Mathematics of Data Science

(oder: Mathematische Verfahren für hochdimensionale Daten)

Matthias Birkner

Hauptseminar, WS 2023/24

Ein (kleiner) Beispieldatensatz

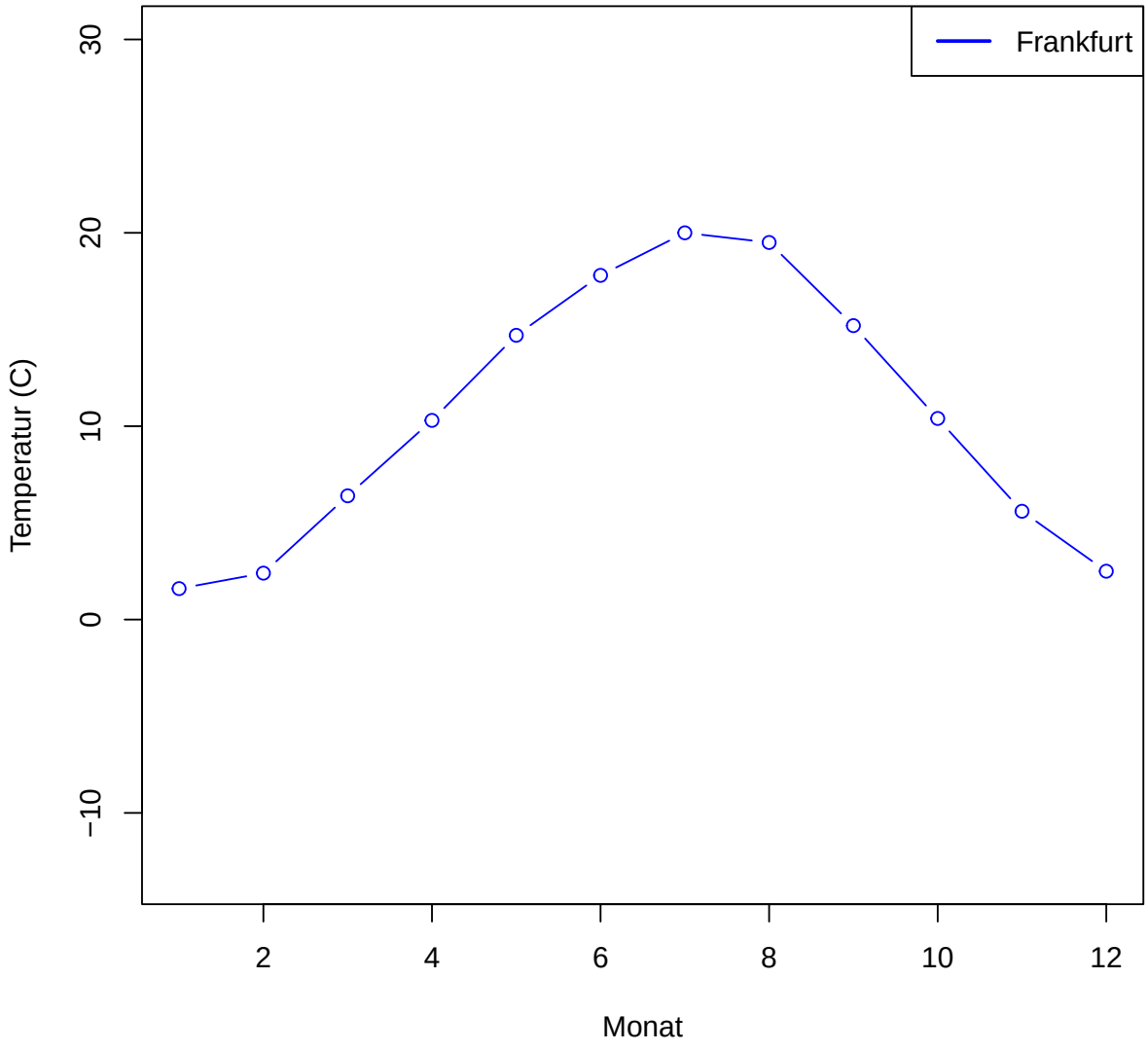
Monatliche Durchschnittstemperaturen in 73 europäischen Städten

($n = 73$ Beobachtungen im p-dimensionalen Raum, $p = 12$)

