4. Parameterschätzung

Aufgabenstellung der Parameterschätzung

Gegeben:

- Mathematisches Modell einer WV $p(\mathbf{m}|\boldsymbol{\theta})$ für eine beobachtbare Größe $\mathbf{m}$
- Stichprobe $\mathrm{D} = \{\mathbf{m}_1, ..., \mathbf{m}_N\}$ für festen, aber unbekannten Parameter-vektor $\boldsymbol{\theta}$.

Gesucht:

- „Wahrer“ Wert des Parametervektors $\boldsymbol{\theta}$.

Ansatz:

- Schätzfunktion $\hat{\boldsymbol{\theta}}(\mathrm{D})$ (Schätzer), welche die Stichprobe D statistisch auswertet und in gewissem Sinne dem „wahren“ Wert $\boldsymbol{\theta}$ möglichst nahe kommt.

Methodiken der Parameterschätzung

Likelihoodmethodik (Klassische Statistik)

- θ wird als unbekannte, konstante, nicht-stochastische Größe angesehen.
- Maximum-Likelihood-Schätzung: Man wählt $\hat{\theta}(D)$ so, dass die gegebenen Beobachtungen $D = \{\mathbf{m}_1, ..., \mathbf{m}_N\}$ maximal wahrscheinlich werden.

Bayes'sche Methodik (Bayes'sche Statistik)

- θ wird als Zufallsvariable angesehen und über eine WV beschrieben.
- Z.B.: Maximum-A Posteriori-Schätzung: Man wählt $\hat{\theta}(D)$ als den Wert θ mit der höchsten A Posteriori Wahrscheinlichkeit nach Beobachtung von D.

Axiome nach Kolmogorov:

A1: Nichtnegativität $\Pr(A) \geq 0$

A2: Normierung $\Pr(\text{Sicheres Ereignis}) = 1$

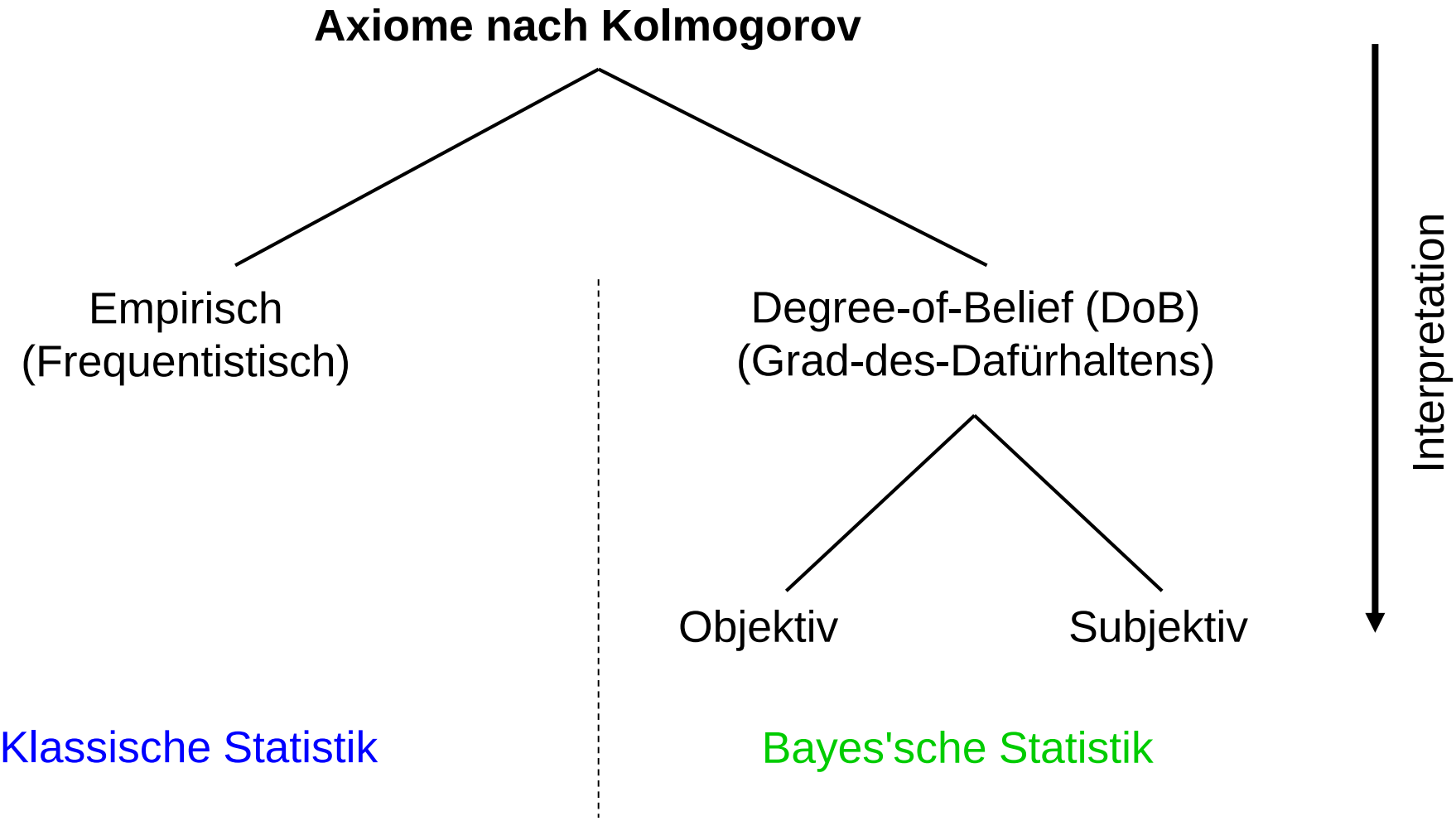
A3: Additivität $A \cap B = \varnothing \quad \Rightarrow \quad \Pr(A \cup B) = \Pr(A) + \Pr(B)$

Bedingte Wahrscheinlichkeit: $\Pr(A|B) := \frac{\Pr(A \cap B)}{\Pr(B)}$

Die Kolmogorovschen Axiome legen fest, wie mit Wahrscheinlichkeiten gerechnet wird (Syntax), machen aber keine Aussage darüber, was Wahrscheinlichkeiten bedeuten (Semantik).

4. Parameterschätzung – Einschub: Bedeutung von Wahrscheinlichkeit

Mit den Axiomen verträgliche Interpretationen bezüglich der Bedeutung von Wahrscheinlichkeit:



Parameterschätzung im Kontext der Mustererkennung

- Schätzung der Parameter der klassenspezifischen WVen $p(\mathbf{m}|\boldsymbol{\theta}_i, \omega_i)\ i = 1,\ldots, c$ aus den Daten D.

Annahmen:

- Struktur der klassenspezifischen WVen ist gegeben. Nur die Parameter $\boldsymbol{\theta}_i$ sind unbekannt.
- Überwachtes Lernen: die Klassenzugehörigkeiten der einzelnen Stichproben in D sind bekannt. → Partitionierung von D in $\mathrm{D}_1,\ldots,\mathrm{D}_c$ mit den zur jeweiligen Klasse gehörenden Stichproben.
- Die Stichproben in D_j haben keinen Einfluss auf $p(\mathbf{m}|\boldsymbol{\theta}_i, \omega_i) \quad i \neq j$.

Konsequenz:

- Schätzung der Parameter der klassenspezifischen WVen zerfällt in c separate Parameterschätzaufgaben, die unabhängig voneinander behandelt werden können. → Klassenunterscheidende Kennzeichnung kann in diesem Kapitel weitgehend unterdrückt werden.

Eigenschaften von Schätzern: Erwartungtreue

Klassische Statistik

Erwartungstreue (unbiased estimator): $\mathrm{E}\{\hat{\boldsymbol{\theta}}\} = \boldsymbol{\theta}$

$$\mathrm{E}\{\hat{\boldsymbol{\theta}}\} = \int_{\Theta} \hat{\boldsymbol{\theta}}\, p(\hat{\boldsymbol{\theta}}) \mathrm{d}\,\hat{\boldsymbol{\theta}} = \int_{\mathbb{M}} \hat{\boldsymbol{\theta}}(\mathbf{m})\, p(\mathbf{m} \mid \boldsymbol{\theta}) \mathrm{d}\,\mathbf{m}$$

Bias (systematischer Fehler): $b(\boldsymbol{\theta}) = \mathrm{E}\{\hat{\boldsymbol{\theta}}\} - \boldsymbol{\theta}$

Bemerkung: Simple Korrektur durch Subtraktion nur möglich, falls b nicht vom Parameter $\boldsymbol{\theta}$ abhängt.

