Model Fitting: Anwendung Bayes Theorem

Dr. Michael Thrun

thrun@deepbionics.de

Probalistisch-Generativer Ansatz

- Es wird davon ausgegangen, dass die Daten durch einen Prozess erzeugt wurden, welcher mit Wahrscheinlichkeiten beschrieben werden kann
- Die Erzeugung eines Datensatzes wird dabei in zwei Schritten vollzogen

Schritt 1:

- Der Daten erzeugende Prozess befindet sich mit einer gewissen Wahrscheinlichkeit, der sog. á priori Wahrscheinlickeit in einem bestimmten Zustand.
  - Zustände entsprechen de späteren Klassen c_i und haben Wahrscheinlichkeiten $p(c_i)$
  - „Gewichte der Klassen“: $\sum_i p(c_i) = 1$

Schritt 2:

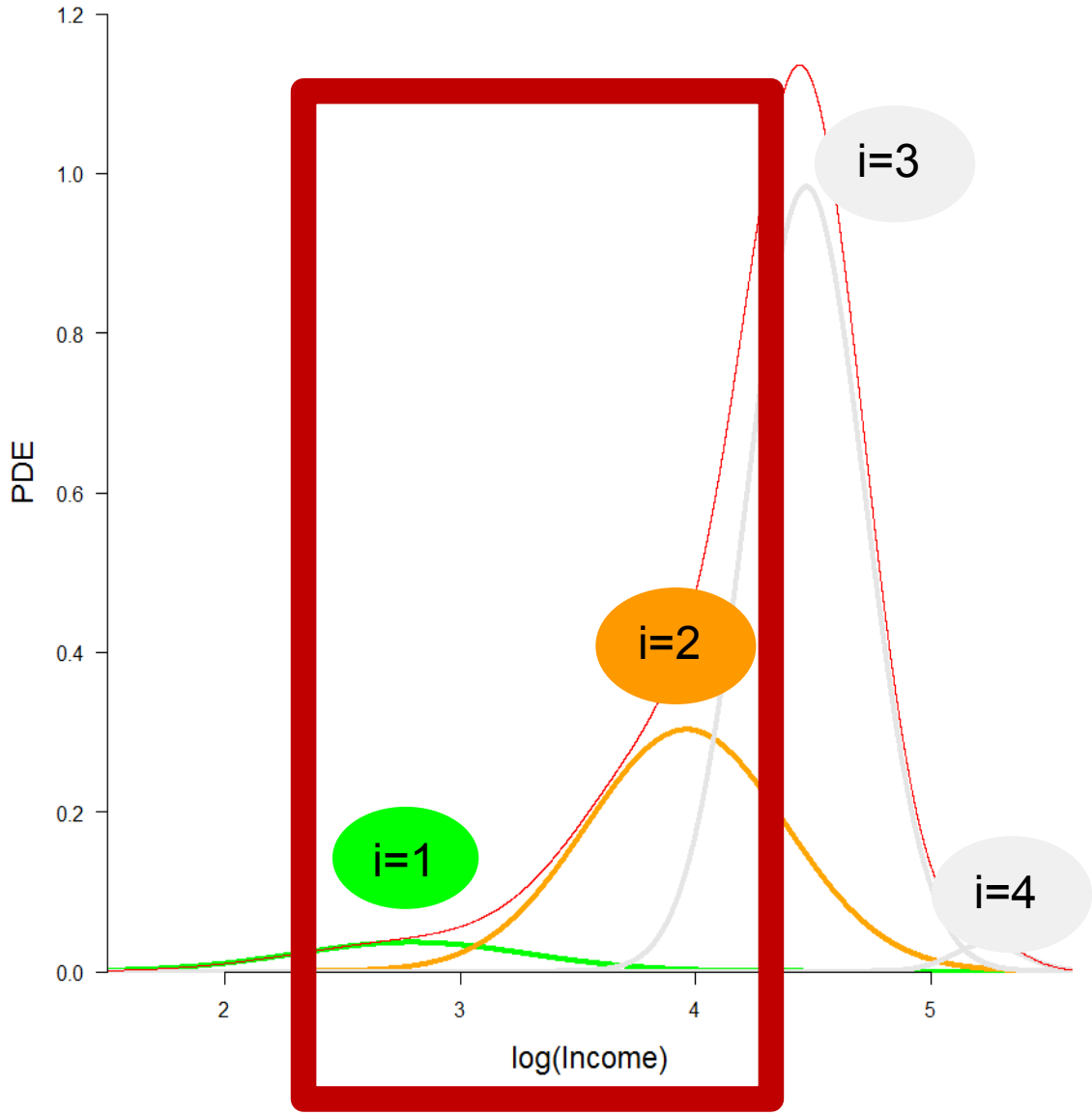
- Wenn eine Klasse c_i gewählt wurde wird im zweiten Schritt der Datenerzeugung ein Datenpunkt gemäß der speziellen Bedingungen der Klasse erzeugt.
- Dies wird mit bedingten Wahrscheinlichkeiten modelliert $\mathrm{p}(\mathrm{x}|\ \mathrm{c_i})$
  - Beschreibt die Wahrscheinlichkeiten p, mit der ein Prozess in einem Zustand $\mathrm{c_i}$ Daten x produziert

