC_2	3_IC_3
mittlere T	0.21	0.01	0.61	0.05	0.08	0.64
Summe N	0.05	0.23	-0.49	-0.02	**0.86**	-0.34
Max N	-0.01	0.17	-0.53	0.01	**0.79**	-0.38
Eisen	**0.92**	**0.92**	0.16	**-0.93**	0.09	0.18
pH	0.45	0.30	0.51	-0.16	0.45	0.60
Sulfat	-0.50	-0.28	**-0.73**	0.18	-0.26	**-0.79**
TOC	**0.92**	**0.86**	0.34	**-0.85**	0.08	0.36
Succinat	0.01	0.00	0.04	0.16	0.56	0.15
Lactat	0.51	0.49	0.15	-0.44	0.22	0.20
Formiat	0.19	0.09	0.31	-0.14	-0.21	0.28
Acetat	**0.91**	**0.95**	0.05	**-0.95**	0.11	0.07
Propionat	**0.89**	**0.92**	0.05	**-0.93**	0.08	0.07
Butyrat	**0.81**	**0.85**	0.02	**-0.86**	0.07	0.04
Nitrat	-0.28	-0.05	**-0.69**	-0.02	-0.12	**-0.73**
Ammonium	0.11	0.09	0.09	-0.05	0.12	0.12

ICA: Bewertung

Vorteile:

- Verteilungsfrei
- ICs als Linearkombinationen der Parameter analog zur PCA interpretierbar
- Liefert oft ähnliche Ergebnisse wie PCA
- Sukzessive Verfeinerung durch Erhöhung der Zahl der ICs möglich

Nachteile:

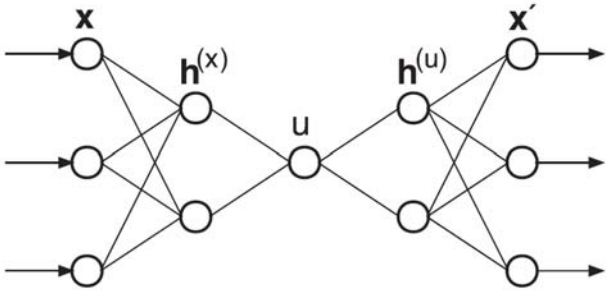
- Keine klaren Kriterien zur Festlegung der Anzahl der ICs
- Ergebnis u.U. stark von einzelnen Ausreißern abhängig
- Nicht in Standard-Software-Paketen enthalten

Nicht-lineare Hauptkomponentenanalyse

(Kramer 1991, Hsieh 2001)

Ansatz: Mehrschicht-Perzeptron mit „Flaschenhals“

hier: 3:2:1:2:3-Netz

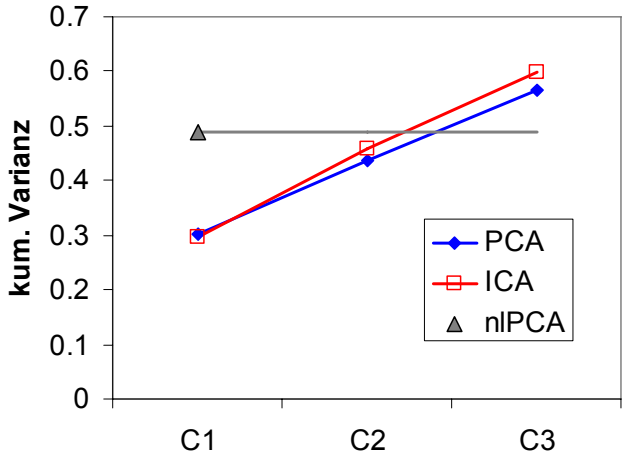


(Hsieh 2004)