Bayes'sche Statistik

Erwartungstreue (unbiased estimator): $\mathrm{E}\{\hat{\boldsymbol{\theta}}\} = \mathrm{E}\{\boldsymbol{\theta}\}$

$$\mathrm{E}\{\hat{\boldsymbol{\theta}}\} = \int_{\Theta} \hat{\boldsymbol{\theta}}\, p(\hat{\boldsymbol{\theta}}) \mathrm{d}\,\hat{\boldsymbol{\theta}} = \int_{\Theta} \int_{\mathbb{M}} \hat{\boldsymbol{\theta}}(\mathbf{m})\, p(\mathbf{m}, \boldsymbol{\theta}) \mathrm{d}\,\mathbf{m}\, \mathrm{d}\boldsymbol{\theta}$$

Bias: $b = \mathrm{E}\{\hat{\boldsymbol{\theta}}\} - \mathrm{E}\{\boldsymbol{\theta}\}$

Eigenschaften von Schätzern: Varianz

Varianz (stochastischer Fehler): $$\mathrm{Var}\{\hat{\boldsymbol{\theta}}\} = \mathrm{E}\{(\hat{\boldsymbol{\theta}} - \mathrm{E}\{\hat{\boldsymbol{\theta}}\})^2\} = \mathrm{E}\{\hat{\boldsymbol{\theta}}^2\} - \mathrm{E}\{\hat{\boldsymbol{\theta}}\}^2$$

Schätzverfahren mit minimaler Varianz werden als effiziente (wirksame) Schätzverfahren bezeichnet.

Cramer-Rao-Schranke (CRB) für erwartungstreue Schätzer (klassische Statistik) für einen skalaren Parameter:

$\mathrm{D} = \{\mathbf{m}_1, ..., \mathbf{m}_N\}$, Beobachtungen $\mathbf{m}_i$ seien stochastisch unabhängig.

$$\mathrm{Var}\{\hat{\theta}(\mathrm{D})\} \geq \frac{1}{N\,\mathrm{E}\left\{\left(\frac{\partial \ln p(\mathbf{m}|\theta)}{\partial \theta}\right)^2\right\}} = \frac{1}{N \int_{\mathbb{M}} \left(\frac{\partial \ln p(\mathbf{m}|\theta)}{\partial \theta}\right)^2 p(\mathbf{m}\,|\,\theta)\,\mathrm{d}\,\mathbf{m}}$$

Details und Verallgemeinerungen siehe z.B.: K. Kroschel, „Statistische Informationstechnik“, Springer 2004

- CRB ist eine untere Schranke für die Varianz aller erwartungstreuen Schätzer.
- Effiziente Schätzverfahren erreichen die CRB.

Beweisskizze CRB: $\mathbf{m} \in \mathbb{R}^d, \theta \in \mathbb{R}, \; N = 1$

Voraussetzungen: $\frac{\partial}{\partial \theta} \int p(\mathbf{m} \mid \theta) \mathrm{d}\mathbf{m} = \int \frac{\partial p(\mathbf{m} \mid \theta)}{\partial \theta} \mathrm{d}\mathbf{m}$

$\hat{\theta}(\mathbf{m})$ ist erwartungstreu: $\mathrm{E}\{\hat{\theta}(\mathbf{m})\} = \theta$

$$\frac{\partial}{\partial \theta} \int \left(\hat{\theta}(\mathbf{m}) - \theta\right) p(\mathbf{m} \mid \theta) \mathrm{d}\mathbf{m} = \int \frac{\partial}{\partial \theta} \left(\hat{\theta}(\mathbf{m}) - \theta\right) p(\mathbf{m} \mid \theta) \mathrm{d}\mathbf{m} = 0$$

mit $\frac{\partial \hat{\theta}(\mathbf{m})}{\partial \theta} = 0$ folgt:

$$-\int p(\mathbf{m} \mid \theta) \mathrm{d}\mathbf{m} + \int \left(\hat{\theta}(\mathbf{m}) - \theta\right) \frac{\partial p(\mathbf{m}|\theta)}{\partial \theta} \mathrm{d}\mathbf{m} = 0$$

mit $\frac{\partial p(\mathbf{m}|\theta)}{\partial \theta} = \frac{\partial \ln p(\mathbf{m}|\theta)}{\partial \theta} p(\mathbf{m} \mid \theta)$ folgt:

$$\int \left(\hat{\theta}(\mathbf{m}) - \theta\right) p(\mathbf{m} \mid \theta) \frac{\partial \ln p(\mathbf{m}|\theta)}{\partial \theta} \mathrm{d}\mathbf{m} = 1$$

4. Parameterschätzung

$$\int \left(\hat{\theta}(\mathbf{m}) - \theta\right) p(\mathbf{m} \mid \theta) \frac{\partial \ln p(\mathbf{m}|\theta)}{\partial \theta} \mathrm{d}\mathbf{m} = 1$$

Mit der Schwarzschen Ungleichung: $\int x^2(\mathbf{m})\mathrm{d}\mathbf{m} \int y^2(\mathbf{m})\mathrm{d}\mathbf{m} \geq \left(\int x(\mathbf{m}) y(\mathbf{m})\mathrm{d}\mathbf{m}\right)^2$

und $\left(\int \left(\hat{\theta}(\mathbf{m}) - \theta\right) \sqrt{p(\mathbf{m} \mid \theta)} \sqrt{p(\mathbf{m} \mid \theta)} \frac{\partial \ln p(\mathbf{m}|\theta)}{\partial \theta} \mathrm{d}\mathbf{m} \right)^2 = 1$ folgt:

$$\underbrace{\int \left(\hat{\theta}(\mathbf{m}) - \theta\right)^2 p(\mathbf{m} \mid \theta)\mathrm{d}\mathbf{m}}_{\mathrm{Var}\{\hat{\theta}(\mathbf{m})\}} \underbrace{\int \left(\frac{\partial \ln p(\mathbf{m}|\theta)}{\partial \theta}\right)^2 p(\mathbf{m} \mid \theta)\mathrm{d}\mathbf{m}}_{\mathrm{E}\left\{\left(\frac{\partial \ln p(\mathbf{m}|\theta)}{\partial \theta}\right)^2\right\}\text{: Fisher-Information}} \geq 1$$

⟶ Cramer-Rao-Schranke

4. Parameterschätzung

Plausibilität: Wie stark ändert sich $\mathbf{m}$ mit θ?

$$\mathbf{m} \sim p(\mathbf{m}\,|\,\theta) \xrightarrow[\frac{\partial}{\partial\theta}]{} \frac{\partial p(\mathbf{m}|\theta)}{\partial\theta} \xrightarrow[\cdot\frac{1}{p(\mathbf{m}|\theta)}]{} \frac{\partial p}{\partial\theta}\frac{1}{p} = \frac{\partial \ln p(\mathbf{m}|\theta)}{\partial\theta}$$

$$\xrightarrow[(\,)^2]{} \left(\frac{\partial \ln p(\mathbf{m}|\theta)}{\partial\theta}\right)^2 \xrightarrow[\mathrm{E}\{\,\}]{} \mathrm{E}\left\{\left(\frac{\partial \ln p(\mathbf{m}|\theta)}{\partial\theta}\right)^2\right\} = J(\theta)$$

Quantifizierung der Empfindlichkeit von $\mathbf{m}$ bezüglich Veränderungen von θ.