Anwendung des Bayes Theorems

- Klassen bedingte Wahrscheinlichkeit
- $p(x|\ c_i)$ definiert das Vorkommen der Daten x in Klasse c_i

- Daraus lassen sich die Posterioris $p(c_i|x)$ bestimmen, also die die Wahrscheinlichkeit der Zugehörigkeit eines Datensatzes x zu einer Klasse

Beispiel: Betrachten wir das rote Fenster mit Komponente c_1 and Komponente c_2



Anwendung Bayes-Theorems

A-Priori:
Wahrscheinlichkeit, sich in dieser Klasse zu befinden

Bedingte Wahrscheinlichkeit:
Wahrscheinlichkeit, Daten in dieser Klasse zu erzeugen

Posterior:

Wahrscheinlichkeit, dass Daten x der Klasse c_i zugehörig sind

$$p(c_i|x) = \frac{p(c) * p(x|c_i)}{\sum_{i=1}^{L} p(c_i) * p(x|c_i)}$$

$$\sum_{i=1}^{L} p(c\_i) = 1$$

$$\sum_{i=1}^{L} p(c\_i \mid x) = 1$$

Normalisierung, entspricht

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

Erste Bayes Entscheidungsgrenze in GMM

Wahrscheinlichkeit, dass Daten x in Klasse c_1 aufgetreten sind ist der Posteriori