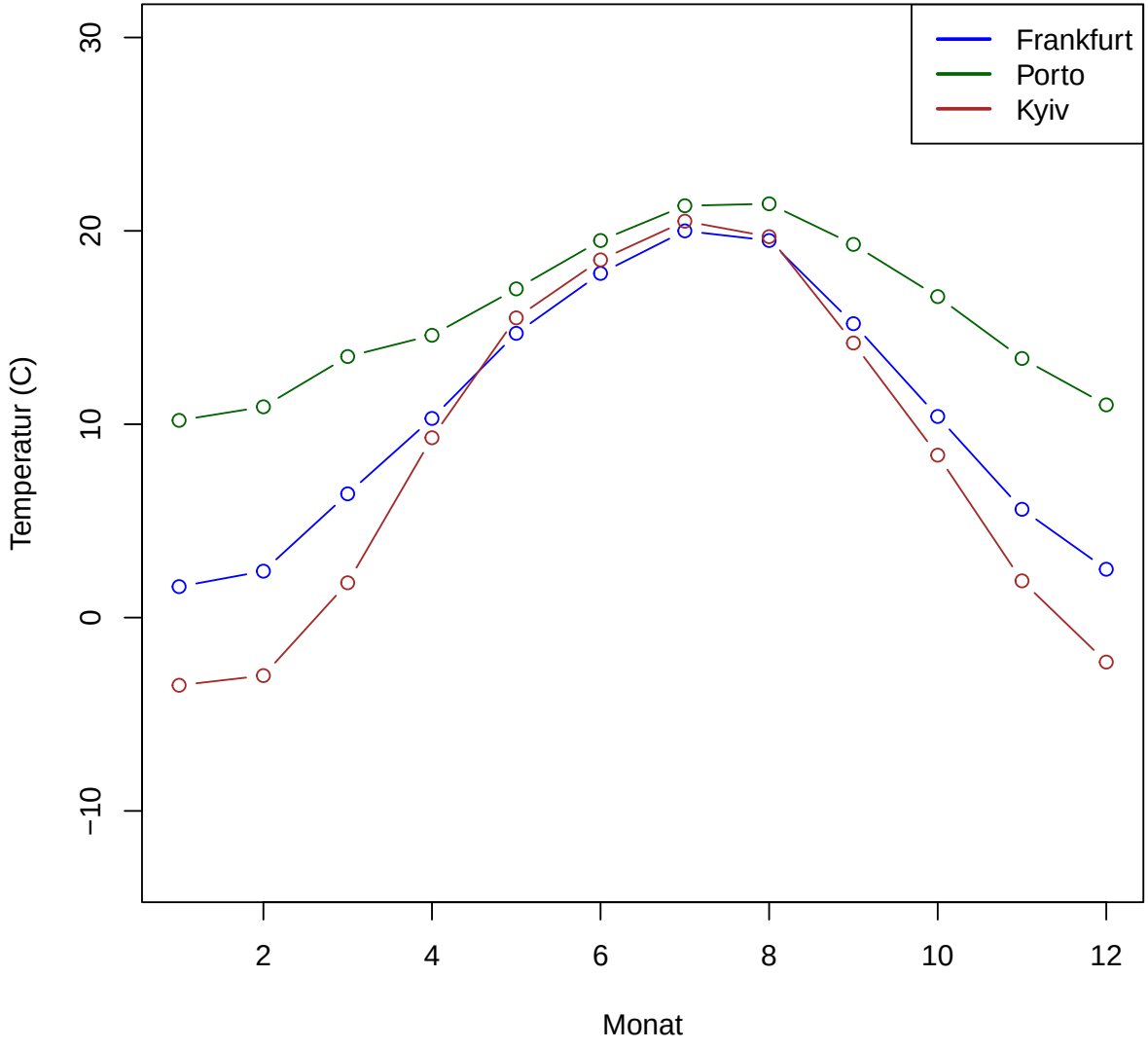
aus `https://en.wikipedia.org/wiki/List_of_cities_by_average_temperature#Europe`

	Jan	Feb	Mar	Apr	Mai	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Tirana	6.7	7.8	10.0	13.4	18.0	21.6	24.0	23.8	20.7	16.0	11.7	8.1
Andorra	2.2	3.5	5.8	7.5	11.5	15.4	18.8	18.5	14.9	10.3	5.7	3.0
Vienna	0.3	1.5	5.7	10.7	15.7	18.7	20.8	20.2	15.4	10.2	5.1	1.1
Minsk	-4.5	-4.4	0.0	7.2	13.3	16.4	18.5	17.5	12.1	6.6	0.6	-3.4
Brussels	3.3	3.7	6.8	9.8	13.6	16.2	18.4	18.0	14.9	11.1	6.8	3.9
Sarajevo	-0.5	1.4	5.7	10.0	14.8	17.7	19.7	19.7	15.3	11.0	5.4	0.9
Sofia	-0.5	1.1	5.4	10.6	15.4	18.9	21.2	21.0	16.5	11.3	5.1	0.7
Zagreb	0.3	2.3	6.4	10.7	15.8	18.8	20.6	20.1	15.9	10.5	5.0	1.4
Split	8.0	8.4	10.6	13.7	18.9	22.8	25.7	25.4	21.2	16.8	12.0	9.1
Nicosia	10.6	10.6	13.1	17.1	22.3	26.9	29.7	29.4	26.2	22.3	16.3	12.0
Prague	-1.4	-0.4	3.6	8.4	13.4	16.1	18.2	17.8	13.5	8.5	3.1	-0.3
Copenhagen	1.4	1.4	3.5	7.7	12.5	15.6	18.1	17.7	13.9	9.8	5.5	2.5
Tallinn	-2.9	-3.6	-0.6	4.8	10.2	14.5	17.6	16.5	12.0	6.5	2.0	-0.9
Helsinki	-3.9	-4.7	-1.3	3.9	10.2	14.6	17.8	16.3	11.5	6.6	1.6	-2.0
Kuopio	-9.2	-9.2	-4.1	2.0	9.1	14.5	17.5	15.0	9.7	4.1	-2.0	-6.7
Oulu	-9.6	-9.3	-4.8	1.4	7.8	13.5	16.5	14.1	8.9	3.3	-2.8	-7.1
Lyon	3.4	4.8	8.4	11.4	15.8	19.4	22.1	21.6	17.6	13.4	7.5	4.3
Marseille	8.4	8.9	11.6	13.8	17.9	21.3	24.5	24.1	20.7	16.9	11.8	9.3
Paris	4.9	5.6	8.8	11.4	15.1	18.2	20.4	20.2	16.9	12.9	8.1	5.4
Berlin	0.6	2.3	5.1	10.2	14.8	17.9	20.3	19.7	15.3	10.5	6.0	1.3
Frankfurt	1.6	2.4	6.4	10.3	14.7	17.8	20.0	19.5	15.2	10.4	5.6	2.5
Athens	10.2	10.9	13.2	16.9	21.8	26.6	29.3	29.3	25.0	20.1	15.5	11.5
Heraklion	12.1	12.2	13.6	16.6	20.4	24.5	26.4	26.3	23.7	20.3	16.8	13.8
Thessaloniki	5.8	7.1	10.1	14.1	19.3	24.2	26.6	26.5	22.0	16.9	11.8	7.2
Budapest	0.4	2.3	6.1	12.0	16.6	19.7	21.5	21.2	16.9	11.8	5.4	1.8
Reykjavík	-0.5	0.4	0.5	2.9	6.3	9.0	10.6	10.3	7.4	4.4	1.1	-0.2
Dublin	5.3	5.3	6.8	8.3	10.9	13.6	15.6	15.3	13.4	10.5	7.4	5.6
Milan	2.5	4.7	9.0	12.2	17.0	20.8	23.6	23.0	19.2	13.4	7.2	3.3
Palermo	12.5	12.6	13.5	15.7	18.9	22.4	25.6	26.2	24.1	20.3	16.8	13.7
Rome	7.5	8.2	10.2	12.6	17.2	21.1	24.1	24.5	20.8	16.4	11.4	8.4
Napoli	8.7	8.8	11.1	13.2	17.8	21.4	24.3	24.7	21.4	17.1	12.4	9.8
Riga	-4.7	-4.2	0.5	5.1	11.4	15.5	16.9	16.2	12.0	7.4	2.1	-2.3

