Model Fitting: Von Empirischen Daten Zur Theoretischen Verteilung

Dr. Michael Thrun

thrun@deepbionics.de

Aufteilung

1. Powerpoint Slides: Kurze Einführung in theoretische Elemente und theoretische Beispiele
2. Rmarkdown: Praktische Beispiele in R

⇒Um Datenwissenschaften zu lernen empfehle ich „learning bei doing!

⇒Reproduktion der Beispiele als „Hausaufgabe“

⇒Sowie Anwendung des Gelernten Anhand weiteren Daten

⇒Alles ist online verfügbar unter https://github.com/Mthrun/ModelFittingData2PDF2021/

Heutige Lernziele

- Verständnis der Wahrscheinlichkeitsdichteverteilung (PDF)
  - Kurzer Überblick über 3 gängige Methoden zur Schätzung von PDFs
- Unterschied des Vorgehens von Datenwissenschaftlern im Gegensatz zur üblichen Statistik

- Kür: (Interaktive) Modellierung von Gaußmixturenmodellen (GMM) anhand von Beispielen

Verteilung eines Merkmals X_1

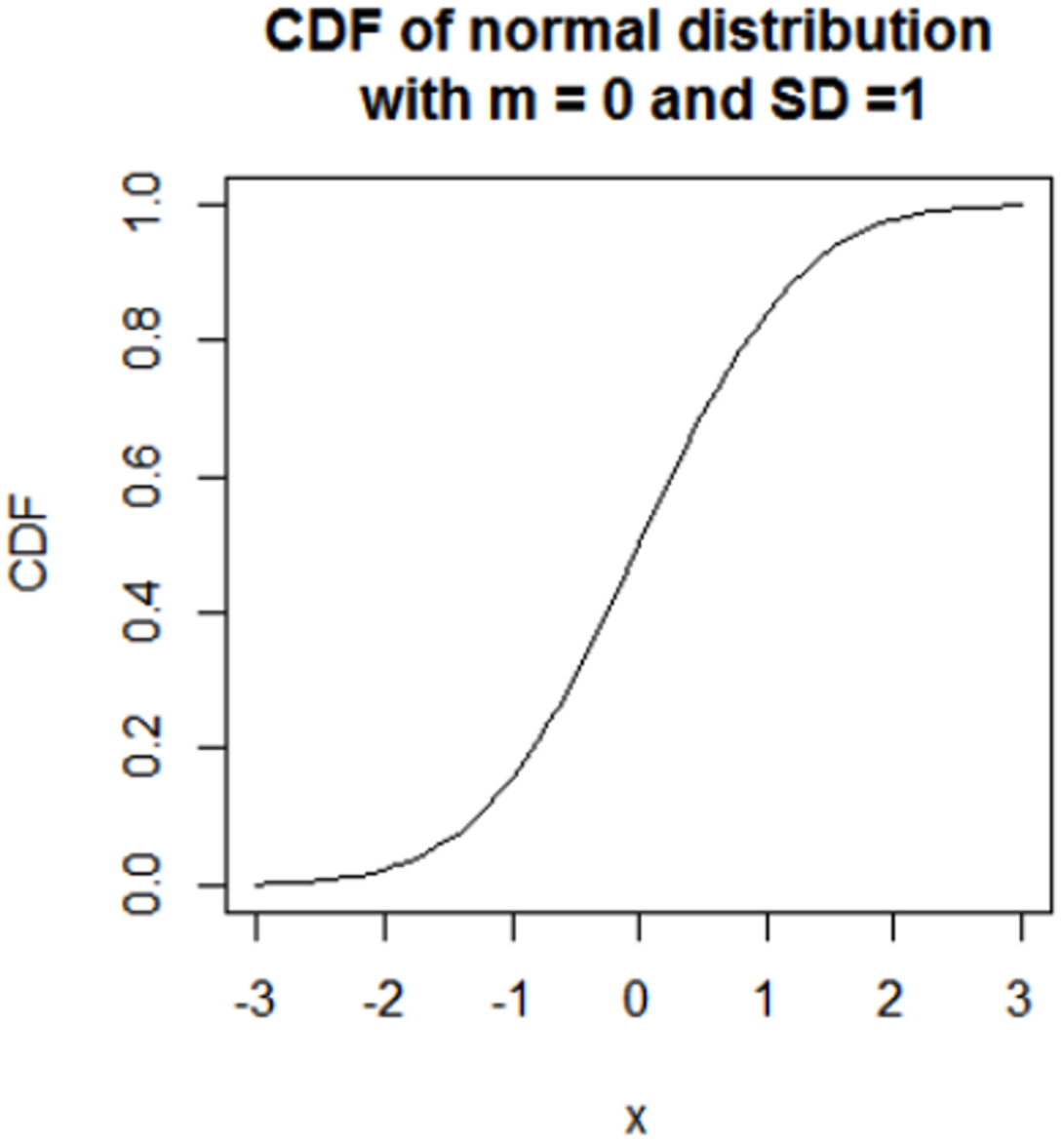
- In der Statistik geht man davon aus, dass Daten X_1 durch ein Zufallsexperiment mit einer Wahrscheinlichkeit erzeugt werden
- F(t) heißt *Verteilungsfunktion der Zufallsvariablen* X_1, wenn

$$F(t) = p(x \leq t)$$

- F(t) gibt für eine Schwelle t an, wie wahrscheinlich es ist einen Wert $x \leq$ t zu erhalten
- Im Englischen werden Verteilungsfunktionen auch cumulative distribution functions (cdf(t) = F(t)) genannt

CDF - *Cumulative Distribution Function*

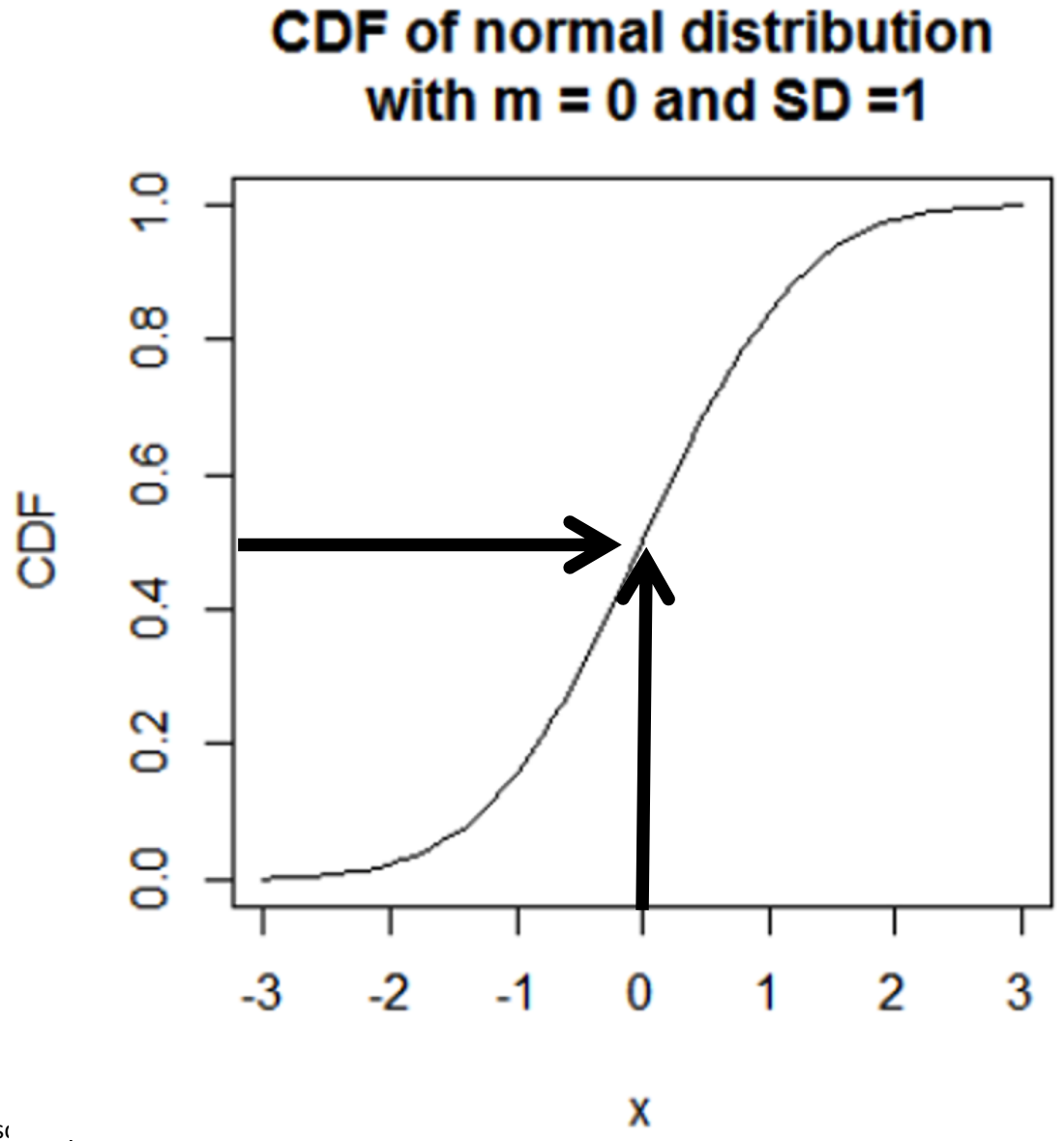
- Verteilungsfunktionen sind monoton wachsende Funktionen in [0,1]
- Das rechte Bild zeigt die Verteilungsfunktionen eines Merkmals, welches einer Gauss-Verteilung bzw. Normalverteilung ist