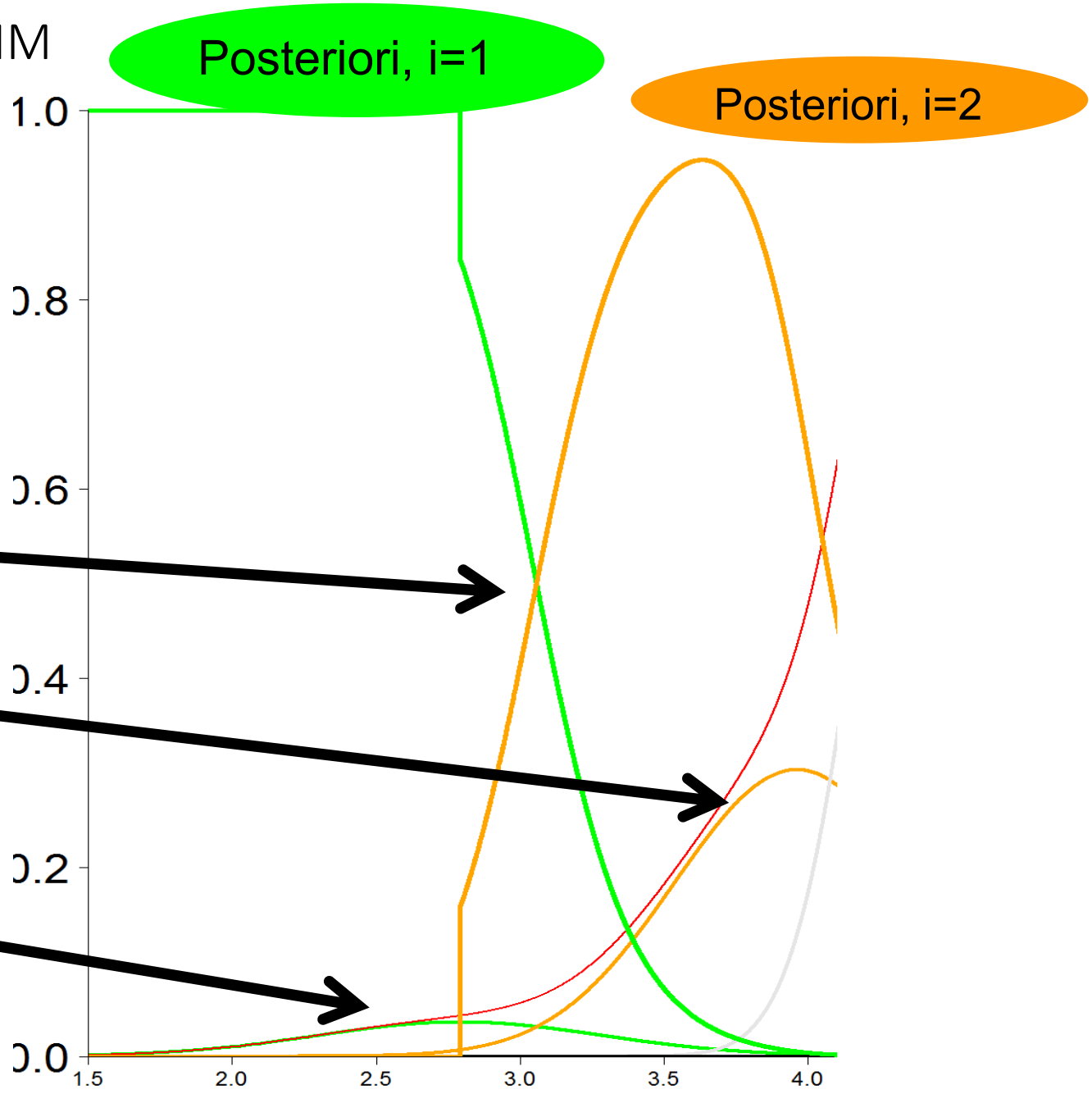
$$p(c_1|x)$$

Posteriori = $p(c_1|x)$=0.5

Mixtur der Komponenten in rot:

Orange: $N(m_2, SD_2)$, c_2

Green: $N(m_1, SD_1)$, c_1



Grenzen durch Verwendung des Bayes-Theorems

- Das Bayes Theorem erlaubt die Bestimmung der Posteriori $p(c_i|x)$
- $p(c_i|x)$ ist die Wahrscheinlichkeit der Zugehörigkeit eines Datensatzes zu einer Klasse.

$$p(c_i|x) = \frac{p(c) * p(x|c_i)}{\sum_{i=1}^{L} p(c_i) * p(x|c_i)} = \frac{\boldsymbol{Priori * bedingte\ Wahrscheinlicheit}}{\boldsymbol{Normalisierung}}$$

- Je nach Anwendung definiert man die Bayes Entscheidungsgrenze bei einem bestimmten Wert von p
  - z.B. bei Gesund vs.Krank möchte man sicher sein, das Kranke wirklich krank sind und damit ist p=0.95 => 95% Wahrscheinlichkeit dass wenn dem Datenpunkt das label krank zugeordnet wird, die Person auch wirklich krank ist