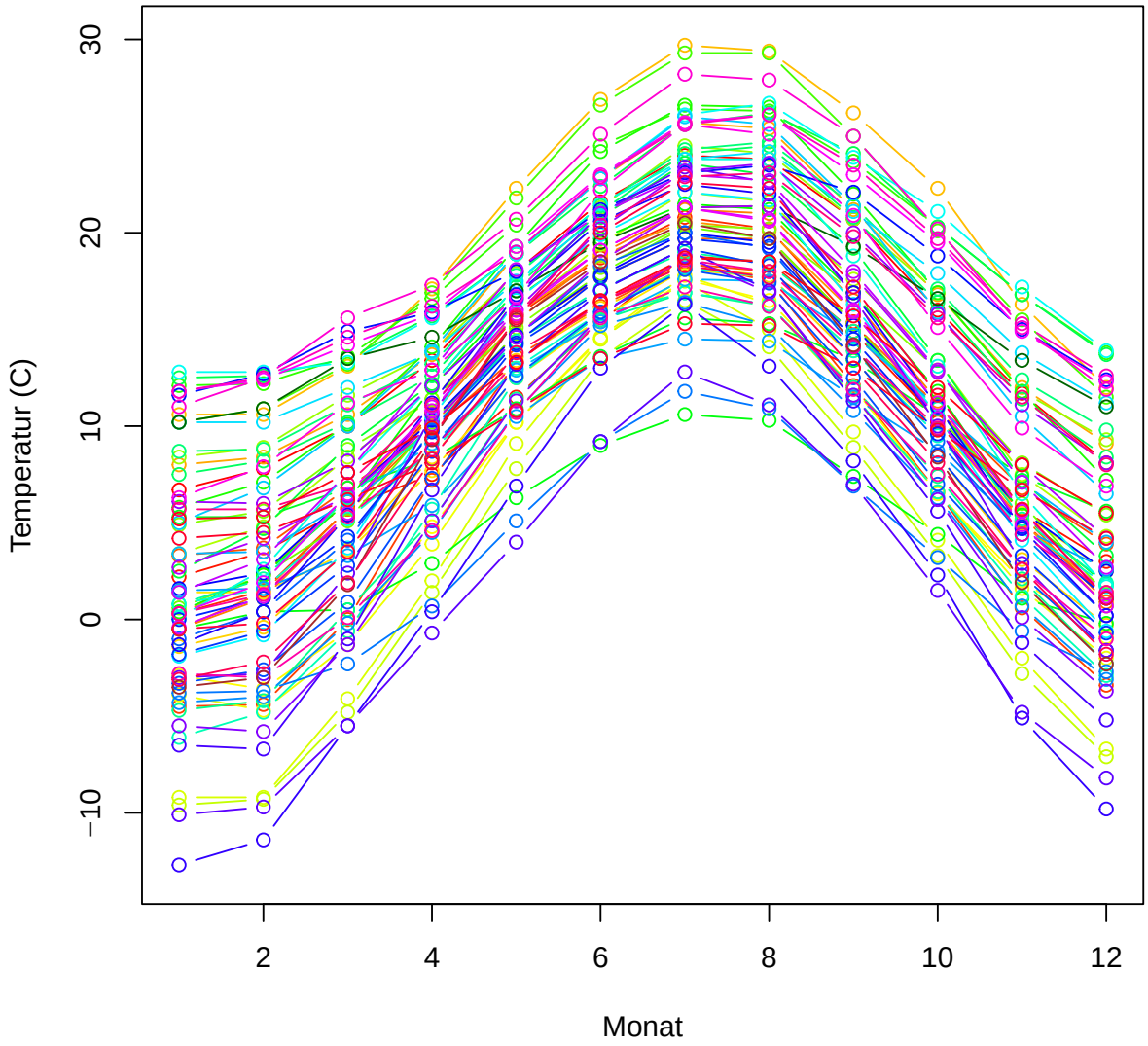
Versuch, die Daten graphisch aufzubereiten