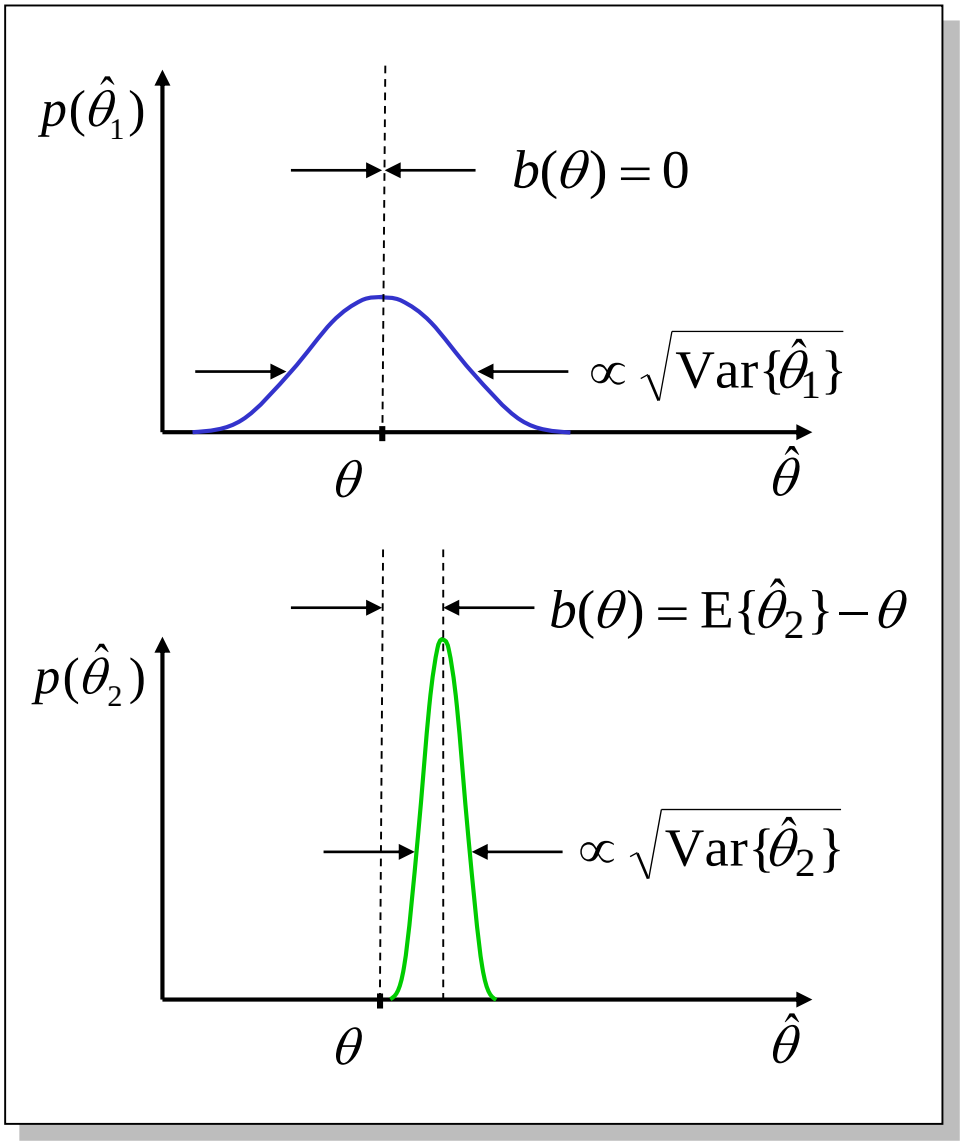
Eigenschaften von Schätzern: Konsistenz

Ein Schätzverfahren heißt konsistent, falls mit wachsender Zahl N der zur Schätzung herangezogenen Beobachtungen die Wahrscheinlichkeit, mit der die Schätzung vom wahren Wert um ein beliebiges ε abweicht, gegen Null konvergiert:

$$\lim_{N\to\infty} P\left\{ \| \hat{\boldsymbol{\theta}}(\mathrm{D}) - \boldsymbol{\theta} \| \geq \varepsilon \right\} = 0, \quad \forall \varepsilon > 0$$

$$\mathrm{D} = \{\mathbf{m}_1, \ldots, \mathbf{m}_N\}$$

Diskussion: Erwartungstreue, Varianz, mittlerer quadratischer Fehler



θ: sei wahrer Wert

Schätzer $\hat{\theta}_1$: erwartungstreu

Schätzer $\hat{\theta}_2$: nicht erwartungstreu, aber:

$$\sqrt{\mathrm{Var}\{\hat{\theta}_2\}} \quad < \quad \sqrt{\mathrm{Var}\{\hat{\theta}_1\}}$$

$\mathrm{Var}\{.\}$: stochastischer Fehler

Bias b: systematischer Fehler

$\mathrm{E}\{(\hat{\theta} - \theta)^2\} = b^2(\theta) + \mathrm{Var}\{\hat{\theta}\}$:
mittlerer quadratischer Fehler

Welcher Schätzer ist besser?

Beispiel: N stochastisch unabhängig gewonnene Stichproben einer Zufallsvariablen m, für die gilt: $m \sim p(m \mid \mu)$ (unabhängige identisch verteilte Stichprobe: i.i.d.). Es gelte:

$$\mu = \mathrm{E}\{m\}$$

$$\sigma^2 = \mathrm{Var}\{m\} = \mathrm{E}\{(m-\mu)^2\}$$

Erwartungswert μ soll durch arithmetischen Mittelwert geschätzt werden.

$$\hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} m_k$$

$$\mathrm{E}\{\hat{\mu}\} = \mathrm{E}\left\{\frac{1}{N}\sum_{k=1}^{N} m_k\right\} = \frac{1}{N}\sum_{k=1}^{N}\mathrm{E}\{m_k\} = \frac{1}{N}\sum_{k=1}^{N}\mu = \mu$$ →Erwartungstreue

Beispiel: Fortsetzung

Varianz des Schätzers:

$$\mathrm{Var}\{\hat{\mu}\} = \mathrm{E}\left\{(\hat{\mu}-\mu)^2\right\} = \mathrm{E}\left\{\left(\frac{1}{N}\sum_{k=1}^{N} m_k - \mu\right)^2\right\} = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\mathrm{E}\left\{(m_k-\mu)(m_l-\mu)\right\}$$

$$\mathrm{Var}\{\hat{\mu}\} = \frac{1}{N^2}\sum_{k=1}^{N}\mathrm{E}\left\{(m_k-\mu)^2\right\} = \frac{1}{N^2}\sum_{k=1}^{N}\sigma^2 = \frac{1}{N}\sigma^2$$

- Varianz verschwindet mit wachsendem N → Konsistenz aufgrund der Ungleichung von Tschebyscheff: $P\left\{\left|\hat{\theta}-\mathrm{E}\{\hat{\theta}\}\right| \geq \varepsilon\right\} \leq \mathrm{Var}\{\hat{\theta}\}/\varepsilon^2, \quad \forall \varepsilon > 0$
- Abnahme der Standardabweichung des Schätzers $\sqrt{\mathrm{Var}\{\hat{\mu}\}}$ mit $1/\sqrt{N}$ ist typisch für die meisten praktisch relevanten Aufgabenstellungen.

4.1. Maximum Likelihood Schätzung

Gegeben:

N stochastisch unabhängig gewonnene Stichproben eines Zufallsvektors $\mathbf{m}$, für den gilt: $\mathbf{m} \sim p(\mathbf{m}|\boldsymbol{\theta})$ (unabhängige identisch verteilte Stichprobe: i.i.d.).

Dann gilt:

$$p(\mathrm{D}|\boldsymbol{\theta}) = \prod_{k=1}^{N} p(\mathbf{m}_k|\boldsymbol{\theta}) \qquad \mathrm{D} = \{\mathbf{m}_1, ..., \mathbf{m}_N\}$$

$p(\mathrm{D}|\boldsymbol{\theta})$ als Funktion von $\boldsymbol{\theta}$ wird als Likelihoodfunktion bezeichnet.

Maximum Likelihood Ansatz: Man schätze $\boldsymbol{\theta}$ derart, dass die Wahrscheinlichkeit (Wahrscheinlichkeitsdichte) $p(\mathrm{D}|\boldsymbol{\theta})$ der vorliegenden Beobachtungen D maximal wird.

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \{p(\mathrm{D}|\boldsymbol{\theta})\}$$

4.1. Maximum Likelihood Schätzung

Wegen der Monotonie der Logarithmusfunktion, kann man äquivalent die logarithmierte Likelihoodfunktion maximieren; das bringt oft analytische Vorteile.

$$l(\boldsymbol{\theta}) := \ln(p(D \mid \boldsymbol{\theta}))$$

Loglikelihoodfunktion

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{N} \ln(p(\mathbf{m}_k \mid \boldsymbol{\theta}))$$

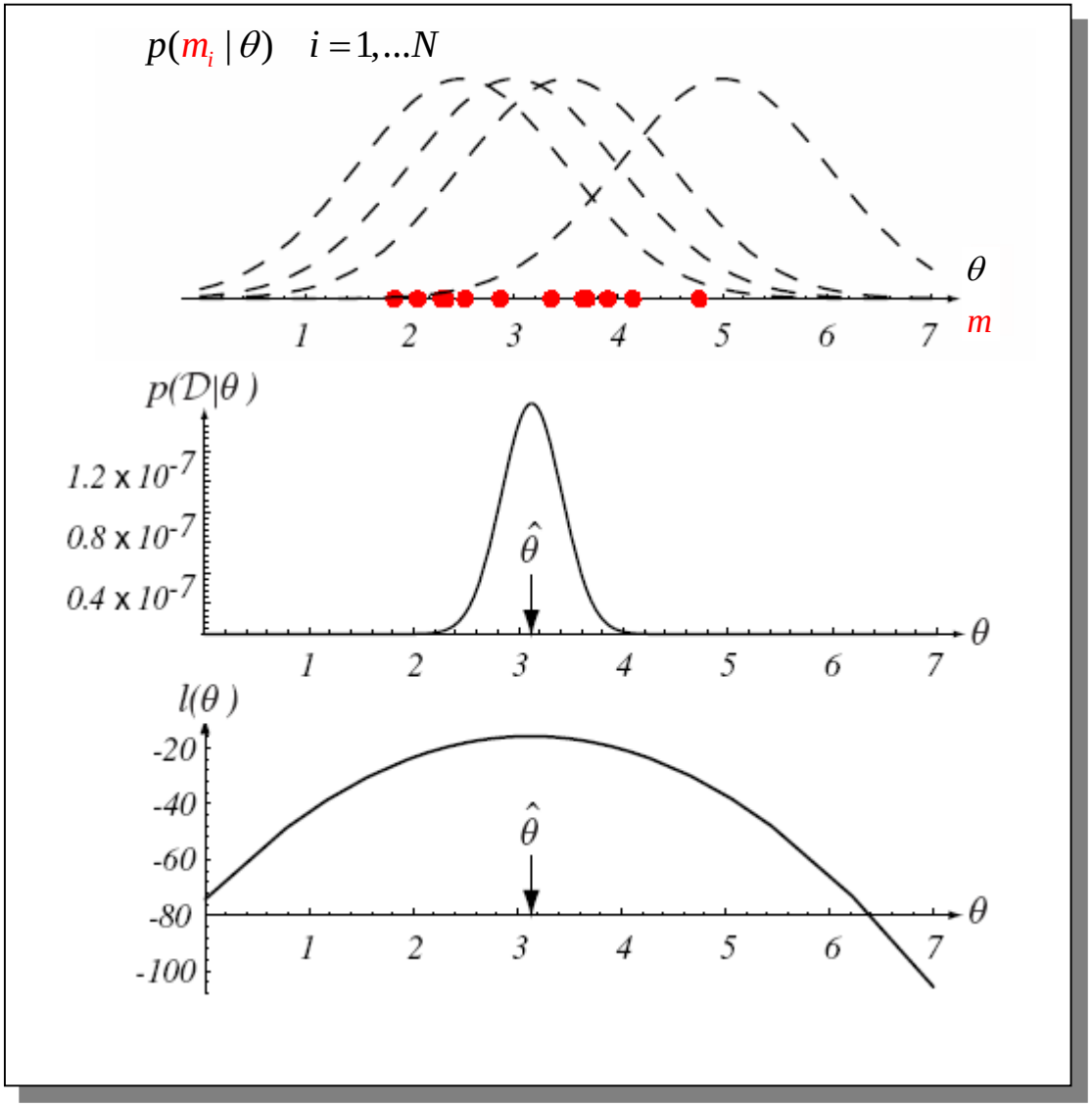
$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \{l(\boldsymbol{\theta})\}$$