CDF - *Cumulative Distribution Function*

- Verteilungsfunktionen sind monoton wachsende Funktionen in [0,1]
- Das rechte Bild zeigt die Verteilungsfunktionen eines Merkmals, welches Gauss-verteilt bzw. Normalverteilt ist
  - Bsp. F(t=0) = p($x_1 \leq 0$) =0.5

=> In dieser Verteilung hat man 50% Wahrscheinlichkeit Werte kleiner oder gleich 0 zu ziehen



Verteilungsfunktion

Wenn die Verteilungsfunktion F(t) dargestellt werden als kann :

$$F(t) = \int_{-\infty}^{t} f(x)dx$$

- Dann nennt man f(x) die Wahrscheinlichkeitsdichtefunktion oder kürzer Dichtefunktion oder Dichte.
- Im Englischen wird die auch probability density function (pdf) oder Likelihood benutzt.
- Durch die Angabe der Dichte wird eine Wahrscheinlichkeitsverteilung und somit auch die Verteilungsfunktionen eindeutig bestimmt.

Beispiele in Rmardown

- Siehe Rmarkdown: 01Verteilung.Rmd

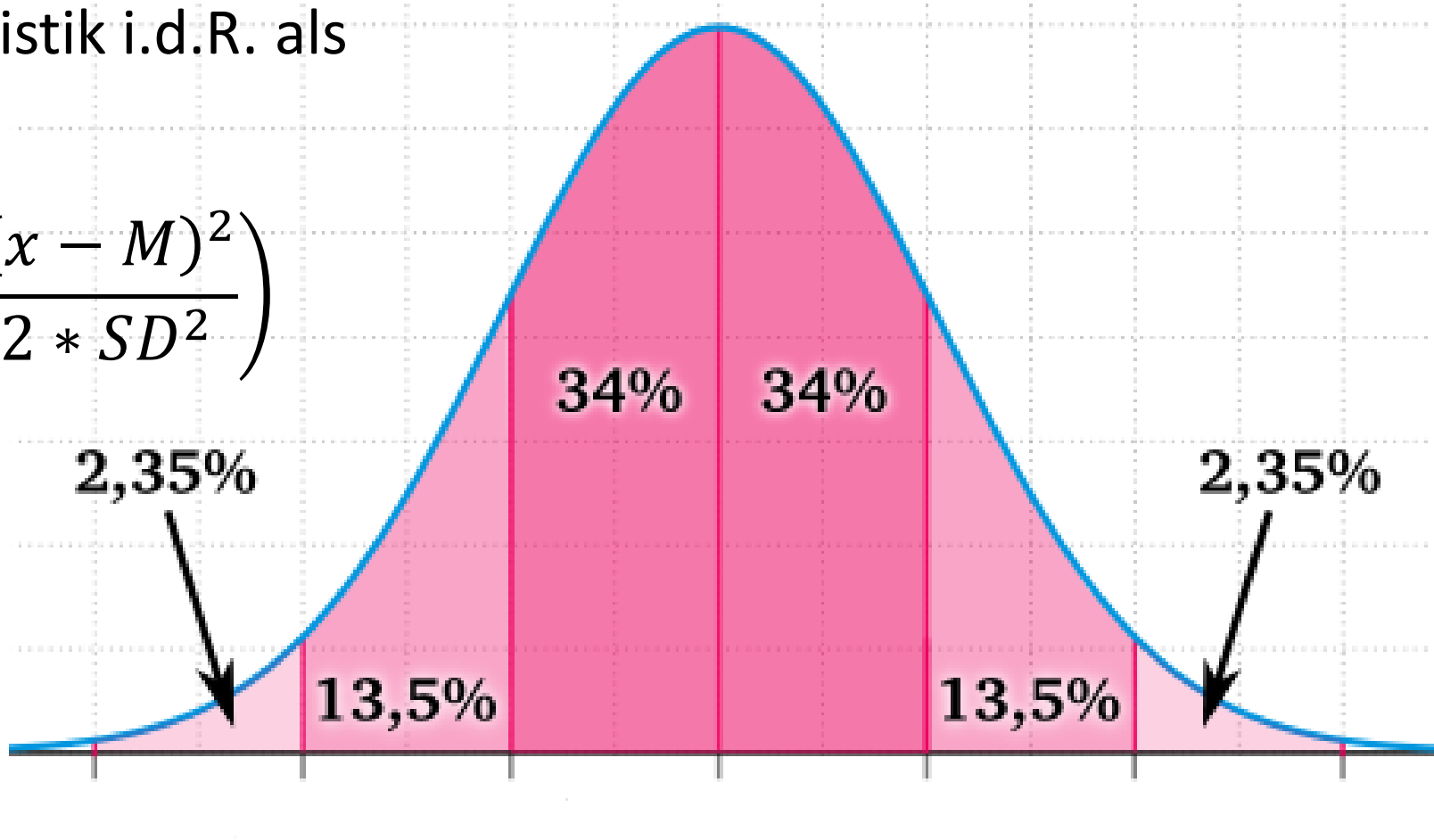
Normalverteilung f(x)=N

- pdf bzw. cdf sind in der Statistik i.d.R. als Formel vorgegeben, z.B

$$f(x) = \mathrm{N}(\mathrm{M}, SD)$$
$$= \frac{1}{\sqrt{2\pi * SD^2}} \exp\left(-\frac{(x-M)^2}{2 * SD^2}\right)$$