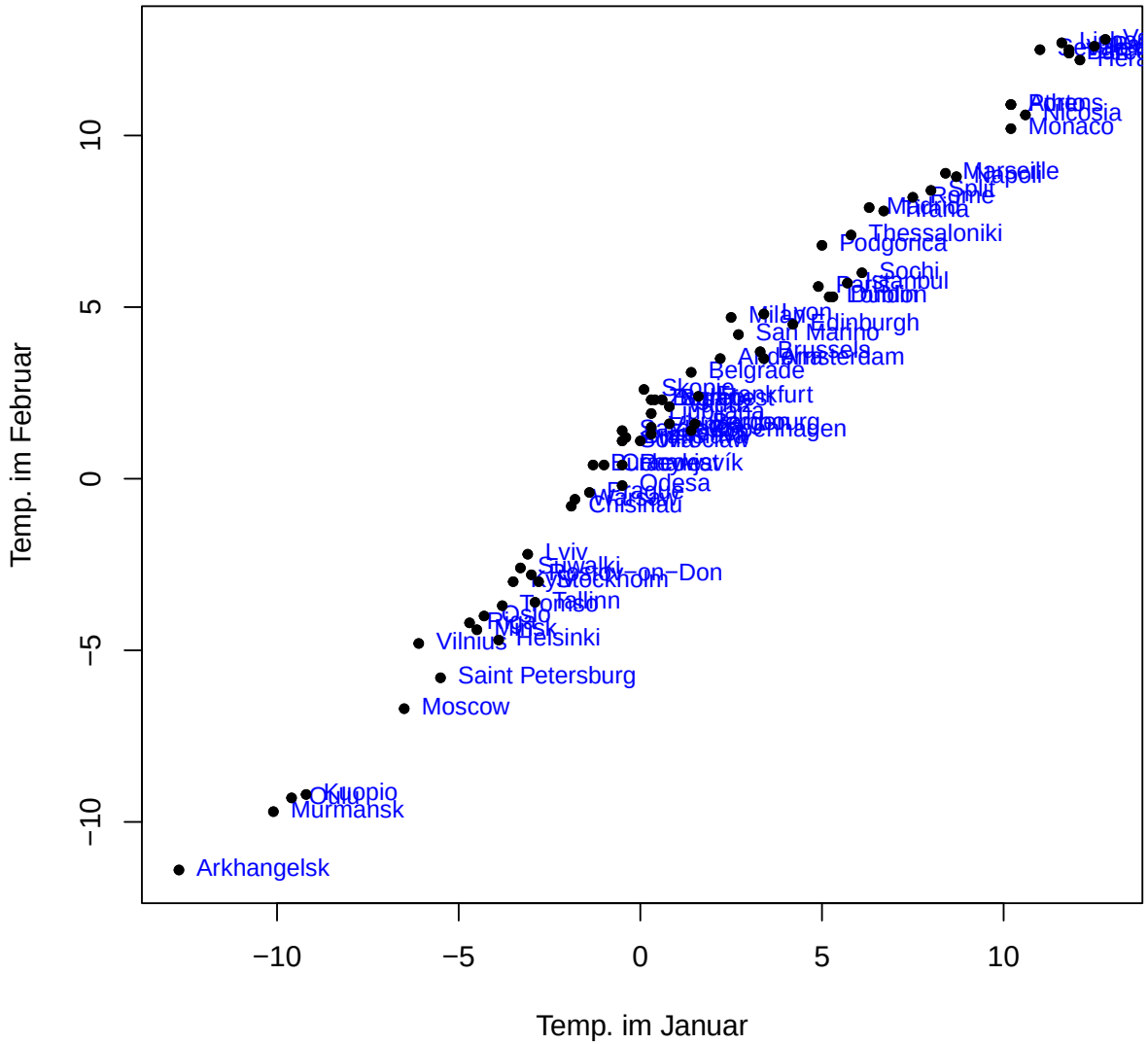
Versuch, die Daten graphisch aufzubereiten