Notwendig für Maximum:

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}} \ln(p(\mathbf{m}_k \mid \boldsymbol{\theta})) \stackrel{!}{=} \mathbf{0} \quad \text{mit} \quad \nabla_{\boldsymbol{\theta}} \equiv \left[\frac{\partial}{\partial \theta_1}, \ \ldots \ , \frac{\partial}{\partial \theta_c} \right]^T$$

4.1. Maximum Likelihood Schätzung

Beispiel: rote Punkte sind Stichproben einer normalverteilten Zufallsvariablen m mit bekannter Standardabweichung σ und unbekanntem, zu schätzendem Erwartungswert θ.



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

4.1. Maximum Likelihood Schätzung

Beispiel: $\mathbf{m} \sim \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ unbekannt

$$\ln p(\mathbf{m}_k \mid \boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{m}_k - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{m}_k - \boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln |\boldsymbol{\Sigma}|$$

$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{m}_k \mid \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{m}_k - \boldsymbol{\mu})$$

$$\sum_{k=1}^{N} \nabla_{\boldsymbol{\mu}} \ln p(\mathbf{m}_k \mid \boldsymbol{\mu}) = \sum_{k=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{m}_k - \hat{\boldsymbol{\mu}}) = 0$$

$$\sum_{k=1}^{N} (\mathbf{m}_k - \hat{\boldsymbol{\mu}}) = 0$$

$$\Rightarrow \quad \boxed{\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} \mathbf{m}_k}$$

4.1. Maximum Likelihood Schätzung

Beispiel: $m \sim \mathrm{N}(m;\mu,\sigma^2)$, μ,σ^2 unbekannt

$$\boldsymbol{\theta} = (\mu,\sigma^2)^{\mathrm{T}} = (\theta_1,\theta_2)^{\mathrm{T}}$$

$$\ln p(m_k \mid \boldsymbol{\theta}) = -\frac{1}{2\theta_2}(m_k-\theta_1)^2 - \frac{1}{2}\ln 2\pi\theta_2$$

$$\sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}} \ln p(m_k \mid \boldsymbol{\theta}) = \sum_{k=1}^{N} \left[\frac{1}{\theta_2}(m_k-\theta_1),\ -\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(m_k-\theta_1)^2 \right]^{\mathrm{T}} \overset{!}{=} \mathbf{0} \quad \Leftrightarrow$$

$$\left.\begin{aligned} &\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2}(m_k-\hat{\theta}_1) = 0 \\ &\sum_{k=1}^{N} \left(-\frac{1}{2\hat{\theta}_2} + \frac{1}{2\hat{\theta}_2^2}(m_k-\hat{\theta}_1)^2 \right) = 0 \end{aligned}\right\} \quad \Leftrightarrow$$

ML-Schätzer:

$$\hat{\theta}_1 = \hat{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} m_k$$

$$\hat{\theta}_2 = \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N} (m_k-\hat{\mu})^2$$

4.1. Maximum Likelihood Schätzung

Beispiel: $\mathbf{m} \sim \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ unbekannt

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})^{\mathrm{T}}$$

$$\ln p(\mathbf{m}_k \mid \boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{m}_k - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{m}_k - \boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln |\boldsymbol{\Sigma}|$$

$$\sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{m}_k \mid \boldsymbol{\theta}) =$$

$$\sum_{k=1}^{N} \left[\boldsymbol{\Sigma}^{-1}(\mathbf{m}_k - \boldsymbol{\mu}), \quad -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}(\mathbf{m}_k - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{m}_k - \boldsymbol{\mu}) - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} \ln |\boldsymbol{\Sigma}| \right]^{\mathrm{T}} \stackrel{!}{=} \mathbf{0}$$

$$\Leftrightarrow \quad \begin{cases} \displaystyle\sum_{k=1}^{N} \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{m}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0} \\ \displaystyle\sum_{k=1}^{N} \left[-\frac{1}{2}\frac{\partial}{\partial \hat{\boldsymbol{\Sigma}}}(\mathbf{m}_k - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{m}_k - \hat{\boldsymbol{\mu}}) - \frac{1}{2}\frac{\partial}{\partial \hat{\boldsymbol{\Sigma}}} \ln |\hat{\boldsymbol{\Sigma}}| \right] = \mathbf{0} \end{cases}$$

Beispiel, Fortsetzung

$$\sum_{k=1}^{N}\left[-\frac{1}{2}\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}(\mathbf{m}_k-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{m}_k-\hat{\boldsymbol{\mu}})-\frac{1}{2}\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\ln|\hat{\boldsymbol{\Sigma}}|\right]=\mathbf{0}$$

$$-\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\sum_{k=1}^{N}\left[(\mathbf{m}_k-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{m}_k-\hat{\boldsymbol{\mu}})\right]-N\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\ln|\hat{\boldsymbol{\Sigma}}|=\mathbf{0}$$

$$\overline{\mathbf{m}}=\frac{1}{N}\sum_{k=1}^{N}\mathbf{m}_k$$

$$\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\sum_{k=1}^{N}\left[(\mathbf{m}_k-\overline{\mathbf{m}}+\overline{\mathbf{m}}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{m}_k-\overline{\mathbf{m}}+\overline{\mathbf{m}}-\hat{\boldsymbol{\mu}})\right]+N\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\ln|\hat{\boldsymbol{\Sigma}}|=\mathbf{0}$$

$$\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\sum_{k=1}^{N}\left[(\mathbf{m}_k-\overline{\mathbf{m}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{m}_k-\overline{\mathbf{m}})\right]+N\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}(\overline{\mathbf{m}}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{m}}-\hat{\boldsymbol{\mu}})+N\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\ln|\hat{\boldsymbol{\Sigma}}|=\mathbf{0}$$

Skalar

$$\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\mathrm{tr}\left[\hat{\boldsymbol{\Sigma}}^{-1}\sum_{k=1}^{N}(\mathbf{m}_k-\overline{\mathbf{m}})(\mathbf{m}_k-\overline{\mathbf{m}})^{\mathrm{T}}\right]+N\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}(\overline{\mathbf{m}}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{m}}-\hat{\boldsymbol{\mu}})+N\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\ln|\hat{\boldsymbol{\Sigma}}|=\mathbf{0}$$

$$N\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{S}}\right)+N\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}(\overline{\mathbf{m}}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{m}}-\hat{\boldsymbol{\mu}})+N\frac{\partial}{\partial\hat{\boldsymbol{\Sigma}}}\ln|\hat{\boldsymbol{\Sigma}}|=\mathbf{0}$$

$$\text{mit}\quad\tilde{\mathbf{S}}:=\frac{1}{N}\mathbf{S}=\frac{1}{N}\sum_{k=1}^{N}(\mathbf{m}_k-\overline{\mathbf{m}})^{\mathrm{T}}(\mathbf{m}_k-\overline{\mathbf{m}})$$