Exakte Entscheidungsgrenze in Magenta

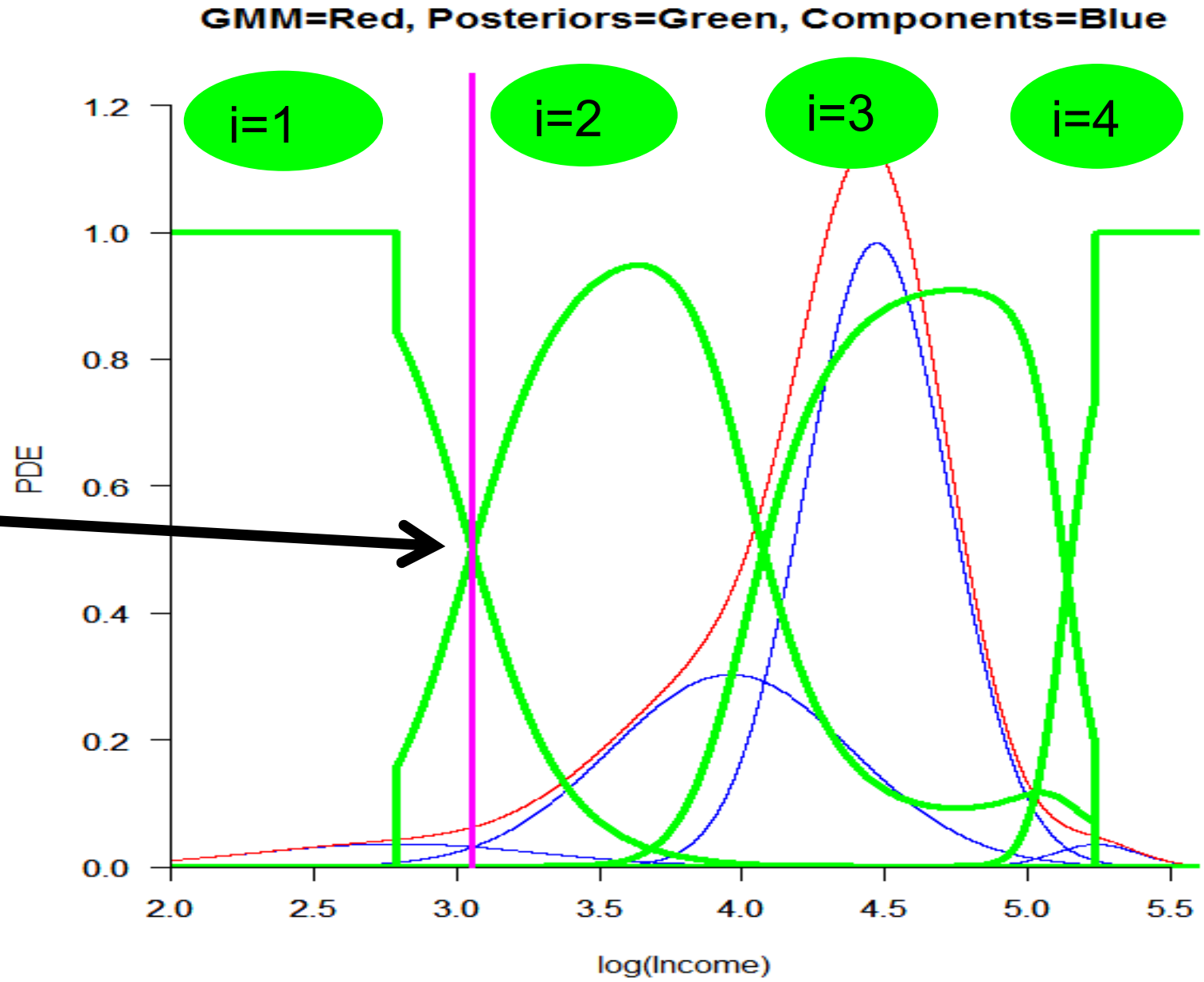
Green: Berechne Posteriori der Gauß-Mixtur der Komponenten

$c_i, i = 1, \dots, 4$

Posteriori = 50%

$\Rightarrow$Bayes Entscheidungsgrenze

zwischen $i = 1$ and $i = 2$ (magenta)