- Ist die Gaußverteilung bzw. Normalverteilung

- Es gilt $f(x) \geq 0 \; \forall x \in \mathbb{R}$



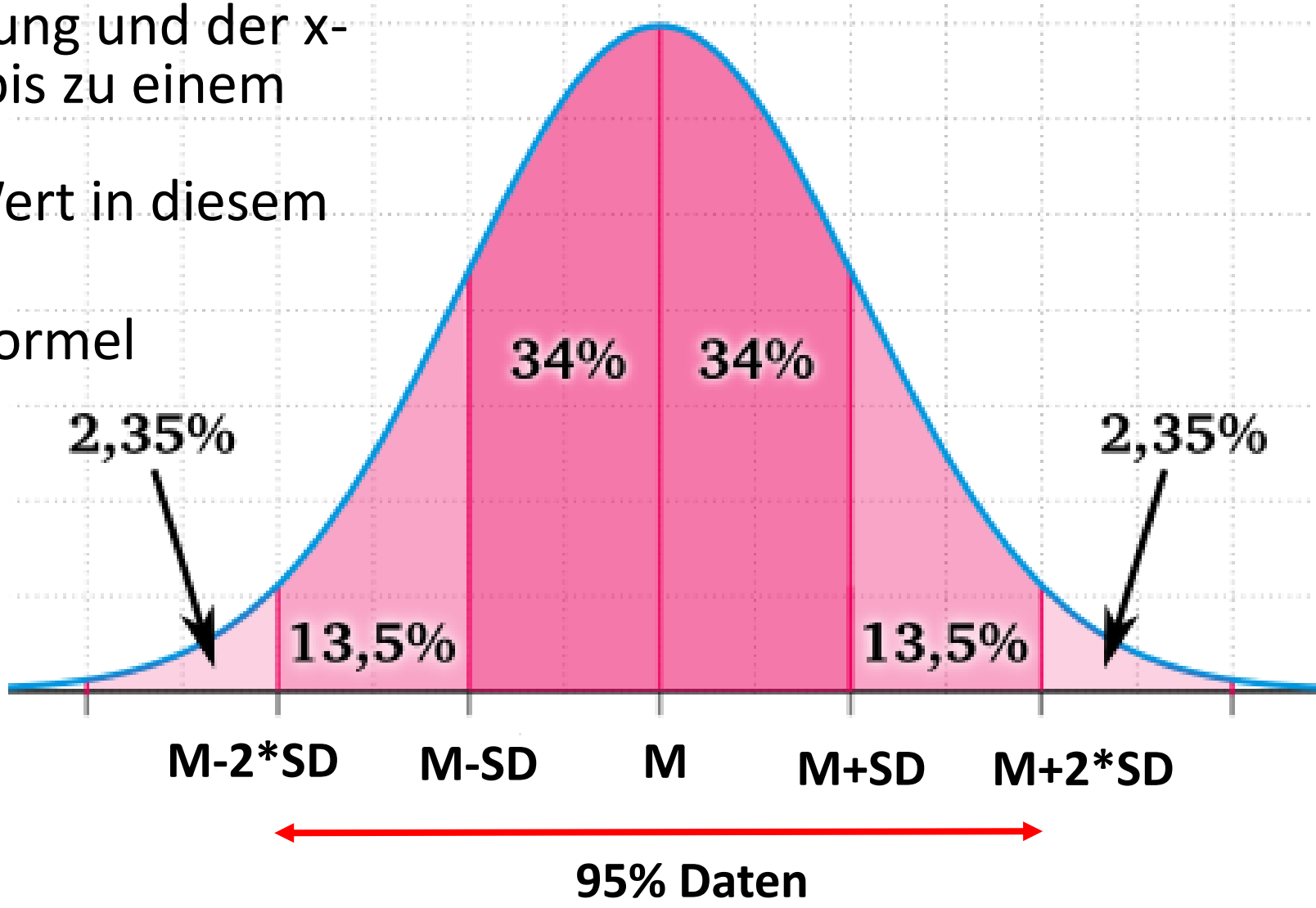
Normalverteilung f(x)=N

- Fläche zwischen der Verteilung und der x-Achse von einem Punkt M bis zu einem Punkt +SD entspricht der Wahrscheinlichkeit einen Wert in diesem Bereich zu erhalten (34%)
- Nota die cdf_N ist nicht als Formel darstellbar

i.d.R.:

$$(i) \int f(x)dx = 1$$

aber nicht f(x)<1 $\forall x \in \mathbb{R}$



Ein Merkmal X_1

- Ein erstes Ziel des Data Mining ist es, sich ein Bild über jedes einzelne Merkmal X_i des vorgelegten Datensatzes zu machen
  - Merkmale werden üblicherweise in Spalten aufgetragen
  - In Zeilen gibt es eine eindeutige Zuordnung zu einem eindeutigen Key der 1. Spalte und jeweils einen „Fall“ pro Zeile
- Vereinfacht nehmen wir in dieser Vorlesung an X_1 sei das Merkmal „waiting“ (2. Spalte)
- Dabei ist im einfachsten Fall der Datensatz durch i=1,..,n Merkmale in Spalten bestimmt

Beispiel in R

> data("faithful")

> View(faithful)

Key	waiting	eruptions
1	79	3.600
2	54	1.800
3	74	3.333
4	62	2.283
5	85	4.533
6	55	2.883
7	88	4.700
8	85	3.600
9	51	1.950
10	85	4.350
11	54	1.833
12	84	3.917
13	78	4.200
14	47	1.750
15	83	4.700
16	52	2.167
17	62	1.750
18	84	4.800
19	52	1.600

Ein Merkmal X_1

- Ein erstes Ziel des Data Mining ist es, sich ein Bild über jedes einzelne Merkmal X_i des vorgelegten Datensatzes zu machen:
  - In welchem Wertebereich die Daten liegen
  - Wie deren „Verteilung“ ist, die für Data Mining interessant sind
  - Identifikation von Ausreißer, Lage- und Streumaßen
  - Kandidaten für Transformationen zum Zwecke der Normierung

Key	waiting	eruptions
1	79	3.600
2	54	1.800
3	74	3.333
4	62	2.283
5	85	4.533
6	55	2.883
7	88	4.700
8	85	3.600
9	51	1.950
10	85	4.350
11	54	1.833
12	84	3.917
13	78	4.200
14	47	1.750
15	83	4.700
16	52	2.167
17	62	1.750
18	84	4.800
19	52	1.600

Ein Merkmal X_1

- Ein erstes Ziel des Data Mining ist es, sich ein Bild über jedes einzelne Merkmal des vorgelegten Datensatzes zu machen
  - In welchem Wertebereich die Daten liegen
  - Wie deren „Verteilung“ ist, die für Data Mining interessant sind
  - Identifikation von Ausreißer, Lage- und Streumaßen
  - Kandidaten für Transformationen zum Zwecke der Normierung

Key	waiting	eruptions
1	79	3.600
2	54	1.800
3	74	3.333
4	62	2.283
5	85	4.533
6	55	2.883
7	88	4.700
8	85	3.600
9	51	1.950
10	85	4.350
11	54	1.833
12	84	3.917
13	78	4.200
14	47	1.750
15	83	4.700
16	52	2.167
17	62	1.750
18	84	4.800
19	52	1.600

Für Empirische Daten im Data Mining gilt

- Beim Data Mining sind die Verteilungen der Merkmale i.d.R. nicht bekannt.
- Eine erste Aufgabe ist es daher, eine Hypothese über die Verteilung der Daten zu entwickeln
- Hierzu ist die Beurteilung eines angetroffenen Resultates einer Verteilungsschätzung wichtig.
  - Dies geschieht am Besten indem eine bildliche Darstellung der Verteilung so erzeugt wird, dass sie von einem Menschen beurteilt werden kann
- Dieses Prinzip der Visualisierung von Sachverhalten wird als wichtige Methode im Folgenden immer wieder eingesetzt werden

Empirische kumulative Verteilungsfunktion (ecdf)

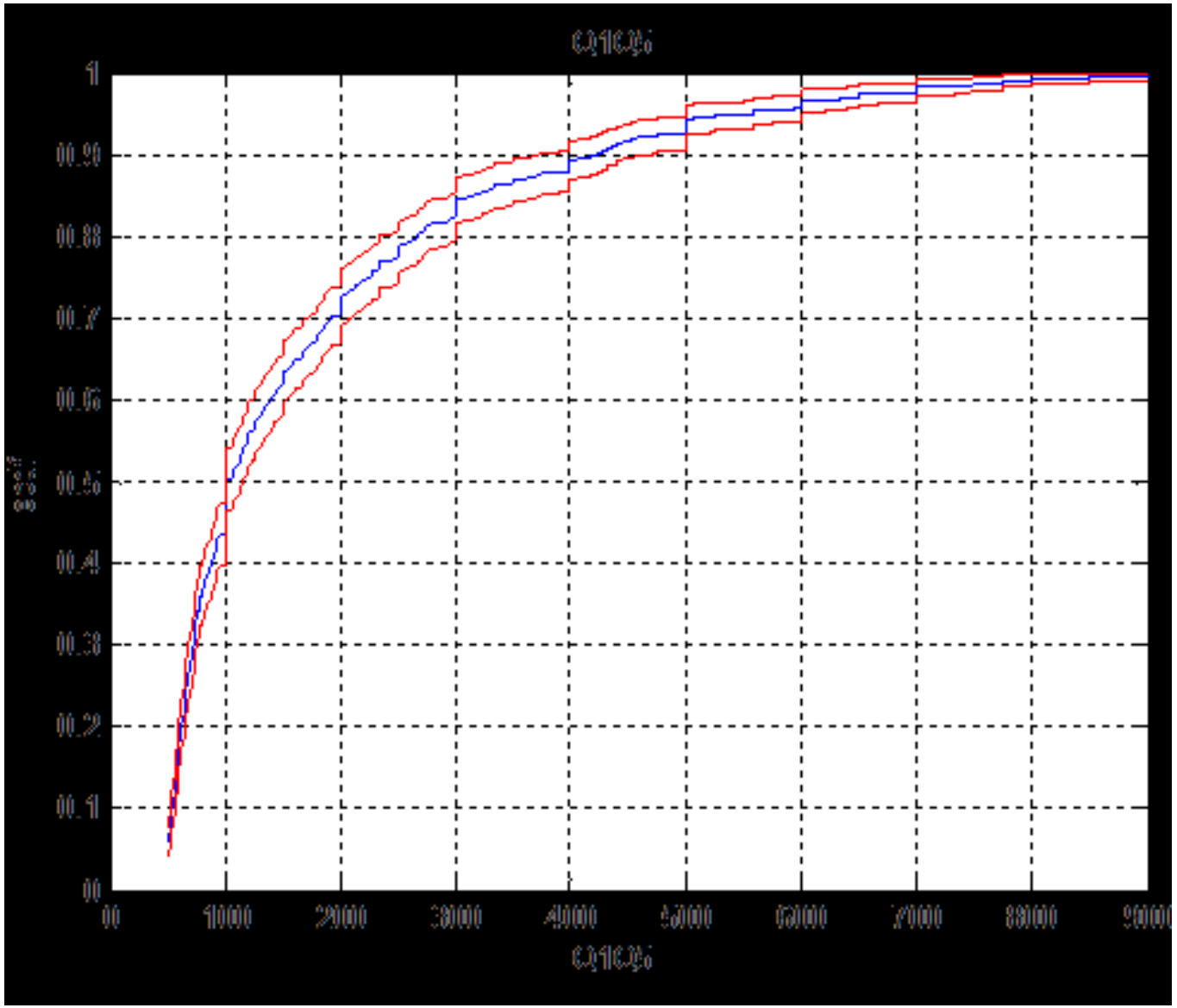
- Als empirische cdf kann die Funktion berechnet und visualisiert werden über

$$\text{ecdf}(x) = \frac{\#\,\boldsymbol{Beobachtungen \leq x}}{\#\,\boldsymbol{aller\ Beobachtungen}}$$