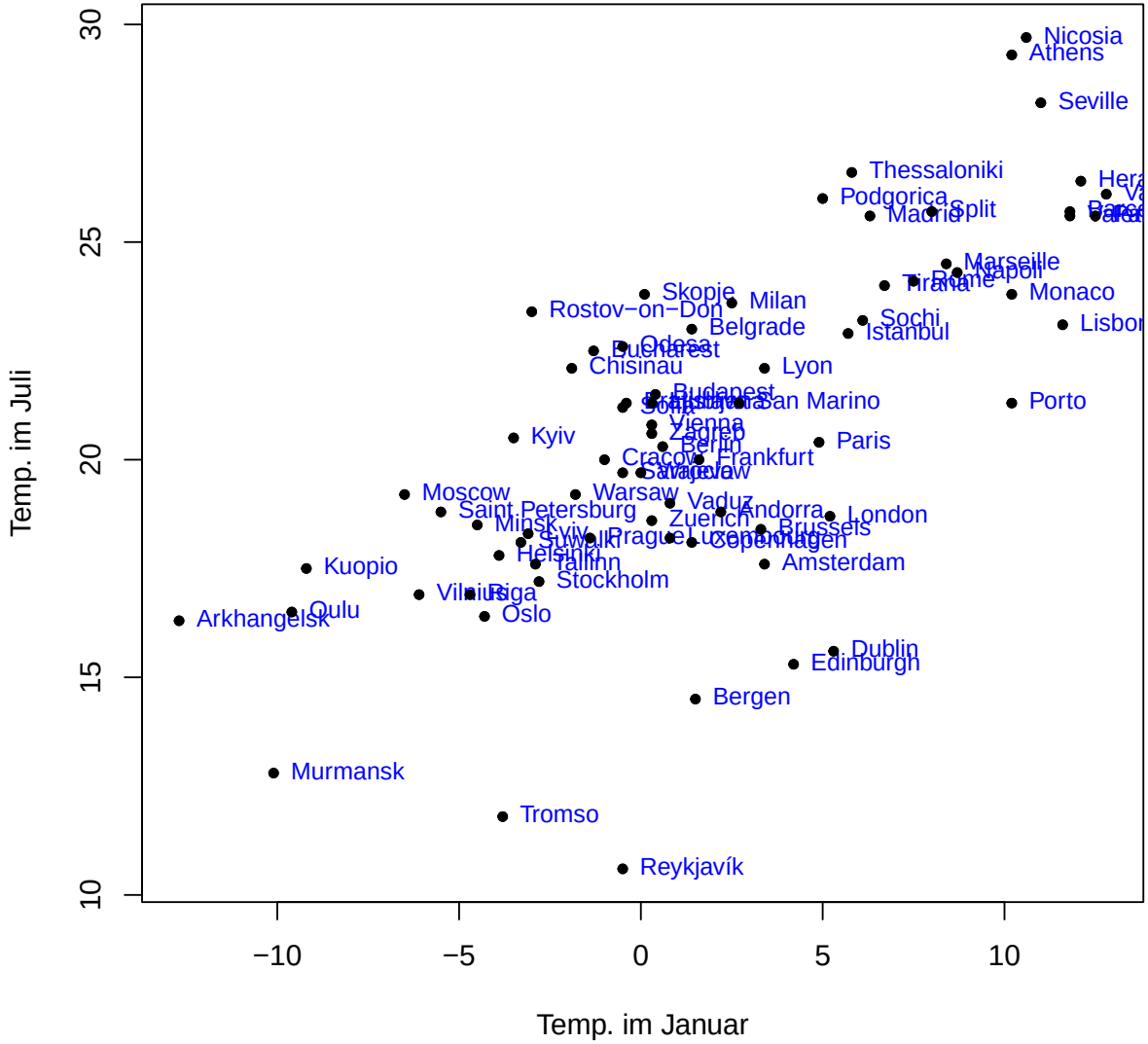
Versuch, die Daten graphisch aufzubereiten: gescheitert?



Ein zweidimensionale Projektion



Ein andere zweidimensionale Projektion



Idee der Hauptkomponentenanalyse

$x = (x_{ij})$, $x_{ij} =$ (Durchschnitts-)Temperatur in Stadt i in Monat j

$x^{(i)} = (x_{i,1}, \ldots, x_{i,12})$ i-te Zeile der Datenmatrix ($i = 1, \ldots, n$, $n = 73$)

Betrachte zunächst 1-dimensionale Projektionen:

$u \in \mathbb{R}^{12}$ mit $||u||_2 = 1$, bilde $x^{(1)} \cdot u, x^{(2)} \cdot u, \ldots, x^{(73)} \cdot u$

Ziel: möglichst große (empirische) Varianz dieser Werte (denn so bleibt möglichst viel der ursprünglichen Variabilität der Daten erhalten)

Welches u soll man wählen?

Idee der Hauptkomponentenanalyse

$x = (x_{ij})$, $x_{ij} =$ (Durchschnitts-)Temperatur in Stadt i in Monat j

$x^{(i)} = (x_{i,1}, \ldots, x_{i,12})$ i-te Zeile der Datenmatrix ($i = 1, \ldots, n$, $n = 73$)

Betrachte zunächst 1-dimensionale Projektionen:

$u \in \mathbb{R}^{12}$ mit $||u||_2 = 1$, bilde $x^{(1)} \cdot u, x^{(2)} \cdot u, \ldots, x^{(73)} \cdot u$

Ziel: möglichst große (empirische) Varianz dieser Werte (denn so bleibt möglichst viel der ursprünglichen Variabilität der Daten erhalten)

Welches u soll man wählen?

Es stellt sich heraus: Die Varianz dieser Werte ist $u^T \widehat{\Sigma} u$ mit

$$\widehat{\Sigma} = (\widehat{\Sigma}_{jk}), \quad \widehat{\Sigma}_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_{\cdot,j})(x_{ik} - \overline{x}_{\cdot,k})$$

die (empirische) Kovarianzmatrix

Idee der Hauptkomponentenanalyse, 2

$x^{(i)} = (x_{i,1}, \ldots, x_{i,12})$ i-te Zeile der Datenmatrix ($i = 1, \ldots, n$, $n = 73$)

Projektionen in Richtung Einheitsvektor $u \in \mathbb{R}^{12}$ haben (empirische) Varianz $u^T \widehat{\Sigma} u$ ($=^!$ groß)

Demnach: u sollte Eigenvektor von $\widehat{\Sigma}$ zum größten Eigenwert sein

Idee der Hauptkomponentenanalyse, 2

$x^{(i)} = (x_{i,1}, \ldots, x_{i,12})$ i-te Zeile der Datenmatrix ($i = 1, \ldots, n$, $n = 73$)

Projektionen in Richtung Einheitsvektor $u \in \mathbb{R}^{12}$ haben
(empirische) Varianz $u^T \widehat{\Sigma} u$ ($=^!$ groß)

Demnach: u sollte Eigenvektor von $\widehat{\Sigma}$ zum größten Eigenwert sein

Die Eigenwerte von $\widehat{\Sigma}$ sind im Beispiel

262.19334643, 19.83611122, 2.37770120, 0.62443159, 0.25743423, 0.13861559,
0.10674104, 0.06414481, 0.04278174, 0.03640898, 0.02528569, 0.02047160