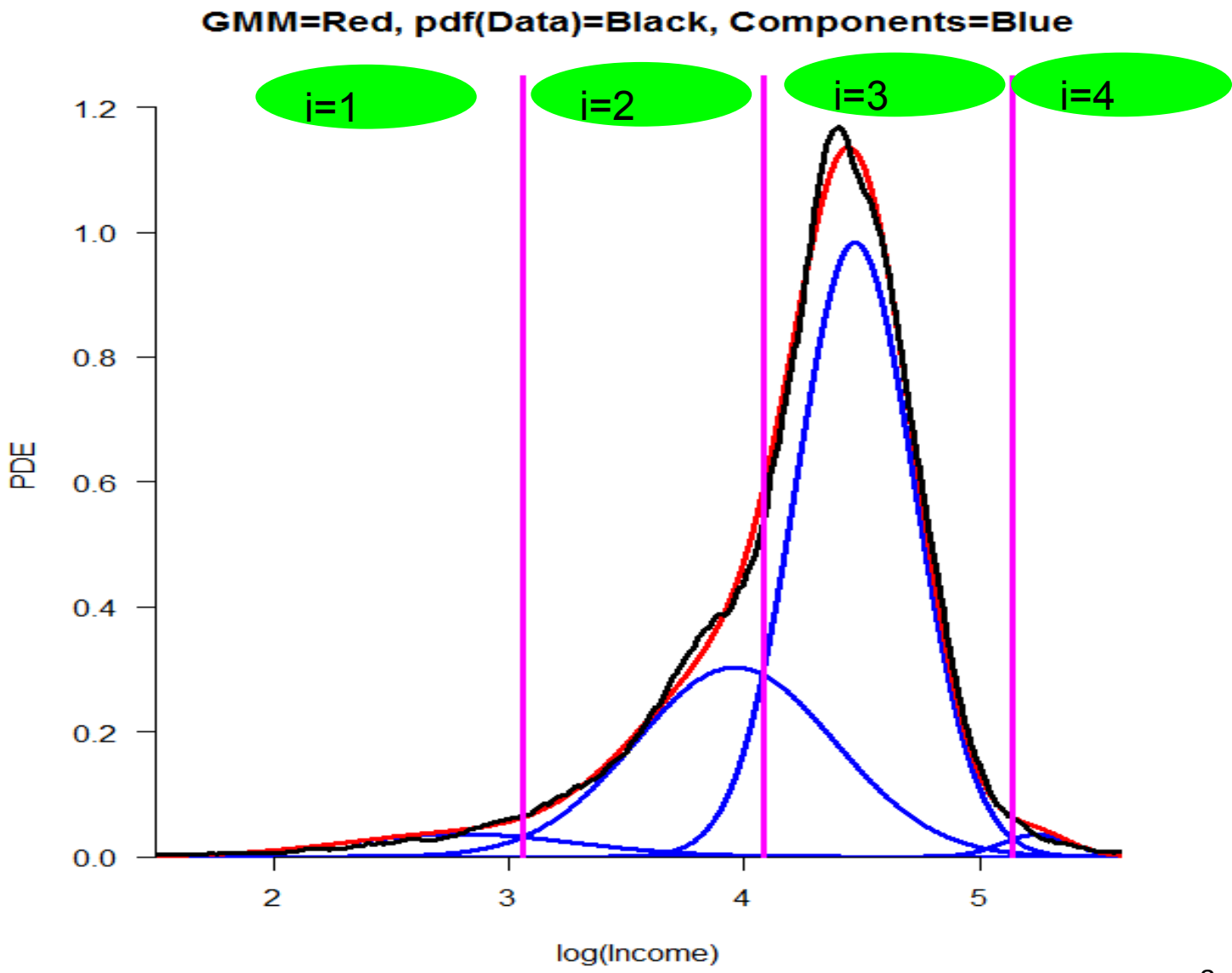
Klassifizierung durch Anwendung des Bayes Theoremes

Schwarz= pdf(log(Data))