- Die ecdf liefert eine stückweise konstante Treppenfunktion.
- Beispiel in Rmarkdown: 021SchätzungVonVerteilungen_ECDF.Rmd

Irrtumswahrscheinlichkeit der ECDF

- Da die Berechnung der cdf empirisch erfolgt, kann sie mit Fehlern behaftet sein
  - Wenn man sich eine „Irrtumswahrscheinlichkeit“ alpha vorgibt, z.B. alpha = 5%, so kann man mit statistischen Methoden Grenzen finden innerhalb derer die tatsächliche cdf mit der Wahrscheinlichkeit 1-alpha (im Beispiel 95%) zu finden ist



pdf: Verteilungsdichte

- Um eine Vorstellung von der Verteilungsdichte der Datenmenge eines Merkmales zu bekommen muss für jeden Wert x bestimmt werden:

$$pdf(x) = \lim_{r \to 0} |d \in [x - r, x + r]|$$

- Es muss also gezählt werden, wie viele Datenpunkte im Intervall [x-r, x+r] liegen
- Dieses Intervall wird Parzen-Fenster, r der Radius des Parzenfensters genannt
- Für gegebene Daten wird die pdf nur an endlich vielen Punkten $x_1,\ldots x_n$ bestimmt

=> Wahl von r ist kritisch

Histogramme

- Histogramme schätzen die pdf in einer Folge von sich berührenden Intervallen, den sog. Bins.
- Anzahl der Daten pro Bin bzw. die Häufigkeiten des Auftretens in den Bins wird ermittelt.
- Die so ermittelten Werte werden als aneinander stoßende Rechtecke, deren Flächeninhalt proportional zur Anzahl der Daten (Häufigkeiten) in den jeweiligen Bins ist, dargestellt.
- In der Regel werden dabei gleichgroße Intervalllängen angenommen, so dass die Häufigkeit der Daten in den Bins direkt an der Höhe der Rechtecke abgelesen werden kann.

Histogramm Beispiel mit 2 Binbreiten

- Die Wahl eines geeigneten Radius und der richtigen Bingrenzen ist dabei kritisch
- Ungeeignete Wahl der Binparameter ein falscher Eindruck von der Verteilung der Daten entstehen

- Beispiele in 022SchätzungVonVerteilungen_Histogram.Rmd

Kerneldichteschätzer

- Histogramme sind sog. „Kerndichteschätzer“ mit festem, üblicherweise nicht überlappendem Radius r
- Für jedes Bin eines Histogramms wird die Anzahl der Daten gezählt, die in dieses Bin fallen.
- Es gibt viele alternative Kerneldichteschätzer und Schätzverfahren.

Kerneldichteschätzer mit Überlappendem Radius

- Im Datamining können viele interessante Grundeigenschaften durch die Pareto Density Estimation (PDE), vorgeschlagen von Ultsch 2003,2005, gut erkannt werden [Thrun et al., 2020, PLOS ONE].
- PDE ist eine Schätzung der Wahrscheinlichkeitsdichte (pdf) mit einem Kerndichteschätzer mit überlappendem Radius .

Pareto Density Estimation (PDE)

- Bei den PDE-Plots wird die Datendichte an allen verschiedenen Datenpunkten der Datenmenge abgeschätzt
- Hierzu werden für einen Punkt x alle Datenpunkte y, mit in $d(x,y) < r_p$, der sog. ParetoKugel gezählt.
  - d ist ein Abstand, r_p der so genannte ParetoRadius.
- Pareto Radius rp wird datengetrieben gewählt
  - Details in einer anderen Vorlesung

PDE plot

- Um aus der Dichteschätzung eine Schätzung der Wahrscheinlichkeitsdichte (probability density estimation PDE) zu erhalten wird die Fläche unter der Kurve mit der Trapezmethode bestimmt
- Mit dieser Flächenschätzung wird die Dichtemessung normiert um eine Wahrscheinlichkeitsdichte zu erhalten
- PDE ist insbesondere dazu gedacht, Gruppen in Daten zu identifizieren [Ultsch 2005].

- Beispiele in Rmarkdown: 03DensityEstimation.Rmd

Quantil/Quantil-Plot (QQ-Plot)

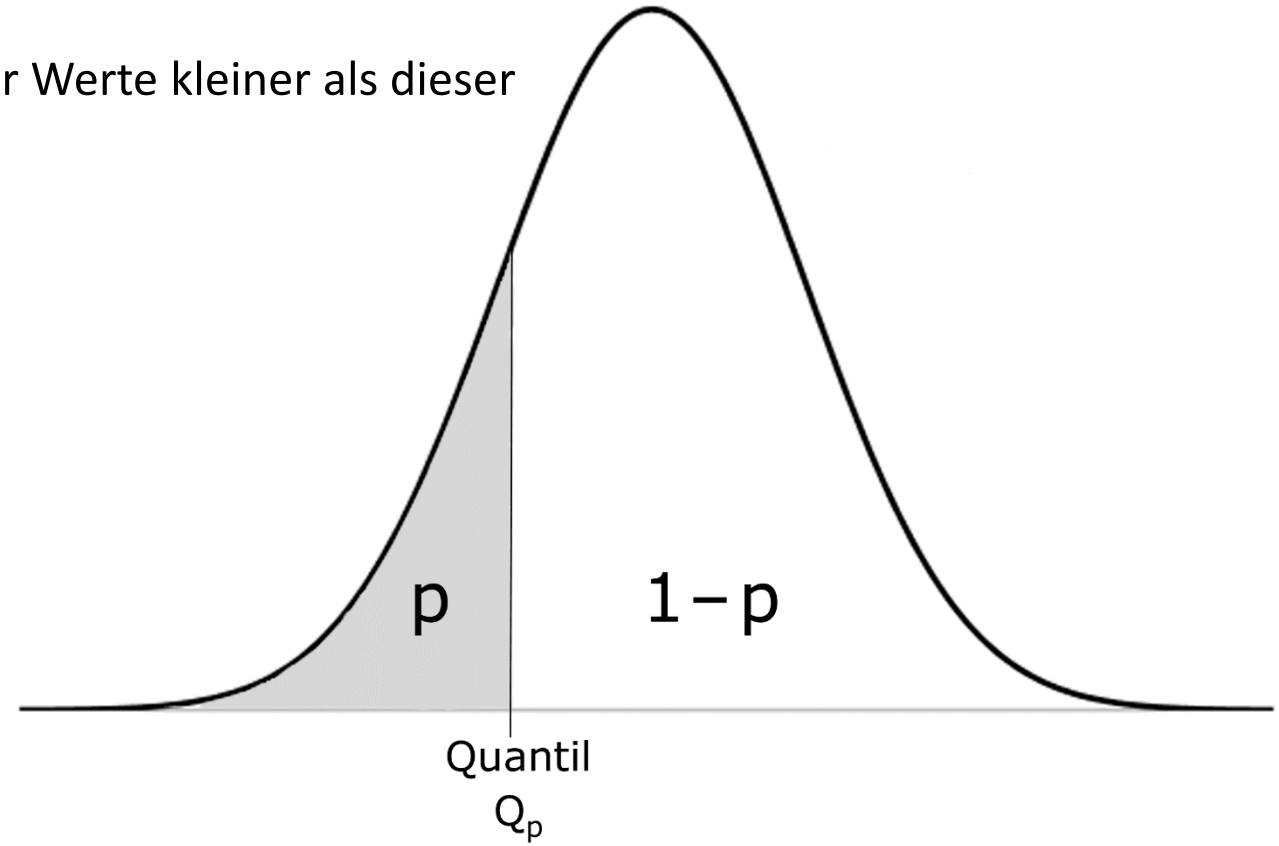
- Dichteschätzer wie Histogramme und PDE-plots sollten jedoch nur als Anhaltspunkt für eine Verteilungsvermutung herangezogen werden.

- Ein fundierteres Bild einer Verteilung liefert ein QQ-plot.
- Dieser erlaubt den auch einen Vergleich mit einer vorgegebenen, bekannten Verteilung.