Beispiel, Fortsetzung

$$\frac{\partial}{\partial \hat{\boldsymbol{\Sigma}}} \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{S}}\right) + \frac{\partial}{\partial \hat{\boldsymbol{\Sigma}}} (\bar{\mathbf{m}} - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1} (\bar{\mathbf{m}} - \hat{\boldsymbol{\mu}}) - \frac{\partial}{\partial \hat{\boldsymbol{\Sigma}}} \ln |\hat{\boldsymbol{\Sigma}}^{-1}| = \mathbf{0}$$

$$\left.\begin{array}{l} \dfrac{\partial}{\partial \hat{\boldsymbol{\Sigma}}} \operatorname{tr}\left(\mathbf{V}\tilde{\mathbf{S}}\right) + \dfrac{\partial}{\partial \hat{\boldsymbol{\Sigma}}} (\bar{\mathbf{m}} - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \mathbf{V} (\bar{\mathbf{m}} - \hat{\boldsymbol{\mu}}) - \dfrac{\partial}{\partial \hat{\boldsymbol{\Sigma}}} \ln |\mathbf{V}| = \mathbf{0} \\ \dfrac{\partial \ln p(\mathbf{m}_k \mid \boldsymbol{\theta})}{\partial \hat{\boldsymbol{\Sigma}}} = \dfrac{\partial \ln p(\mathbf{m}_k \mid \boldsymbol{\theta})}{\partial \mathbf{V}} \dfrac{\partial \mathbf{V}}{\partial \hat{\boldsymbol{\Sigma}}} \\ \dfrac{\partial \mathbf{V}}{\partial \hat{\boldsymbol{\Sigma}}} = -\hat{\boldsymbol{\Sigma}}^{-2} \end{array}\right\} \Rightarrow$$

$$\left.\begin{array}{l} \dfrac{\partial}{\partial \mathbf{V}} \operatorname{tr}\left(\mathbf{V}\tilde{\mathbf{S}}\right) + \dfrac{\partial}{\partial \mathbf{V}} (\bar{\mathbf{m}} - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \mathbf{V} (\bar{\mathbf{m}} - \hat{\boldsymbol{\mu}}) - \dfrac{\partial}{\partial \mathbf{V}} \ln |\mathbf{V}| = \mathbf{0} \\ \dfrac{\partial}{\partial \mathbf{X}} \ln |\mathbf{X}| = 2\mathbf{X}^{-1} - \operatorname{diag}\left(\mathbf{X}^{-1}\right) \\ \dfrac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{A}\mathbf{X}) = \dfrac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{X}\mathbf{A}) = \mathbf{A}^{\mathrm{T}} + \mathbf{A} - \operatorname{diag}(\mathbf{A}) \end{array}\right\} \Rightarrow$$

$$2\tilde{\mathbf{S}} - \operatorname{diag}(\tilde{\mathbf{S}}) + \frac{\partial}{\partial \mathbf{V}} (\bar{\mathbf{m}} - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \mathbf{V} (\bar{\mathbf{m}} - \hat{\boldsymbol{\mu}}) - 2\hat{\boldsymbol{\Sigma}} + \operatorname{diag}(\hat{\boldsymbol{\Sigma}}) = \mathbf{0}$$

Beispiel, Fortsetzung

$$\begin{cases} \hat{\boldsymbol{\mu}} = \dfrac{1}{N}\sum_{k=1}^{N}\mathbf{m}_k \\ 2\tilde{\mathbf{S}} - \operatorname{diag}(\tilde{\mathbf{S}}) + \dfrac{\partial}{\partial \mathbf{V}}(\overline{\mathbf{m}} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}\mathbf{V}(\overline{\mathbf{m}} - \hat{\boldsymbol{\mu}}) - 2\hat{\boldsymbol{\Sigma}} + \operatorname{diag}(\hat{\boldsymbol{\Sigma}}) = \mathbf{0} \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\boldsymbol{\mu}} = \dfrac{1}{N}\sum_{k=1}^{N}\mathbf{m}_k \\ 2\tilde{\mathbf{S}} - 2\hat{\boldsymbol{\Sigma}} + \operatorname{diag}(\hat{\boldsymbol{\Sigma}} - \tilde{\mathbf{S}}) = \mathbf{0} \end{cases} \Leftrightarrow$$

ML-Schätzer:

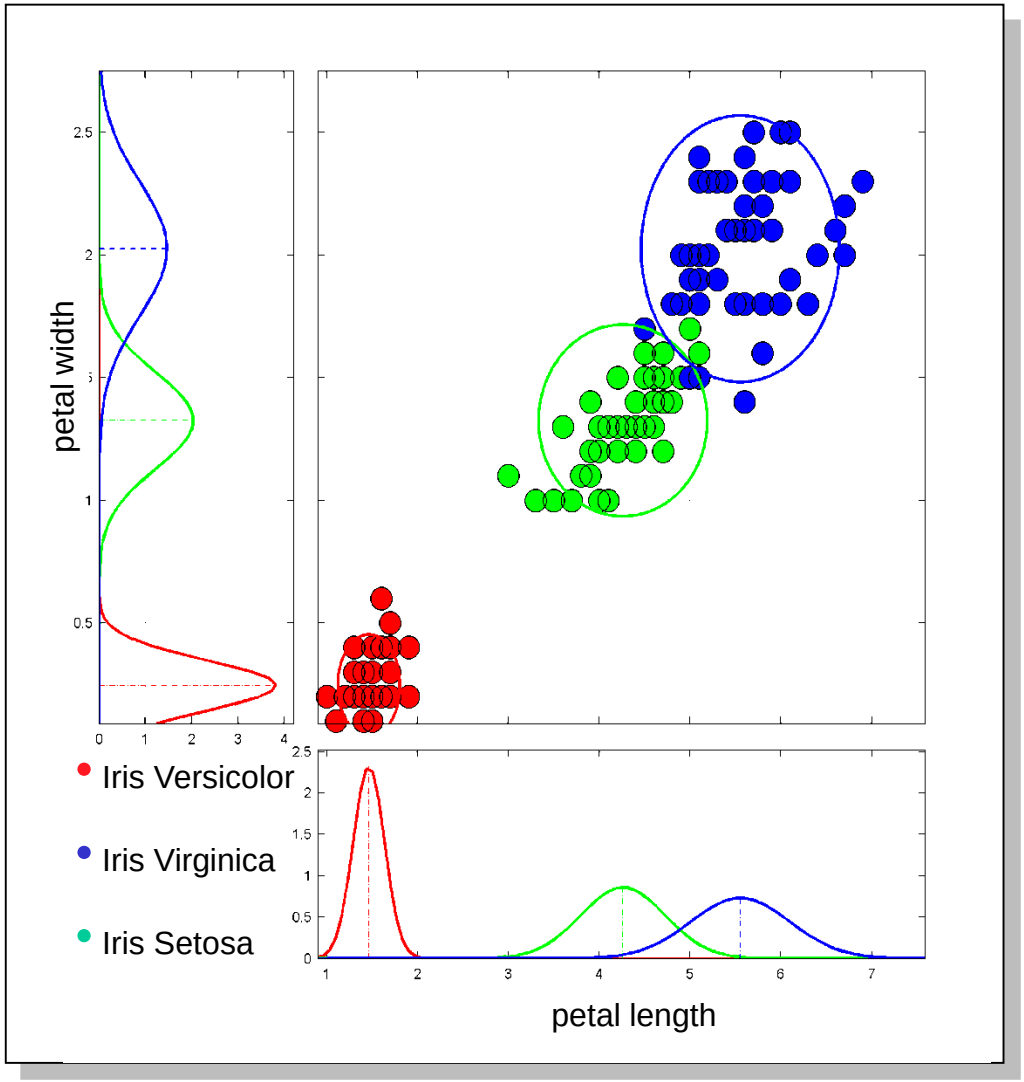
$$\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{m}_k$$

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{m}_k - \hat{\boldsymbol{\mu}})(\mathbf{m}_k - \hat{\boldsymbol{\mu}})^{\mathrm{T}}$$

Differentiation nach Vektoren und Matrizen, siehe z.B.:
T. K. Moon; W. C. Stirling: Mathematical Methods and Algorithms for Signal Processing, Prentice Hall, 2000.
C. Voigt; J. Adamy: Formelsammlung der Matrizenrechnung, Oldenbourg Verlag, 2007.