Magenta=Bayes Boundaries

Rot=GMM

Blau=Komponenten bzw. Moden



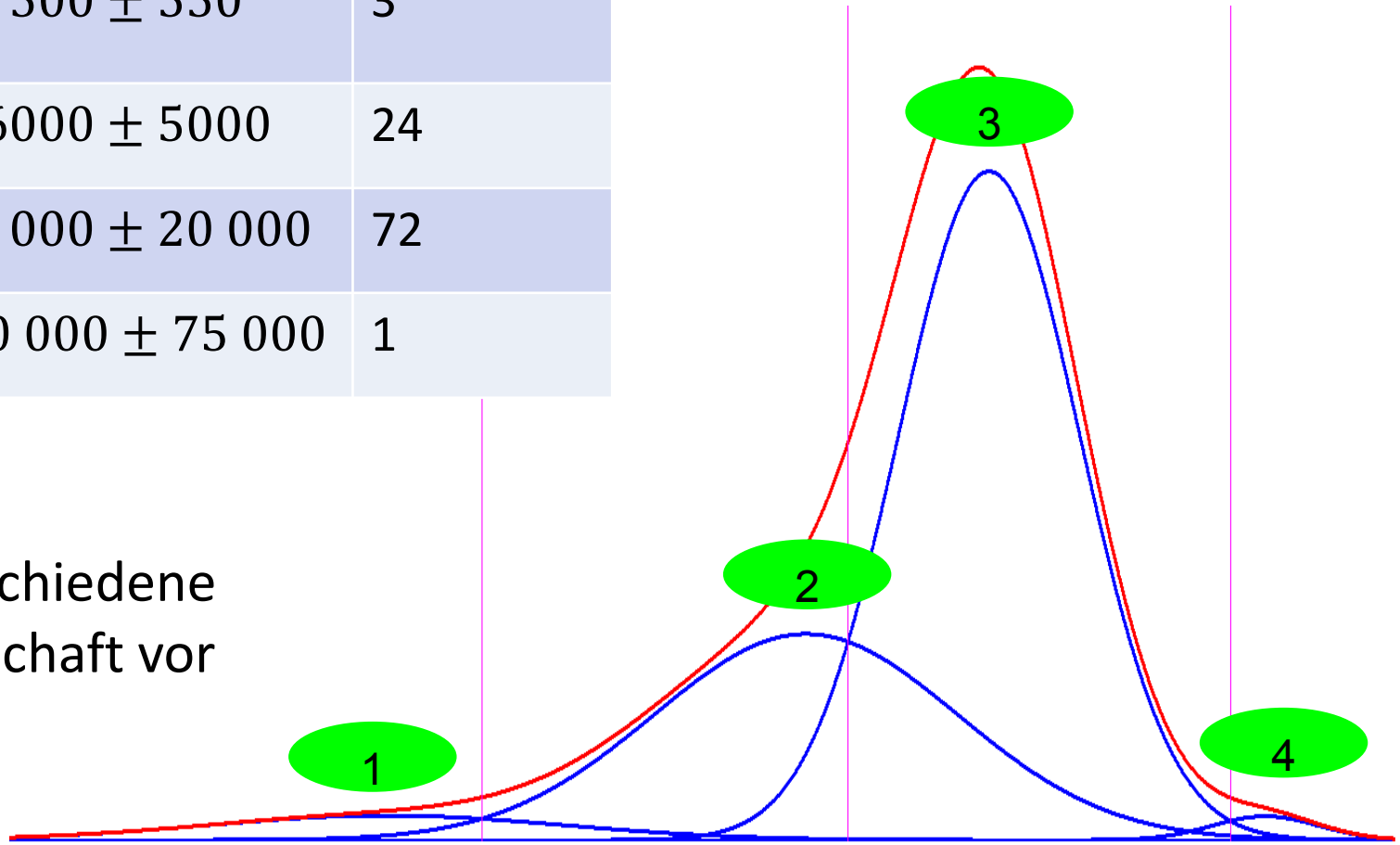
Wertebereich:

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

Knowledge Discovery der Einkommensverteilung

No.	Group	Median ±AMAD in Euro	Population in %
1	Arbeitslos	500 ± 550	3
2	Geringverdiener	6000 ± 5000	24
3	Mittelschicht	$40\,000 \pm 20\,000$	72
4	Oberschicht	$190\,000 \pm 75\,000$	1