Quantile

- Das q-te Quantil Q_p ist derjenige Schwellenwert p, der bestimmt dass q% des Datensatzes kleiner als p sind, z.B.
  - Der Wert p für das 25-%-Quantil bedeutet 25 % aller Werte kleiner als dieser Wert p
- Beispiel

```
> data("faithful")
> waiting=faithful[,2]
> range(waiting)
43 96
> quantile(waiting,probs = 0.25, type = 7)
25%
 58
```



⇒25% der Werte sind (geschätzt) kleiner als p=58 in der Variable „waiting“

- Achtung: Schätzungsansatz hängt von Parameter „type“ ab (9 Optionen alleine in R…)

QQ-Plot

- Mit Quantil/Quantil-Plots oder kurz QQ-Plots können zwei Verteilungen mit einander verglichen werden
- Hierzu werden die Quantile der beiden Verteilungen in einem Koordinatensystem gegeneinander aufgetragen.
  - Meistens 100 Stück in 1% Abständen
- Bilden die so entstandenen Punkte annähernd eine Gerade, so kann davon ausgegangen werden, dass die beiden Verteilungen gleich sind
- Beispiele in Rmarkdown: 04QQplot.Rmd

Grundlegende Verteilungstypen

- Gleichverteilung
- Normalverteilung (Gaußverteilung)
- Schiefe Verteilungen
- Log-Normalverteilungen
- Cauchy Verteilung
- Chi-Quadrat Verteilung
- Multimodale Verteilung
- ….

Zusammenfassung I: Wahrscheinlichkeitsdichtefunktion PDF

- Die Verteilung (PDF) eines Merkmals x beschreibt dieses Merkmal eindeutig
  - Fläche zwischen der Verteilung und der x-Achse von einem Punkt a bis zu einem Punkt b entspricht der Wahrscheinlichkeit einen Wert zwischen a und b zu erhalten

$$\int_a^b f(x)dx = P([a,b])$$

- In der Statistik ist die Verteilung bekannt und definiert
- Im Data Mining muss die Verteilung erst „entdeckt“ werden, der übergeordnete Bereicht heißt „Knowledge Discovery“

Zusammenfassung II: Data Mining

- Statistik:
  - ⇒Verteilung für Analyse (z.B. für t-test) bekannt oder zumindestens Verteilungsannahme existiert und kann statistisch bei **geeigneten Voraussetzungen** geprüft werden (z.B. Shapiro-Wilk-Test)

- Data Mining: Seien unbekannte Daten generiert, wie sind diese verteilt?
  - ⇒Das Prinzip der Visualisierung von Sachverhalten ist hier enorm wichtig
  - ⇒Über viele Visualisierungen (Indizien) muss eine Vermutung über die zugrundeliegende Verteilung der Daten aufgestellt werden
  - ⇒Indizien die auf das Selbe hindeuten führten zu einer Hypothese über die Verteilung der Daten
  - ⇒Aber: Jedes Indiz fußt auf bestimmten Voranahmen und kann in die Irre führen!

Beispiele für zu treffende Vorannahmen

- Histogram
  - Binbreite, d.h. Breite des Intervalles [x-r, x+r], r ist Radius
  - Binposition
- Kerneldichteschätzung
  - Algorithmus
  - Radius r
  - Weitere mögliche Parameter die auf diversen statistischen Annahmen fußen
- QQ plot
  - Gegen welche Verteilung soll geprüft werden

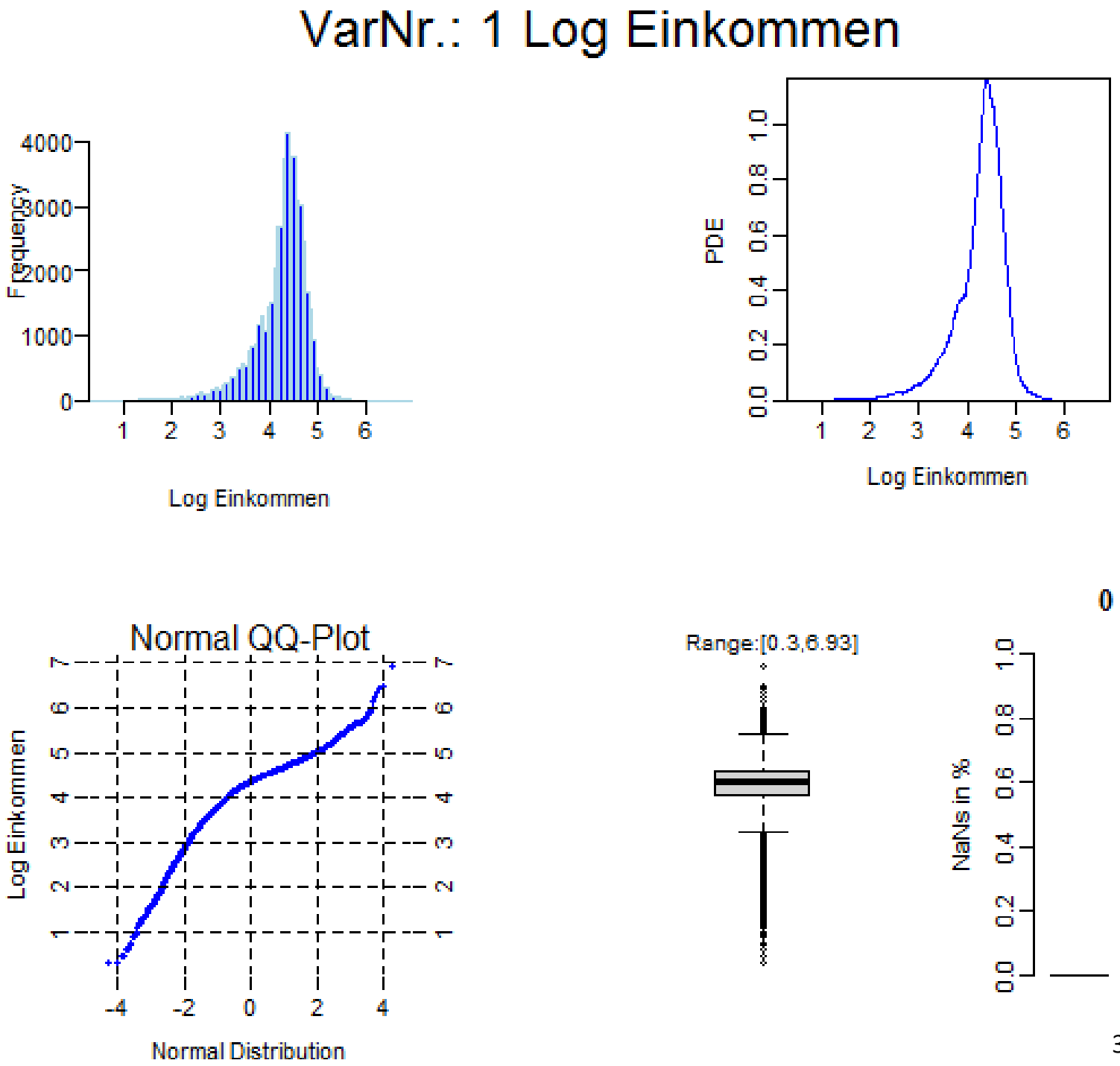
- Wie geht ein Datenwissenschaftlicher nun vor?

- Er kombiniert diverse Verfahren mit wenigen „robusten“ Parametern
- Er ist ein „Detektiv“ und glaubt nicht einfach nur einem Verfahren!

Beispiel:

- Aufruf: DataVisualizations::InspectVariable()

- Rechts Oben: PDF des Bruttoeinkommen deutscher Bürger 2003, im LOG_{10},d.h.



[Thrun M.C.,Ultsch, A, 2015]

Grundlegende Verteilungstypen

- Gleichverteilung
- Normalverteilung (Gaußverteilung)
- Schiefe Verteilungen
- Log-Normalverteilungen
- Chi-Quadrat Verteilung
- Cauchy Verteilung
- …
- Multimodale Verteilung
  - Superposition von Normalverteilungen mit