Der Eigenvektor zum größten EW ist

$$u^{(1)} = (0.3520060, 0.3573429, 0.3171045, 0.2514131, 0.2105548, 0.2051480, 0.2127255, 0.2470177, 0.2813611, 0.3035283, 0.3263267, 0.3378793)^T$$

(zum Vergleich: $1/\sqrt{12} \approx 0.2886751$)

Idee der Hauptkomponentenanalyse, 2

$x^{(i)} = (x_{i,1}, \ldots, x_{i,12})$ i-te Zeile der Datenmatrix ($i = 1, \ldots, n$, $n = 73$)

Projektionen in Richtung Einheitsvektor $u \in \mathbb{R}^{12}$ haben
(empirische) Varianz $u^T \widehat{\Sigma} u$ ($=^!$ groß)

Demnach: u sollte Eigenvektor von $\widehat{\Sigma}$ zum größten Eigenwert sein

Die Eigenwerte von $\widehat{\Sigma}$ sind im Beispiel

262.19334643, 19.83611122, 2.37770120, 0.62443159, 0.25743423, 0.13861559,
0.10674104, 0.06414481, 0.04278174, 0.03640898, 0.02528569, 0.02047160

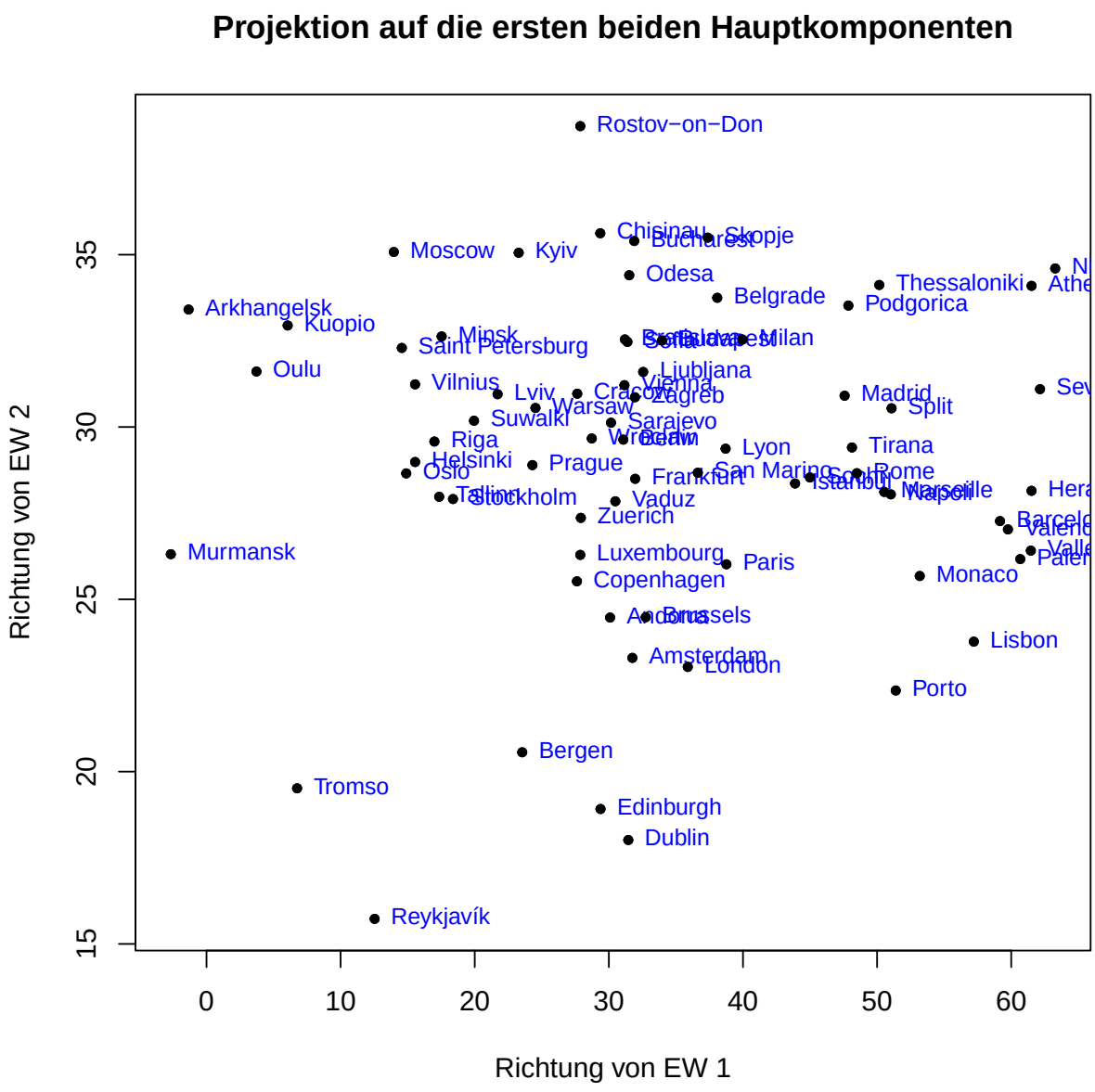
Der Eigenvektor zum größten EW ist

$$u^{(1)} = (0.3520060, 0.3573429, 0.3171045, 0.2514131, 0.2105548, 0.2051480, 0.2127255, 0.2470177, 0.2813611, 0.3035283, 0.3263267, 0.3378793)^T$$

(zum Vergleich: $1/\sqrt{12} \approx 0.2886751$)

Zudem

$$u^{(2)} = (-0.36890735, -0.31622814, -0.11610145, 0.15599227, 0.32665128, 0.39567751, 0.41417584, 0.36619950, 0.19791458, 0.03033259, -0.15048997, -0.30732893)^T$$