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

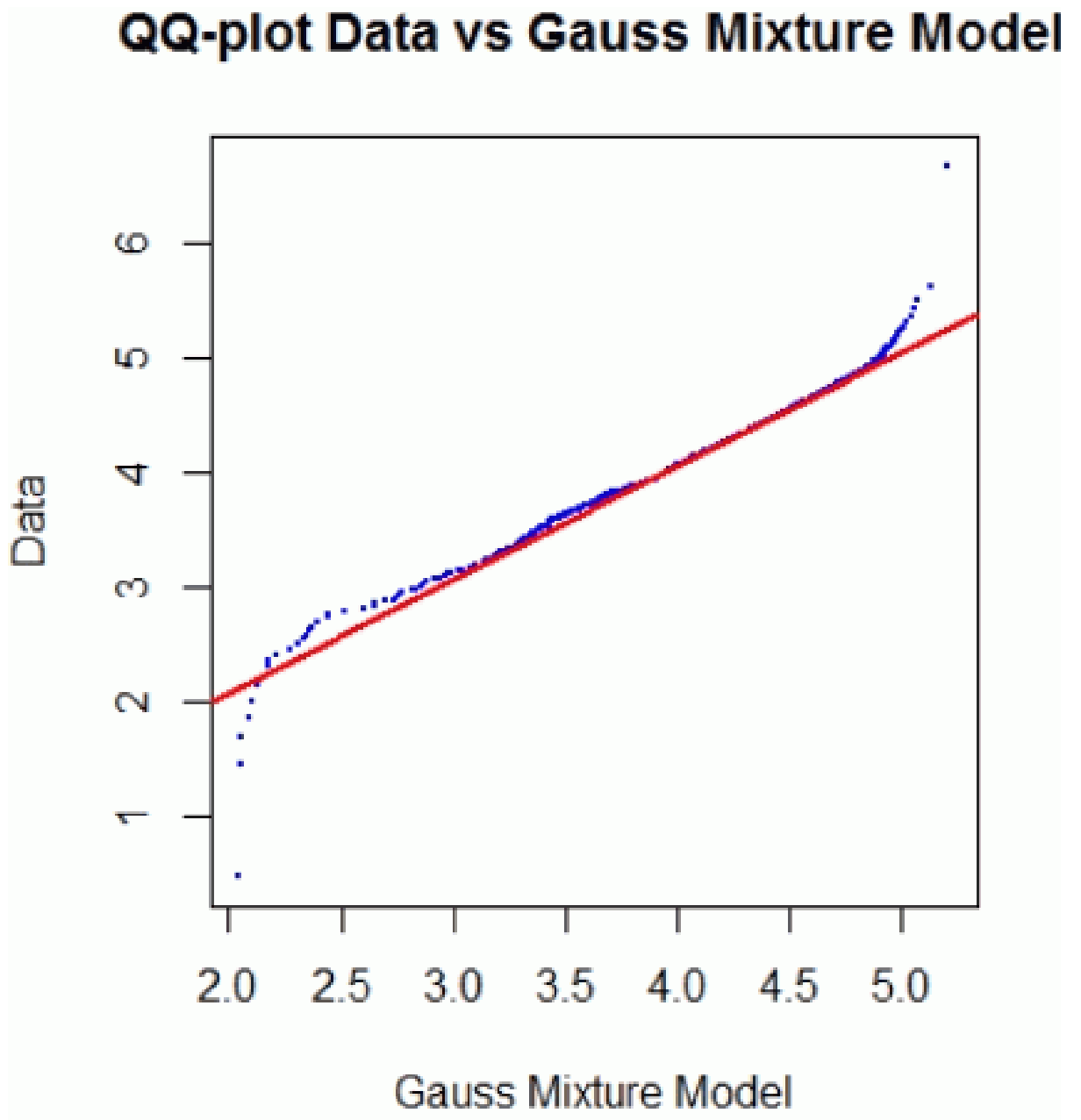


- Model schlägt verschiedene Klassen in der Gesellschaft vor

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 normal verteilten Variablen wird das Gaußmixturen Model (*GMM*) verwendet.
- Über Bayes können empirisch Grenzen zwischen den Moden berechnet und somit den Daten Klassen zugeordnet werden

Danke fürs Zuhören, haben Sie Fragen?

Bücher Empfehlungen für Zwischendurch

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