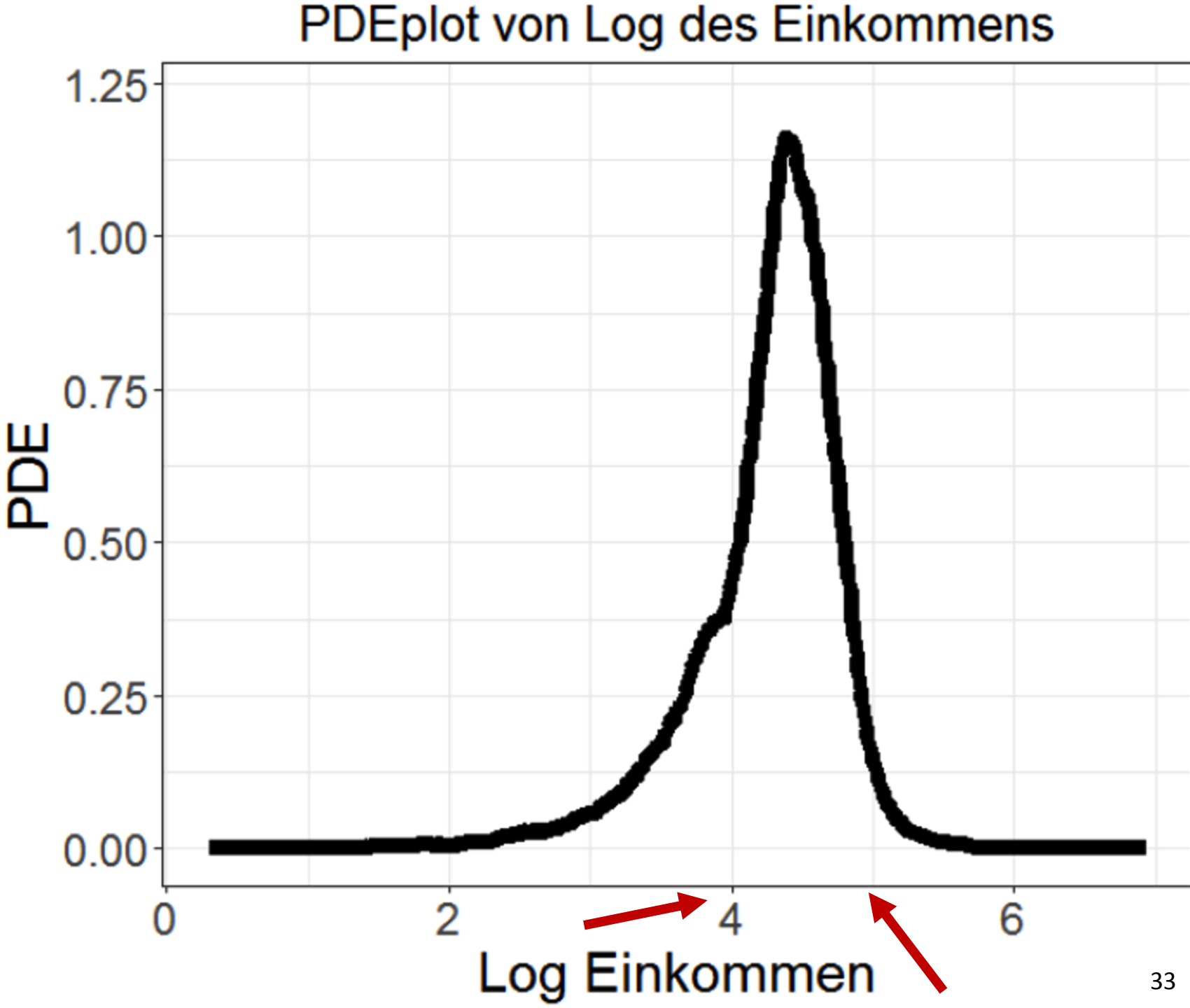
$$\mathrm{GMM}(x) = \sum_{i=1}^{4} w_i * \mathrm{N}(m_i, SD_i)$$

- Model Fitting bei gegebener Verteilungsannahme
- Rechts: PDF des Log des Bruttoeinkommen deutscher Bürger 2003
  - im LOG_{10},d.h.
  - 4 entspricht 10^4=10000
  - 5 entspricht 10^5=100000

- Aufruf:

  DataVisualizations::PDEplot

- Wie modelliert man Dichtezustände innerhalb einer möglicherweise multimodalen Verteilung?



[Thrun M.C.,Ultsch, A, 2015]

Gaussian Mixture Model (GMM)

- Ein Algorithmus schätzt eine Gauß‘sche Mischung aus vier Dichtezuständen (Komponenten)

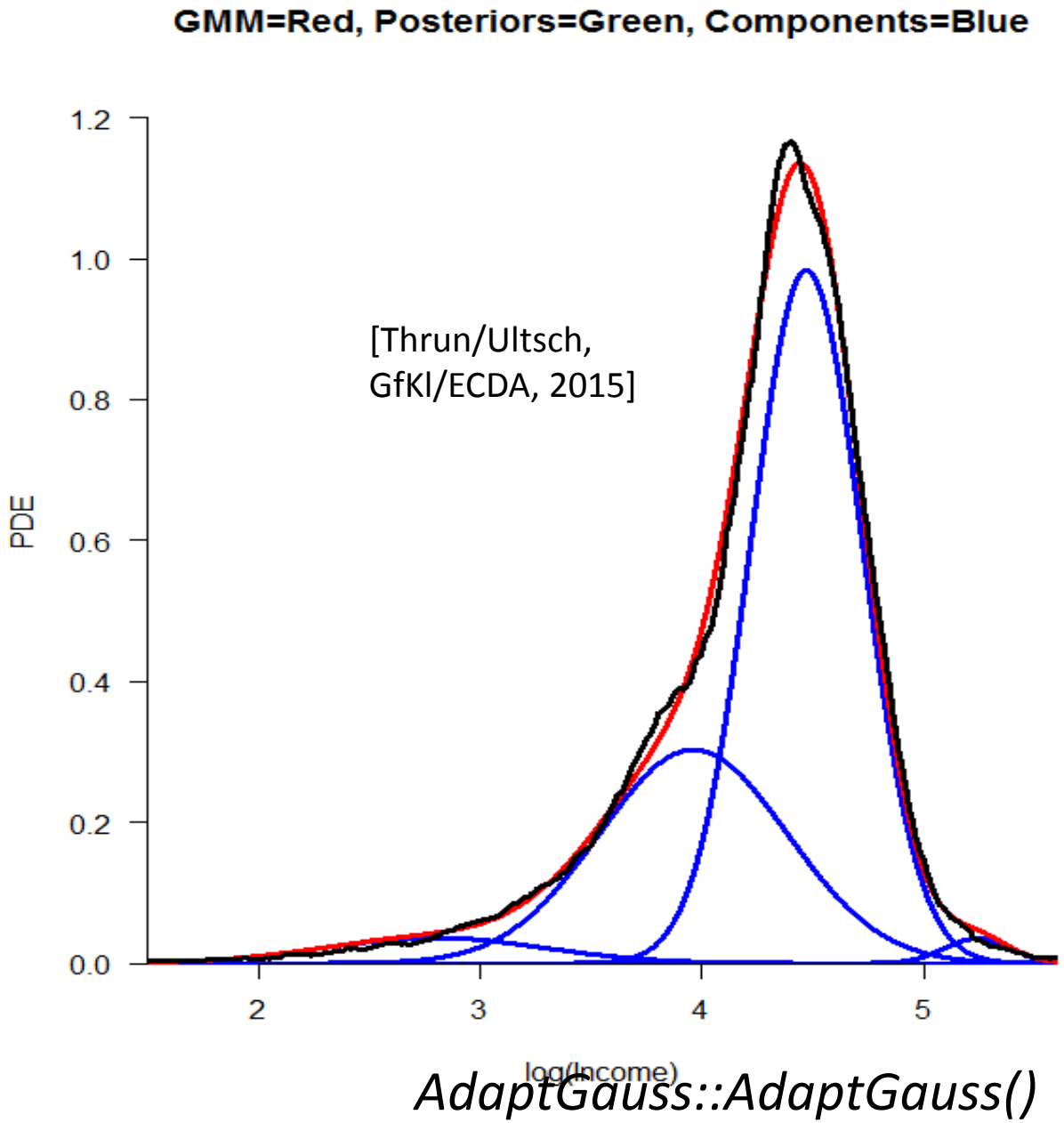
- Blau: Komponenten $\mathrm{N}(m_i, SD_i)$

Rot: $\mathrm{GMM}(x) = \sum_{\mathrm{i}=1}^{4} w_i * \mathrm{N}(m_i, SD_i)$

$$\sum_{i}^{4} w_i = 1$$

$$\int^{i} GMM(x) = 1$$

Wie funktioniert dies?