Projektion auf die ersten beiden Hauptkomponenten
Richtung von EW 2
Richtung von EW 1
15
20
25
30
35
0
10
20
30
40
50
60
Rostov−on−Don
Chisinau
Bucharest
Skopje
Moscow
Kyiv
Odesa
Belgrade
Thessaloniki
Podgorica
Arkhangelsk
Kuopio
Minsk
Saint Petersburg
Milan
Oulu
Ljubljana
Vilnius
Lviv
Vienna
Zagreb
Madrid
Split
Warsaw
Suwalki
Sarajevo
Riga
Lyon
Tirana
Helsinki
Prague
Oslo
Frankfurt
San Marino
Rome
Stockholm
Vaduz
Zuerich
Murmansk
Luxembourg
Paris
Copenhagen
Monaco
Brussels
Lisbon
Amsterdam
London
Porto
Bergen
Tromso
Edinburgh
Dublin
Reykjavík

Das “Spike-Problem”

$p = 1000$, wir wählen einen Einheitsvektor $w \in \mathbb{R}^p$
(die 1-dimensionale “Signalrichtung”)

simulieren $n = 2000$ Kopien von

$$X = \sqrt{\beta} Z_0 w + (Z_1, Z_2, \ldots, Z_p)$$

mit $Z_0, Z_1, \ldots, Z_p$ unabhängig und standard-normalverteilt
($\beta > 0$ ist ein Parameter)

Das “Spike-Problem”

$p = 1000$, wir wählen einen Einheitsvektor $w \in \mathbb{R}^p$
(die 1-dimensionale “Signalrichtung”)

simulieren $n = 2000$ Kopien von

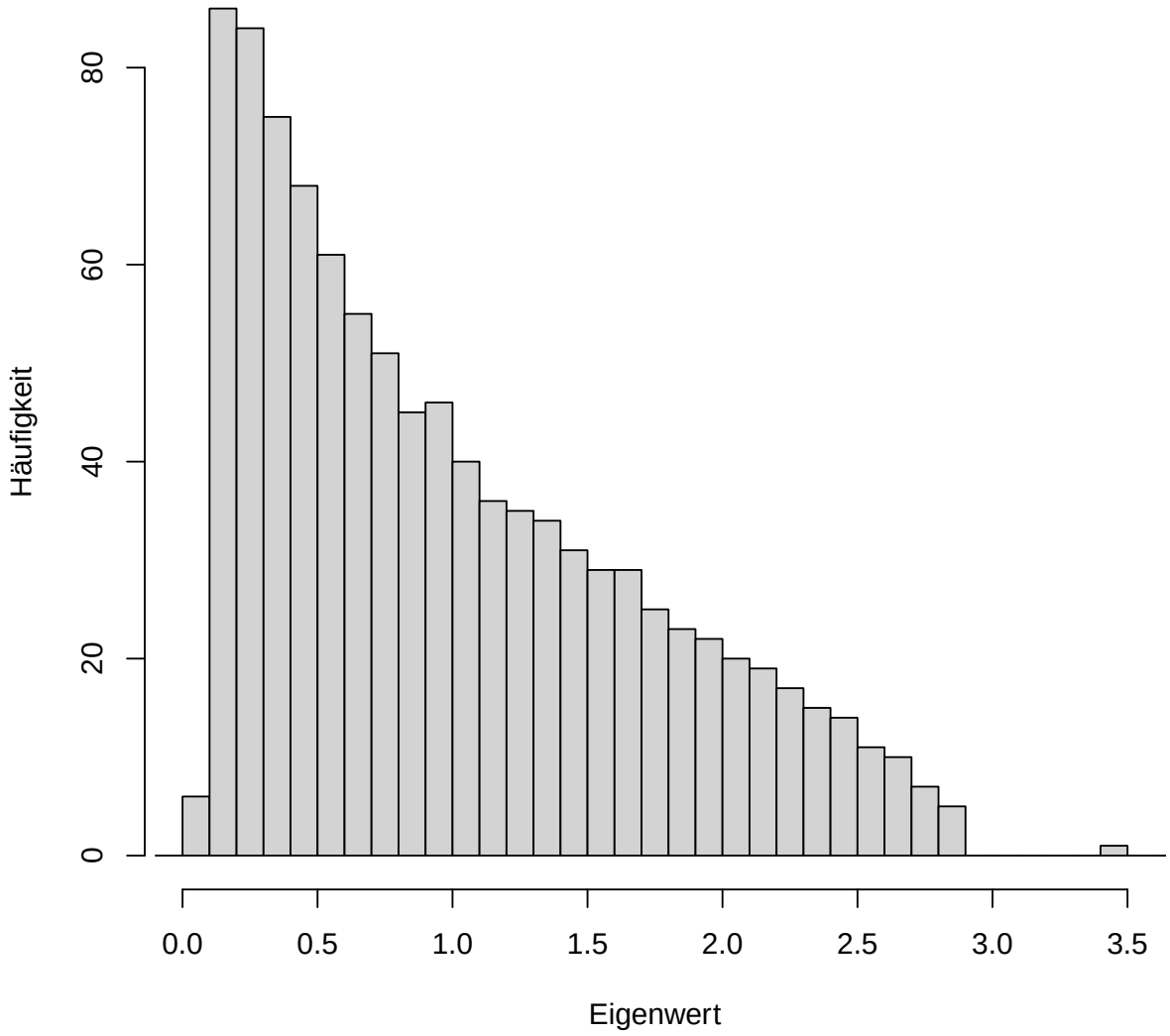
$$X = \sqrt{\beta} Z_0 w + (Z_1, Z_2, \ldots, Z_p)$$

mit $Z_0, Z_1, \ldots, Z_p$ unabhängig und standard-normalverteilt
($\beta > 0$ ist ein Parameter)

Die resultierenden $X^{(1)}, X^{(2)}, \ldots, X^{(2000)}$ fassen wir als Beobachtungsvektoren in einer Hauptkomponentenanalyse auf.
Können wir w wiederfinden?
(Oder zumindest entscheiden, ob $\beta \neq 0$?)