Sukzessive Bestimmung der Hauptkomponenten

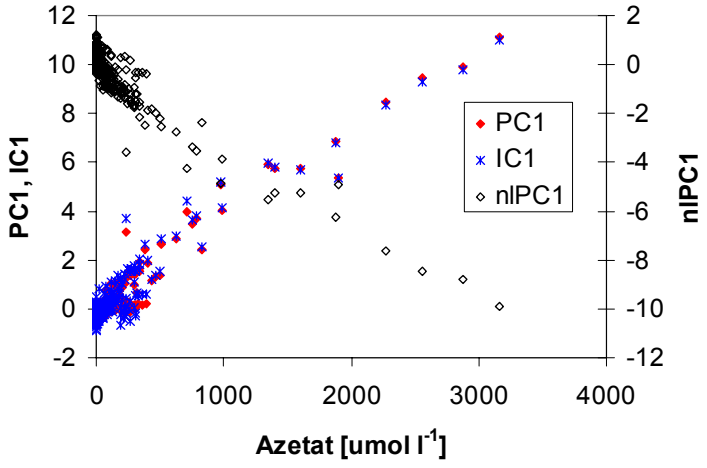
1. Bestimmung der 1. nlPC
2. Bestimmung der 2. nlPC anhand der Residuen von 1.
3. Bestimmung der 3. nlPC anhand der Residuen von 2.
4. ...