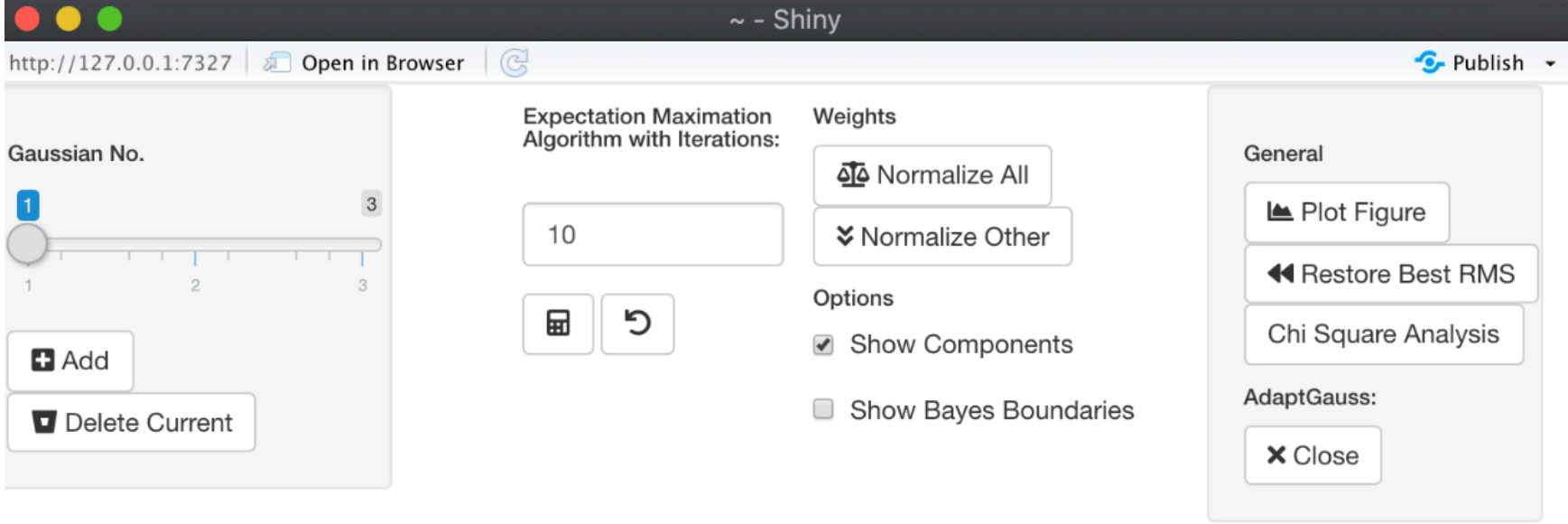


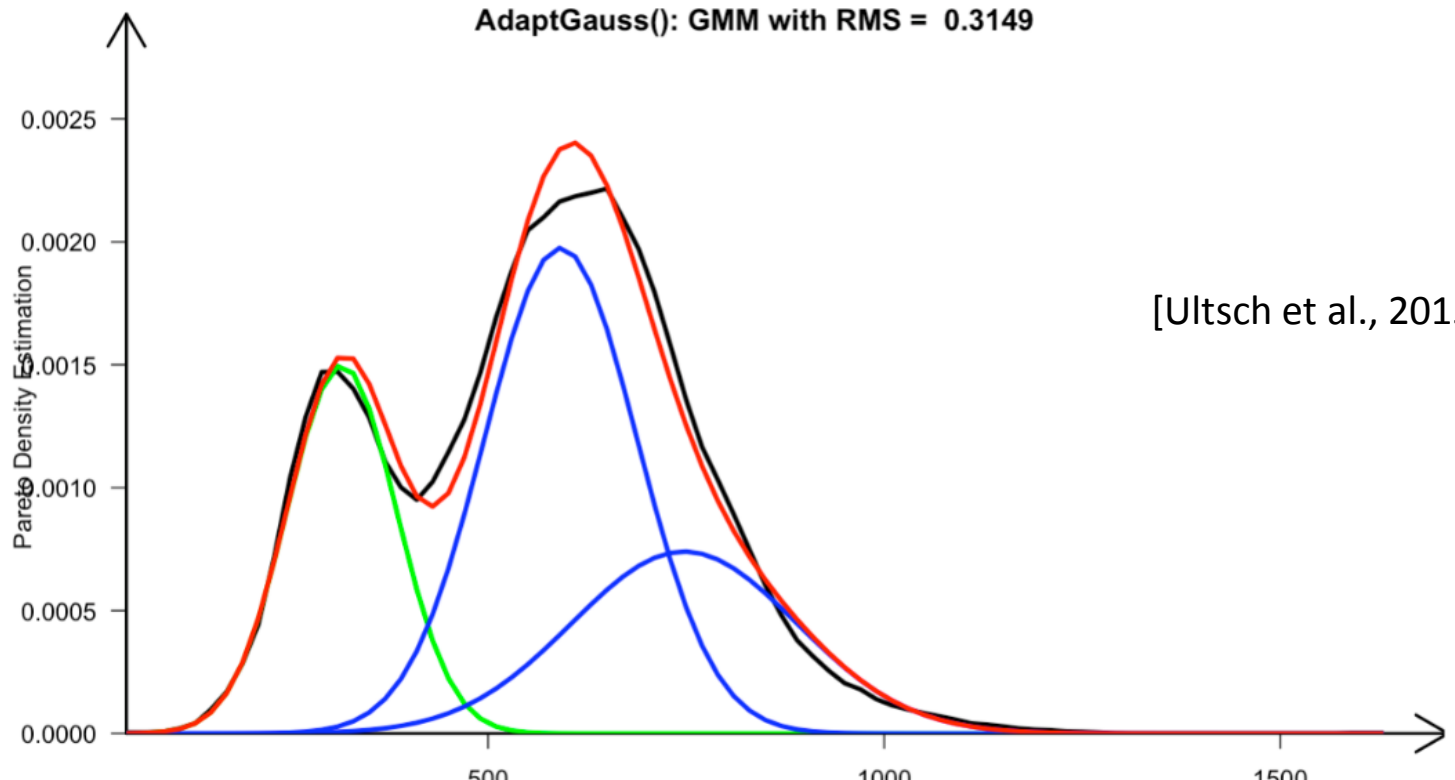
AdaptGauss::AdaptGauss()

Interactive Gaussian Mixture Modelling

- Für die Abschätzungen der Mittelwerte und Standardabweichungen eignet sich der Erwartungs-Maximierung-Algorithmus (EM).

- Beispiele in Rmarkdown: 05ModelleriungEinMerkmal.Rmd

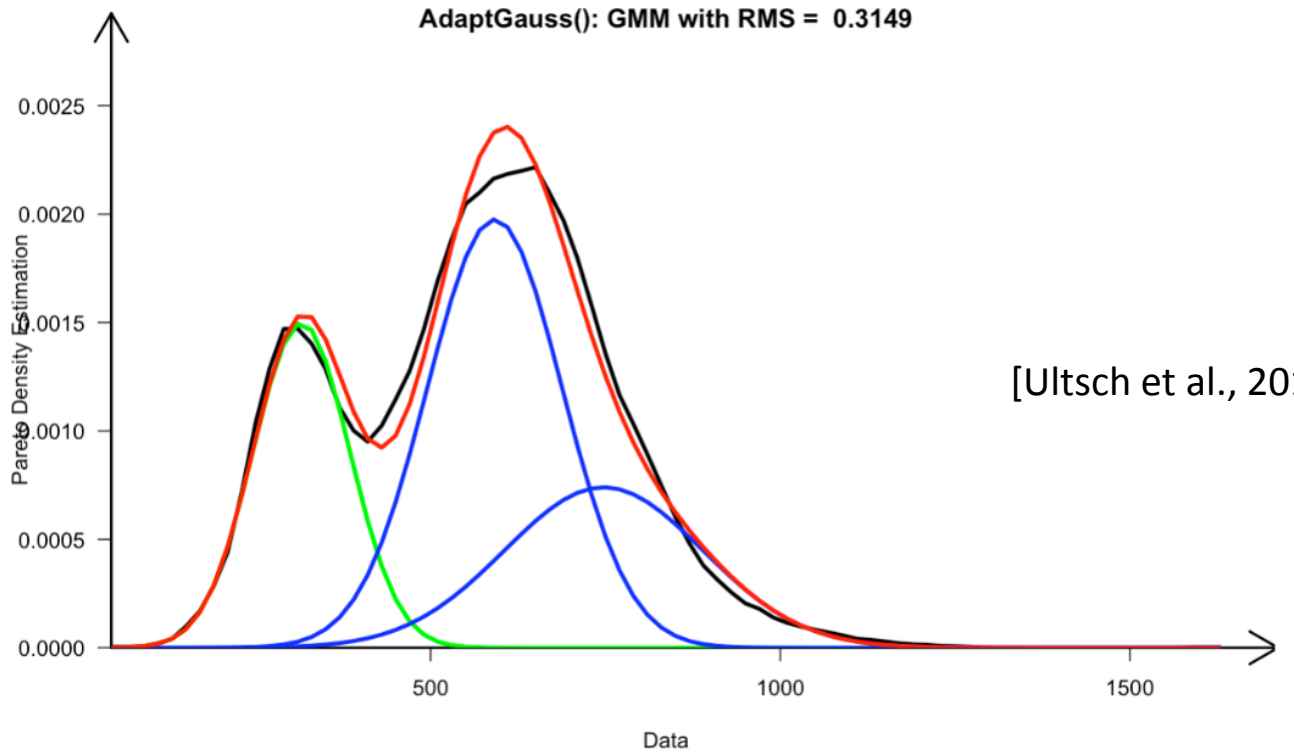




[Ultsch et al., 2015]

Interactive Gaußian Mixture Modelling

- Der EM-Algorithmus sucht ein lokales Maximum dreier Parameter jeder Mode. Er benötigt eine vorgegebene Anzahl an Moden sowie die jeweiligenen
  - Mittelwerte
  - Standardabweichungen
  - Gewichtungen