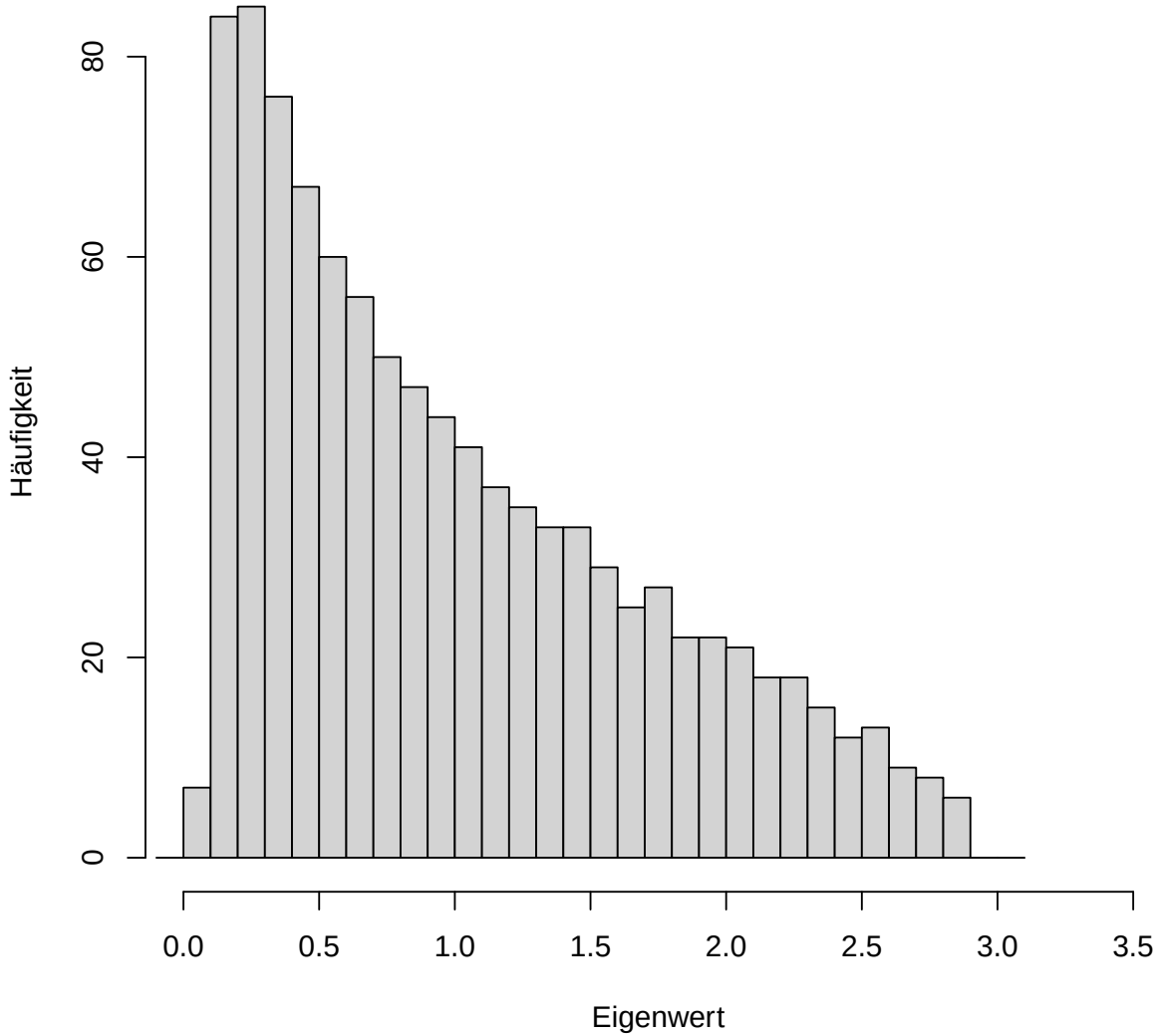
(Bem.: Die (theoretische) Kovarianzmatrix von X ist $I_p + \beta w w^T$.)

$\beta = 1.5$



Es ist $u^{(1)} \cdot w = 0.7750505$

$\beta = 0.5$



Es ist $u^{(1)} \cdot w = 0.0002692055$

Voraussetzungen

Grundvorlesungen Analysis und Lineare Algebra, Grundlagen der Stochastik

Quelle(n)

Afonso S. Bandeira, Amit Singer, Thomas Strohmer, *Mathematics of Data Science*, Book in preparation, 2023.
https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf

Afonso S. Bandeira, Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science. Lecture notes, 2016.
https://people.math.ethz.ch/~abandeira/TenLecturesFortyTwoProblems.pdf

Roman Vershynin, *High dimensional probability. An introduction with applications in Data Science.* Cambridge University Press, 2018.
https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf

Ggf. weitere Quellen aus der aktuellen Literatur und/oder aus den Weiten des Internets

Besprechung & Themenvergabe : Mi., 19. Juli 2023, 14 h

https://www.staff.uni-mainz.de/birkner/MoDS2324/