Erklärte Varianz



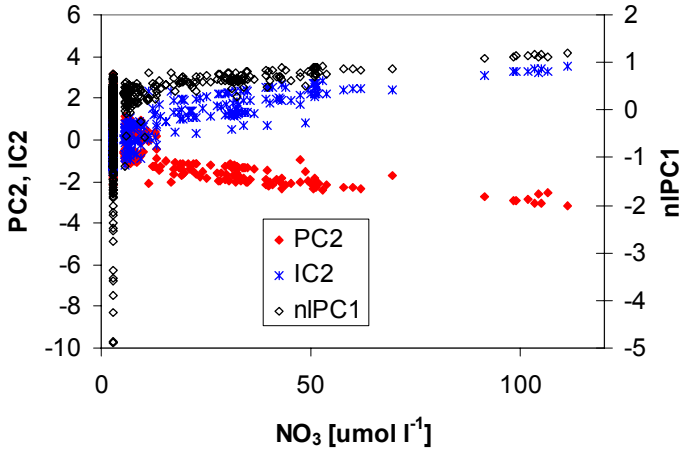
Vergleich PCA – ICA - nlPCA

"Ladungen": Azetat

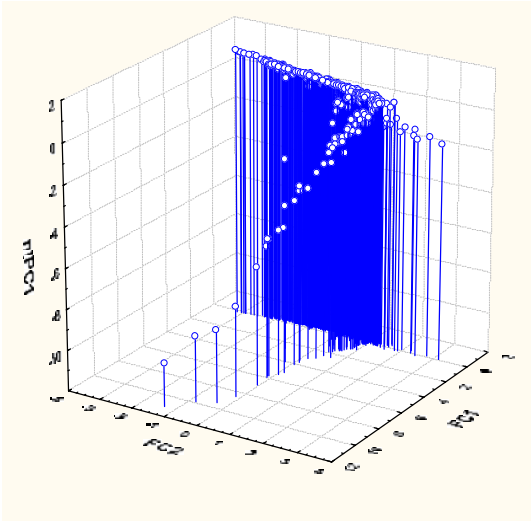


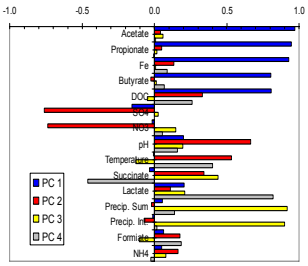
Vergleich PCA – ICA - nlPCA

"Ladungen": NO_3



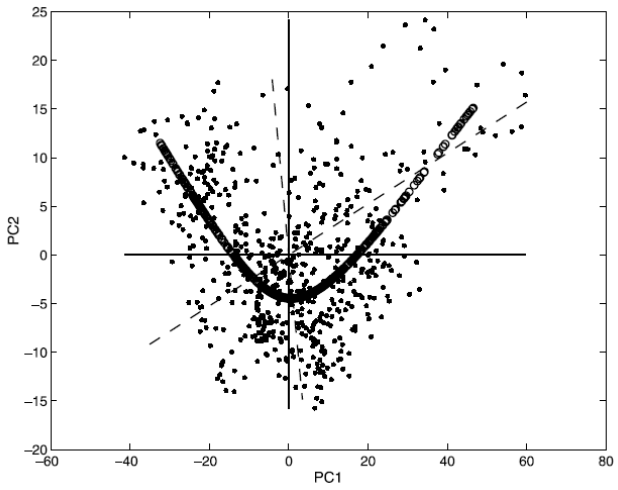
Vergleich PCA – nlPCA





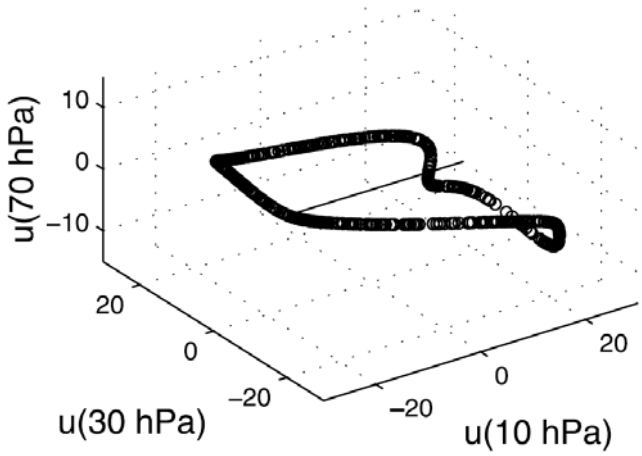
nlPCA: Beispiel

(Sea Surface Temperature; Hsieh 2004)

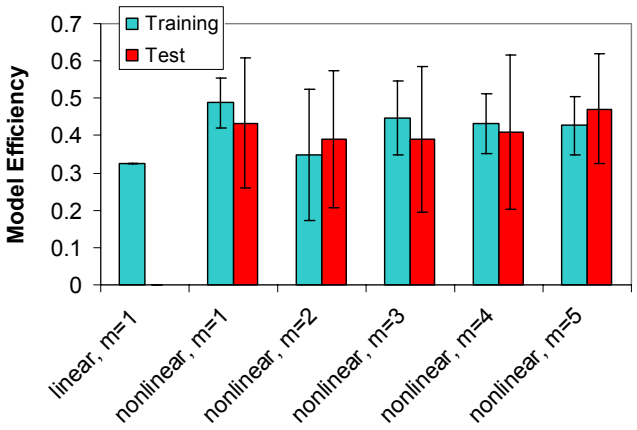


nlPCA: Beispiel

(Äquatorialer stratospherischer zonaler Wind; Hsieh 2004)



Stabilität der Lösung



nlPCA: Bewertung

Vorteile:

- Verteilungsfrei
- Explizite Berücksichtigung nicht-linearer Zusammenhänge, dadurch oft weniger und besser interpretierbare PCs (*im Gegensatz zur stückweisen Linearisierung der PCA*)

Nachteile:

- Vorteile nur bei großen Datensätzen
- Hoher Rechen-Aufwand
- Abhängigkeit vom gewählten Lernverfahren
- Viele Wiederholungen (Kreuzvalidierung) nötig
- Nicht in Standard-Software-Paketen enthalten

Aufgabe

1. Führen Sie mit OOLABSS.exe eine ICA mit einem der Beispielsdatensätze durch. Variieren Sie einzelne Parameter.
2. Importieren Sie den gegebenen Datensatz und führen Sie eine ICA durch. Variieren Sie einzelne Parameter. Exportieren Sie die ICs (= *Sources*).
3. Bestimmen Sie die Ladungen der einzelnen ICs.
4. Untersuchen Sie die Stabilität der Lösung, indem Sie einzelne Ausreißer in den Datensatz einfügen.
5. Vergleichen Sie die Ergebnisse der ICA mit denen der PCA (Varimax-Rotation).