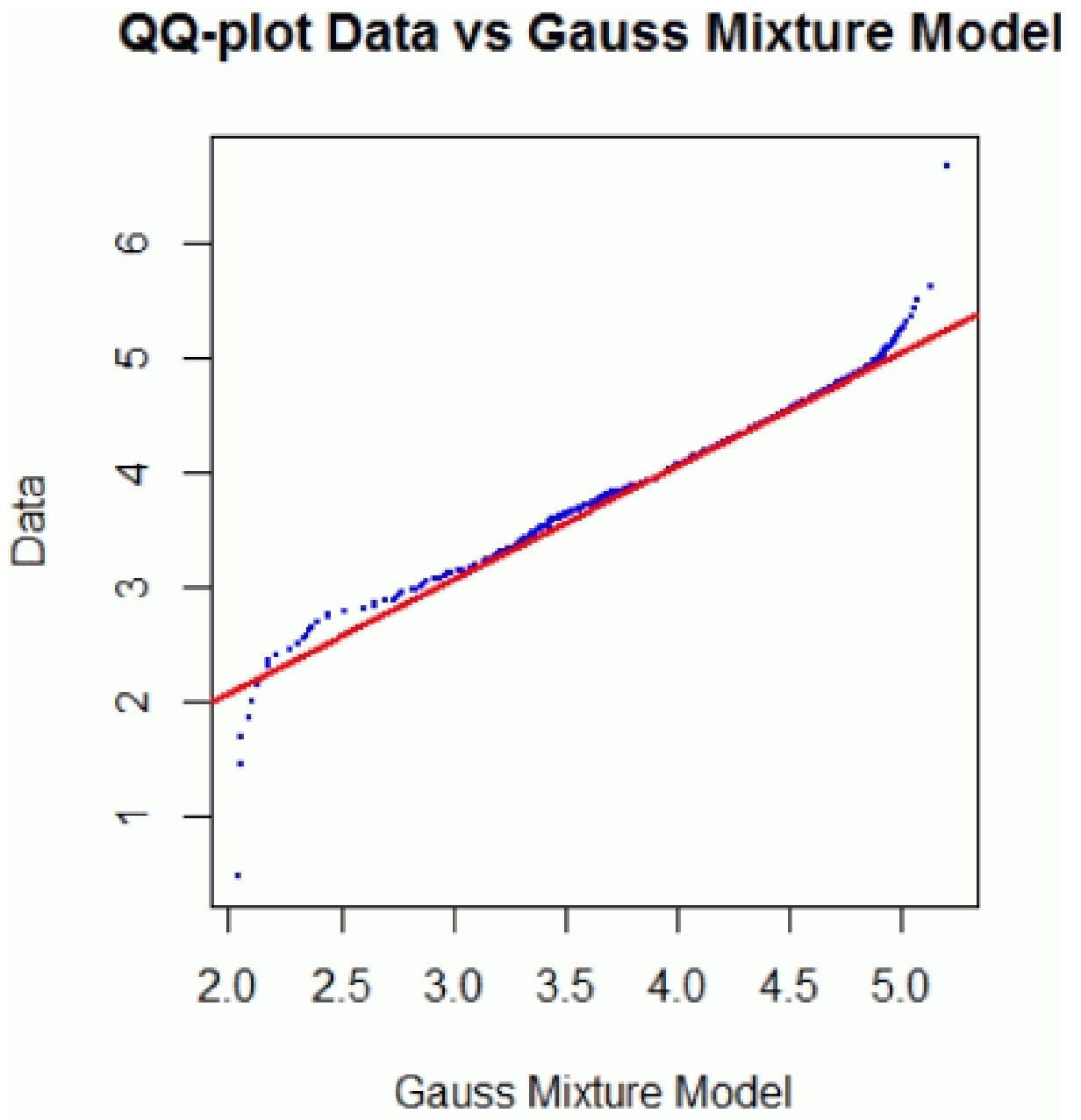


[Ultsch et al., 2015]

Ist der Modelfit gut?

- Statistische Tests:
  - Xi-Quadrat test: p<.001
  - Kolmogorov Smirnov test

- Visuell: QQ plot
  - Vergleicht zwei Verteilungen mit Hilfe von n-Quantilen
  - Empirische Verteilung vs. bekannte Verteilung
  - Wenn gerade Linie: Verteilungen gleich



Zusammenfassung III: GMM

- Mehrere Moden sind ein Hinweis auf eine mögliche Gruppenbildung der Daten.
- Sollten Moden in Daten vorher erkennbar sein oder nach einer Transformation erkennbar werden, ist es möglich Gruppen zu definieren.
  - In einer Variablen, welche nicht normalverteilt ist, ist dies mit leichtverständlichen Ansätzen nur heuristisch möglich.
  - Bei multimodal normal verteilten Variablen wird das Gaußmixturen Model (*GMM*) verwendet.
- Ausblick: Über das Bayes Theorem können empirisch Grenzen zwischen den Moden berechnet und somit den Daten Klassen zugeordnet werden

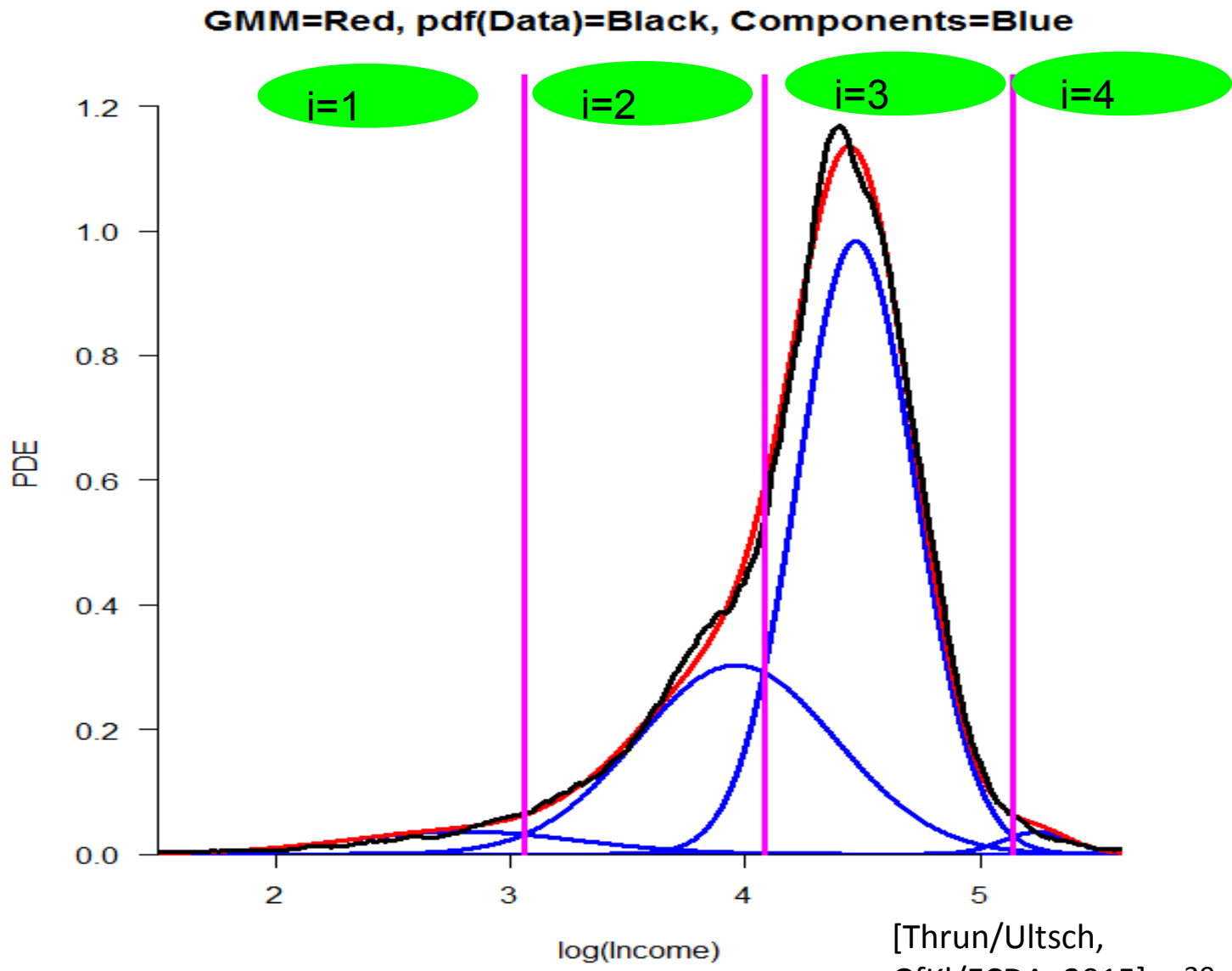
Ausblick: Klassifizierung durch Anwendung des Bayes Theoremes

Schwarz= pdf(log(Data))

Magenta=Bayes Boundaries

Rot=GMM

Blau=Komponenten bzw. Moden



Wertebereich:

1. Gruppe: 0-1100 Euro
2. Gruppe: 1100-12000 Euro
3. Gruppe: 12000 -139000 Euro
4. Gruppe: > 139000 Euro

[Thrun/Ultsch, GfKl/ECDA, 2015]

Danke fürs Zuhören, haben Sie Fragen?

Bücher Empfehlungen für Zwischendurch

- Wenig Mathematik
- Aber einige wichtige Konzepte der Data Science werden anschaulich erklärt