4.1. Maximum Likelihood Schätzung

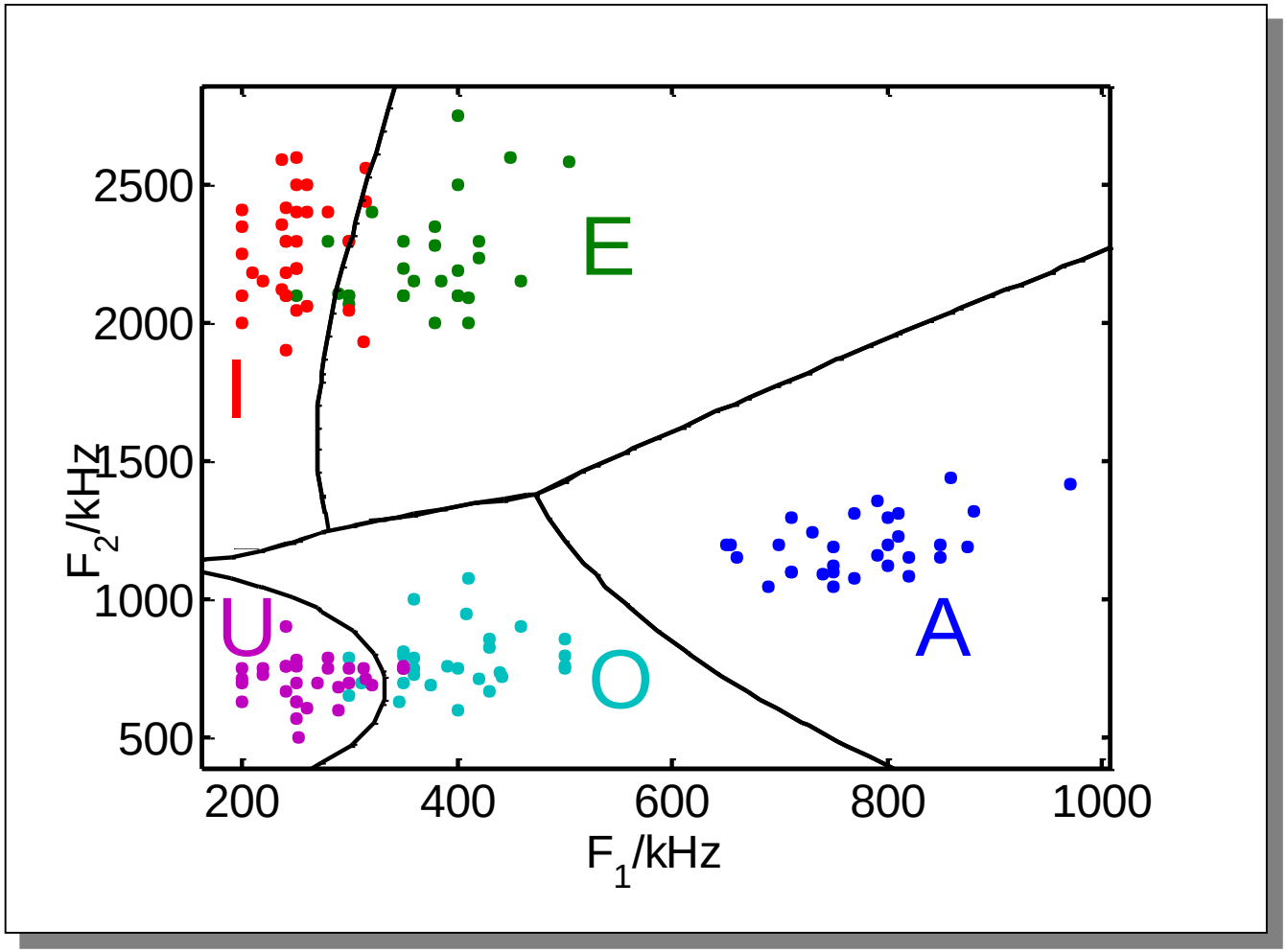
Beispiel: Iris Datensatz nach R. A. Fisher, Modell: Gauß‘sche WDFen mit unkorrelierten Merkmalen, Maximum Likelihood Schätzung der Parameter



Ellipsen kennzeichnen die 2σ - Grenzen der Wahrscheinlichkeitsdichten.

4.1. Maximum Likelihood Schätzung

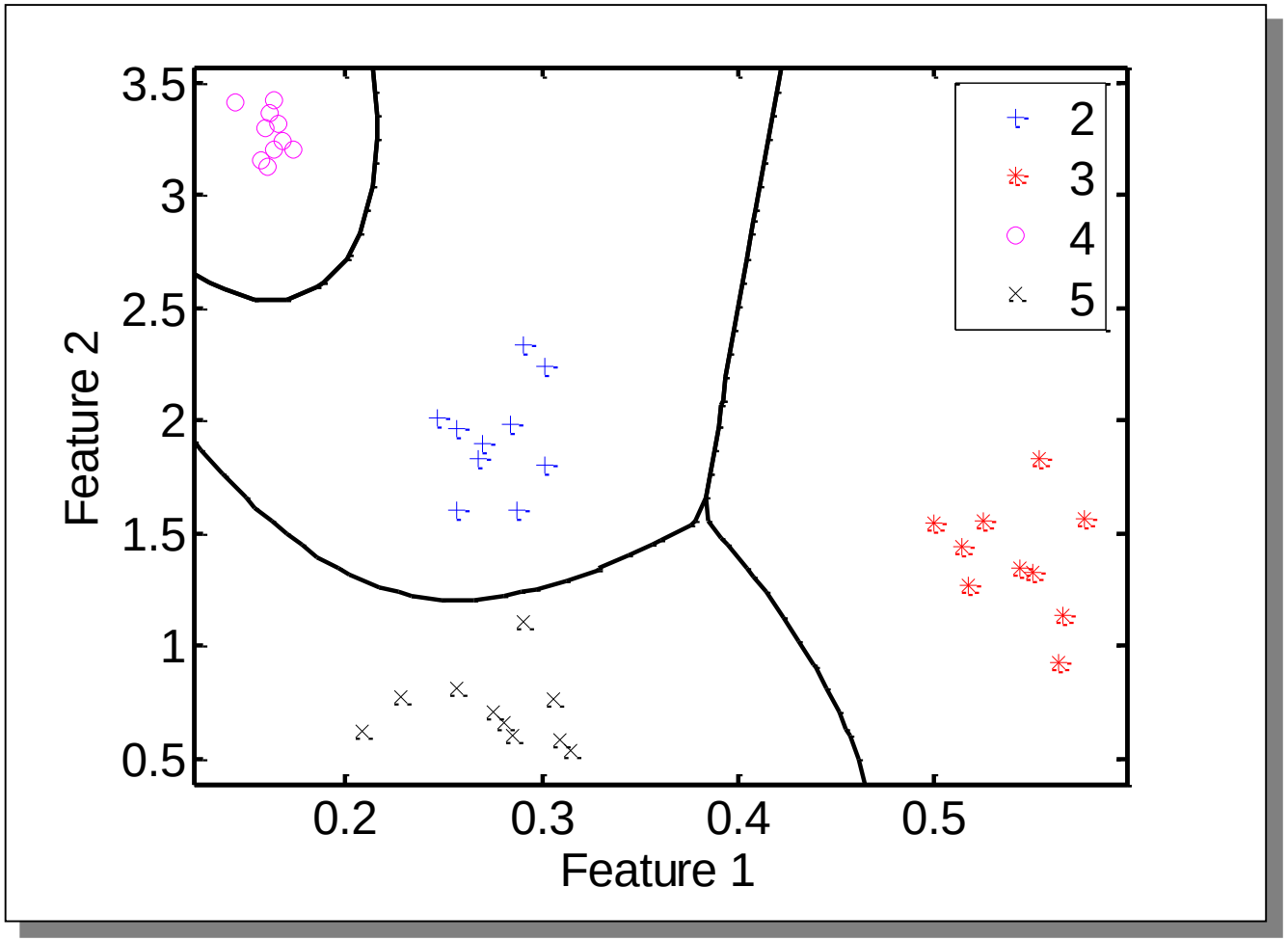
Beispiel: Vokalerkennung



Bayes'sche Klassifikation, Modell: Normalverteilung für klassenbedingte Merkmals-WDFen, Parameterschätzung: Maximum Likelihood

4.1. Maximum Likelihood Schätzung

Beispiel: Grifferkennung Hand



Bayes'sche Klassifikation, Modell: Normalverteilung für klassenbedingte Merkmals-WDFen, Parameterschätzung: Maximum Likelihood

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Fundamentalgröße der Bayes'schen Klassifikation:
A Posteriori Wahrscheinlichkeitsverteilungen der Klassen

$$P(\omega_i \mid \mathbf{m}) = \frac{p(\mathbf{m} \mid \omega_i) P(\omega_i)}{p(\mathbf{m})}$$

Problem: WVen unbekannt → Müssen aus den Daten D geschätzt werden.

Ansatz:

$$P(\omega_i \mid \mathbf{m}, \mathrm{D}) = \frac{p(\mathbf{m} \mid \omega_i, \mathrm{D}) P(\omega_i \mid \mathrm{D})}{\sum_{j=1}^{c} p(\mathbf{m} \mid \omega_j, \mathrm{D}) P(\omega_j \mid \mathrm{D})}$$

Die $p(\mathbf{m} \mid \omega_i, \mathrm{D})$ approximieren die $p(\mathbf{m} \mid \omega_i)$.

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Vereinfachende Annahmen:

- A Priori Wahrscheinlichkeiten seien gegeben oder ausreichend genau bestimmt worden: $P(\omega_i) = P(\omega_i \mid \mathrm{D})$
- Die Stichproben in D_k haben keinen Einfluss auf $p(\mathbf{m} \mid \omega_i, \mathrm{D}) \quad i \neq k$.

$$P(\omega_i \mid \mathbf{m}, \mathrm{D}) = \frac{p(\mathbf{m} \mid \omega_i, \mathrm{D}_i) P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{m} \mid \omega_j, \mathrm{D}_j) P(\omega_j)}$$

- WDFen $p(\mathbf{m} \mid \omega_i)$ haben eine bekannte parametrisierte Form $p(\mathbf{m} \mid \boldsymbol{\theta}_i, \omega_i)$ mit $p(\boldsymbol{\theta}_i \mid \omega_i)$ bekannt.

$$p(\mathbf{m} \mid \omega_i, \mathrm{D}_i) = \int_{\Theta_i} p(\mathbf{m}, \boldsymbol{\theta}_i \mid \omega_i, \mathrm{D}_i)\, \mathrm{d}\boldsymbol{\theta}_i = \int_{\Theta_i} p(\mathbf{m} \mid \boldsymbol{\theta}_i, \omega_i, \cancel{\mathrm{D}_i})\, p(\boldsymbol{\theta}_i \mid \omega_i, \mathrm{D}_i)\, \mathrm{d}\boldsymbol{\theta}_i$$

→

$$p(\mathbf{m} \mid \omega_i, \mathrm{D}_i) = \int_{\Theta_i} p(\mathbf{m} \mid \boldsymbol{\theta}_i, \omega_i)\, p(\boldsymbol{\theta}_i \mid \omega_i, \mathrm{D}_i)\, \mathrm{d}\boldsymbol{\theta}_i$$

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Die Gesamtaufgabe zerfällt in c Teilaufgaben; eine pro Klasse. Für jede Klasse ω_i gilt es, die klassenspezifische WDF

$$p(\mathbf{m} \mid \omega_i, \mathrm{D}_i) = \int_{\Theta_i} p(\mathbf{m} \mid \boldsymbol{\theta}_i, \omega_i)\, p(\boldsymbol{\theta}_i \mid \omega_i, \mathrm{D}_i)\, \mathrm{d}\boldsymbol{\theta}_i \qquad (*)$$

anhand der Daten D_i zu bestimmen.

Zur einfacheren Schreibweise wird im Folgenden die Kennzeichnung der Klassen unterdrückt. Die Überlegungen gelten aber immer nur für eine Klasse ω_i.

→ Vereinfachte Schreibweise für (*):

$$p(\mathbf{m} \mid \mathrm{D}) = \int_{\Theta} p(\mathbf{m} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathrm{D})\, \mathrm{d}\boldsymbol{\theta}$$

Da $p(\mathbf{m} \mid \boldsymbol{\theta})$ als bekannt vorausgesetzt ist, gilt es im Folgenden, $p(\boldsymbol{\theta} \mid \mathrm{D})$ zu bestimmen.

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Beispiel: $m \sim \mathrm{N}(m;\mu,\sigma^2)$, μ unbekannt

Annahme: $\mu \sim \mathrm{N}(\mu;\mu_0,\sigma_0^2)$ (μ_0,σ_0^2 sind sogenannte Hyperparameter)

$$p(\mu \mid \mathrm{D}) = \frac{p(\mathrm{D} \mid \mu)\, p(\mu)}{\int p(\mathrm{D} \mid \mu)\, p(\mu)\, \mathrm{d}\mu} = \alpha(\mathrm{D}) \prod_{k=1}^{N} p(m_k \mid \mu)\, p(\mu)$$

$$p(\mu \mid \mathrm{D}) = \alpha(\mathrm{D}) \prod_{k=1}^{N} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{m_k-\mu}{\sigma}\right)^2\right]}_{p(m_k \mid \mu)} \underbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}_{p(\mu)}$$

$$= \alpha'(\mathrm{D}) \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{N}\left(\frac{\mu-m_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha''(\mathrm{D}) \exp\left[-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{N} m_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

(→ ... kann nur eine Gauß'sche WDF sein! Quadratische Ergänzung → …)

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Beispiel, Fortsetzung

Diese WDF lässt sich schreiben als: $p(\mu \mid \mathrm{D}) = \mathrm{N}(\mu; \mu_N, \sigma_N^2)$

$$p(\mu \mid \mathrm{D}) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_N}{\sigma_N} \right)^2 \right]$$

$$\mu_N := \left(\frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \right) \hat{\mu}_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \qquad \sigma_N^2 := \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}$$

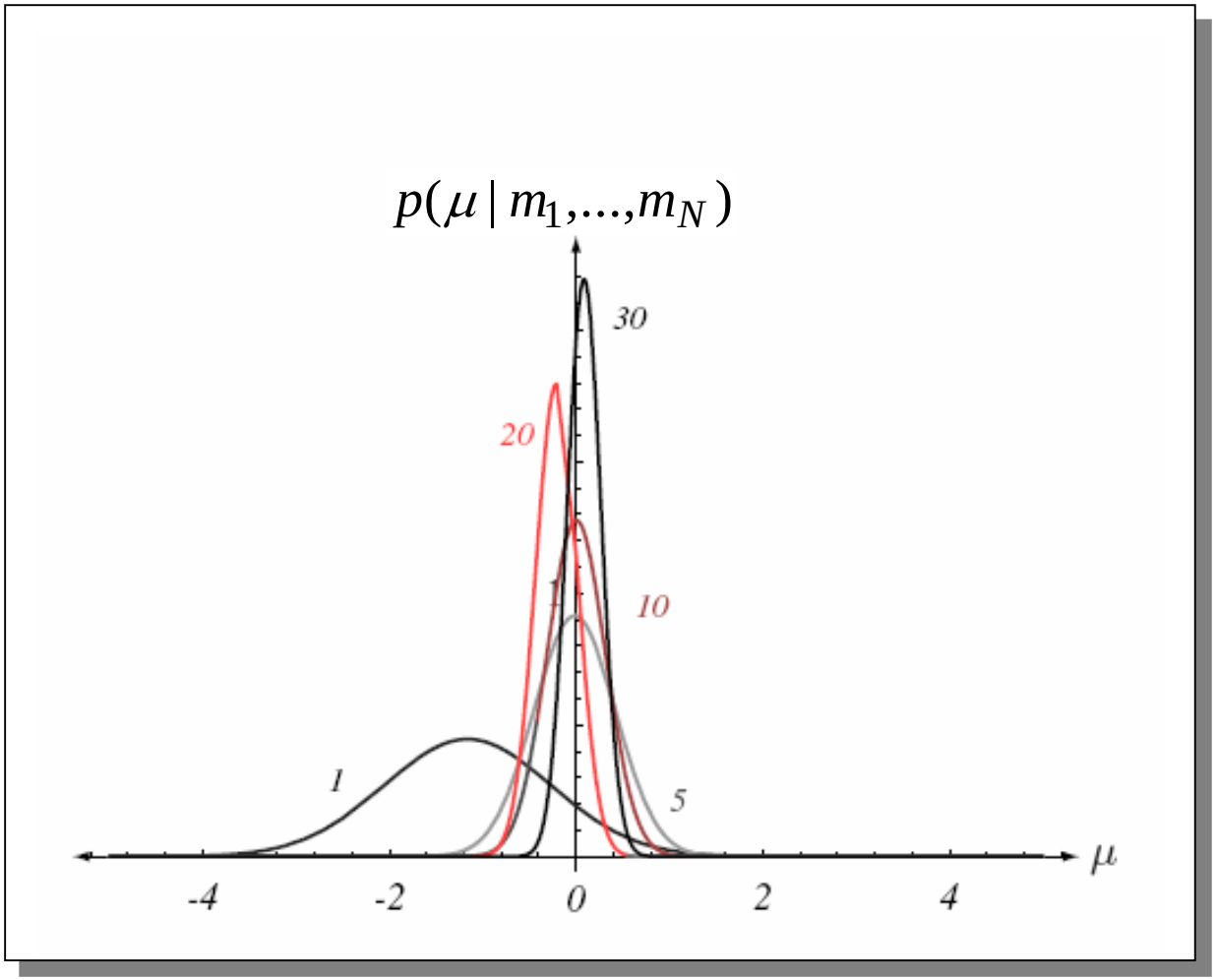
$$\hat{\mu}_N := \frac{1}{N} \sum_{k=1}^{N} m_k$$

Diskussion:

- A Priori + Empirische Information → A Posteriori WDF $p(\mu \mid \mathrm{D})$
- μ_N die beste Schätzung (MAP) für μ nach N beobachteten Werten.
- σ_N^2 beschreibt die Schätzunsicherheit (stochastischer Schätzfehler)
- μ_N liegt zwischen $\hat{\mu}_N$ und μ_0

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Beispiel, Fortsetzung



Quelle: R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification

A Posteriori Verteilungsdichte für wachsende Anzahl von Beobachtungen.

Beispiel, Fortsetzung

$$p(m \mid \mathrm{D}) = \int_{\Theta} p(m \mid \mu)\, p(\mu \mid \mathrm{D})\, \mathrm{d}\mu$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{m-\mu}{\sigma} \right)^2 \right] \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N} \right)^2 \right] \mathrm{d}\mu$$

$$= \frac{1}{2\pi\sigma\sigma_N} \exp\left[-\frac{1}{2} \frac{(m-\mu_N)^2}{\sigma^2+\sigma_N^2} \right] f(\sigma, \sigma_N)$$

$$\text{mit} \quad f(\sigma, \sigma_N) := \int \exp\left[-\frac{1}{2} \frac{\sigma^2+\sigma_N^2}{\sigma^2\sigma_N^2} \left(\mu - \frac{\sigma_N^2 m + \sigma^2 \mu_N}{\sigma_N^2+\sigma^2} \right)^2 \right] \mathrm{d}\mu$$

$$p(m \mid \mathrm{D}) = \mathrm{N}(m; \mu_N, \sigma^2 + \sigma_N^2)$$

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Beispiel: $p(\mathbf{m}\,|\,\boldsymbol{\mu}) = \mathrm{N}(\mathbf{m};\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ unbekannt

Annahme: $p(\boldsymbol{\mu}) = \mathrm{N}(\boldsymbol{\mu};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$

$$p(\boldsymbol{\mu}\,|\,\mathrm{D}) = \alpha \prod_{k=1}^{N} p(\mathbf{m}_k\,|\,\boldsymbol{\mu})\, p(\boldsymbol{\mu})$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^{\mathrm{T}} (N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^{\mathrm{T}} \left(\boldsymbol{\Sigma}^{-1} \sum_{k=1}^{N} \mathbf{m}_k + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \right) \right) \right]$$

$$= \alpha'' \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)^{\mathrm{T}} \boldsymbol{\Sigma}_N^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N) \right]$$

$$p(\boldsymbol{\mu}\,|\,\mathrm{D}) = \mathrm{N}(\boldsymbol{\mu};\boldsymbol{\mu}_N,\boldsymbol{\Sigma}_N)$$

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{N}\boldsymbol{\Sigma} \right)^{-1} \hat{\boldsymbol{\mu}}_N + \frac{1}{N}\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0 + \frac{1}{N}\boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_0$$

$$\boldsymbol{\Sigma}_N = \boldsymbol{\Sigma}_0 (\boldsymbol{\Sigma}_0 + \frac{1}{N}\boldsymbol{\Sigma})^{-1} \frac{1}{N}\boldsymbol{\Sigma}$$

$$\hat{\boldsymbol{\mu}}_N = \frac{1}{N} \sum_{k=1}^{N} \mathbf{m}_k$$

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Beispiel, Fortsetzung

$$p(\mathbf{m} \mid \mathrm{D}) = \int p(\mathbf{m} \mid \boldsymbol{\mu})\, p(\boldsymbol{\mu} \mid D)\, \mathrm{d}\boldsymbol{\mu}$$

$$p(\boldsymbol{\mu} \mid \mathrm{D}) = N(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$$

$$p(\mathbf{m} \mid \boldsymbol{\mu}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(\mathbf{m} \mid \mathrm{D}) = \mathrm{N}(\mathbf{m}; \boldsymbol{\mu}_N, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_N)$$

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Überblick zum Verfahren; für jede Klasse $i = 1,\ldots,c$ separat anzuwenden.

- $p(\mathbf{m} \mid \boldsymbol{\theta}_i, \omega_i)$ strukturell als bekannt angenommen; Wert $\boldsymbol{\theta}_i$ unbekannt.
- $p(\boldsymbol{\theta}_i \mid \omega_i)$ bekannt; enthält das A Priori Wissen über $\boldsymbol{\theta}_i$.
- Das weitere Wissen über $\boldsymbol{\theta}_i$ ist in D_i enthalten; D_i besteht aus N_i unabhängigen Stichproben $\mathbf{m}_1,\ldots,\mathbf{m}_{N_i}$ bezüglich der WDF $p(\mathbf{m}|\omega_i)$.

$$p(\mathrm{D}_i \mid \boldsymbol{\theta}_i, \omega_i) = \prod_{k=1}^{N} p(\mathbf{m}_k \mid \boldsymbol{\theta}_i, \omega_i)$$

$$p(\boldsymbol{\theta}_i \mid \mathrm{D}_i, \omega_i) = \frac{p(\mathrm{D}_i \mid \boldsymbol{\theta}_i, \omega_i)\, p(\boldsymbol{\theta}_i \mid \omega_i)}{\int p(\mathrm{D}_i \mid \boldsymbol{\theta}_i, \omega_i)\, p(\boldsymbol{\theta}_i \mid \omega_i)\, \mathrm{d}\boldsymbol{\theta}_i}$$

$$p(\mathbf{m} \mid \mathrm{D}_i, \omega_i) = \int p(\mathbf{m} \mid \boldsymbol{\theta}_i, \omega_i)\, p(\boldsymbol{\theta}_i \mid \mathrm{D}_i, \omega_i)\, \mathrm{d}\boldsymbol{\theta}_i$$

- Wenn $p(\boldsymbol{\theta}_i \mid \mathrm{D}_i, \omega_i)$ sich stark um einen Wert $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$ konzentriert, dort eine ausgeprägte Spitze ausbildet, dann kann $p(\mathbf{m} \mid \mathrm{D}_i, \omega_i)$ durch $p(\mathbf{m}|\hat{\boldsymbol{\theta}}_i, \omega_i)$ approximiert werden.

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Vergleich: Bayes'sches Verfahren und Maximum Likelihood Schätzung.

Bayes'sches Verfahren:

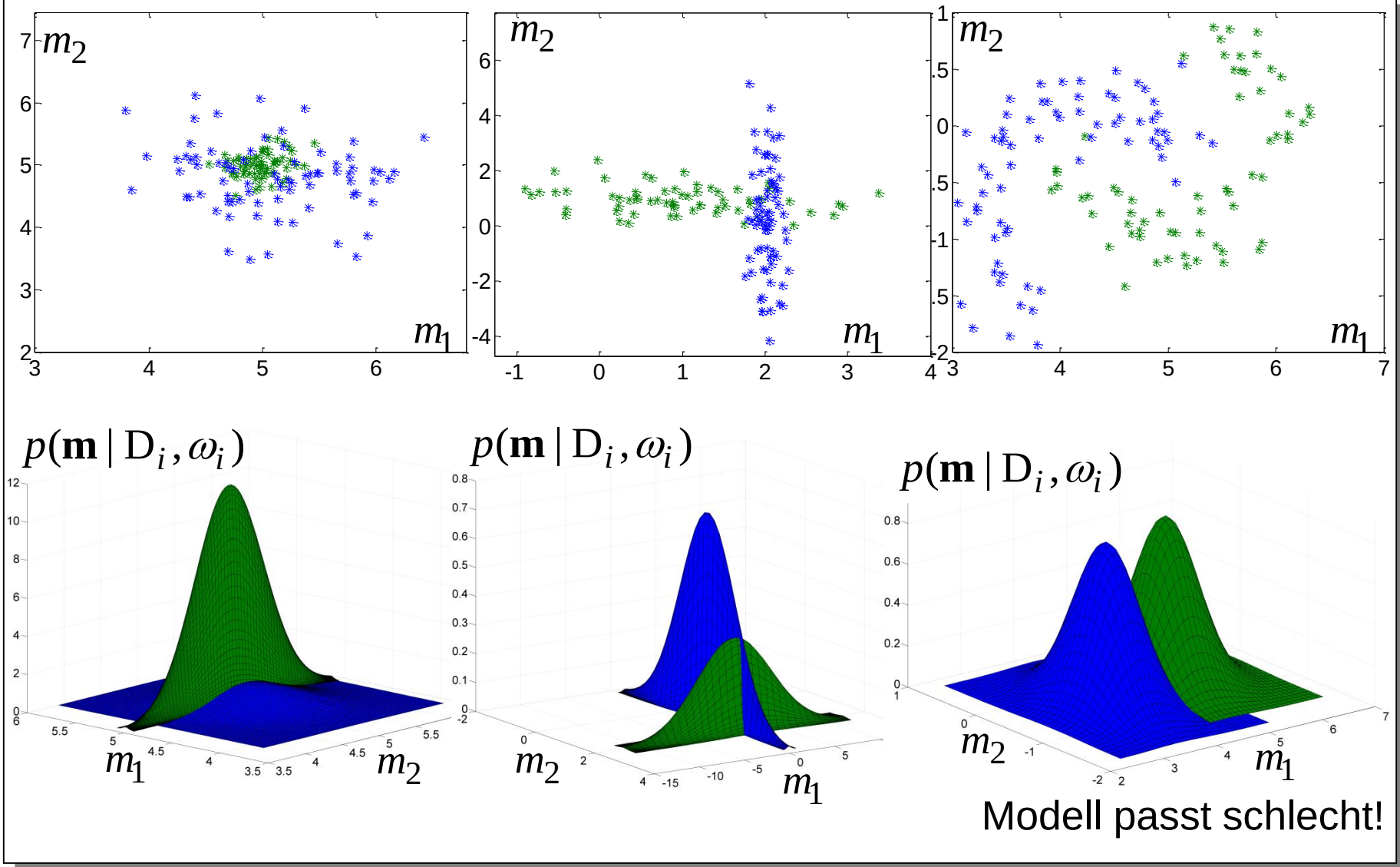
+ A Priori Wissen über θ kann eingebracht werden.
– Numerische Berechung mehrdimensionaler Integrale erforderlich.

Maximum Likelihood Verfahren:

+ Keine mehrdimensionalen Integrale, sondern *nur* Extremwertsuche.
– A Priori Wissen über θ kann nicht eingebracht werden.

4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

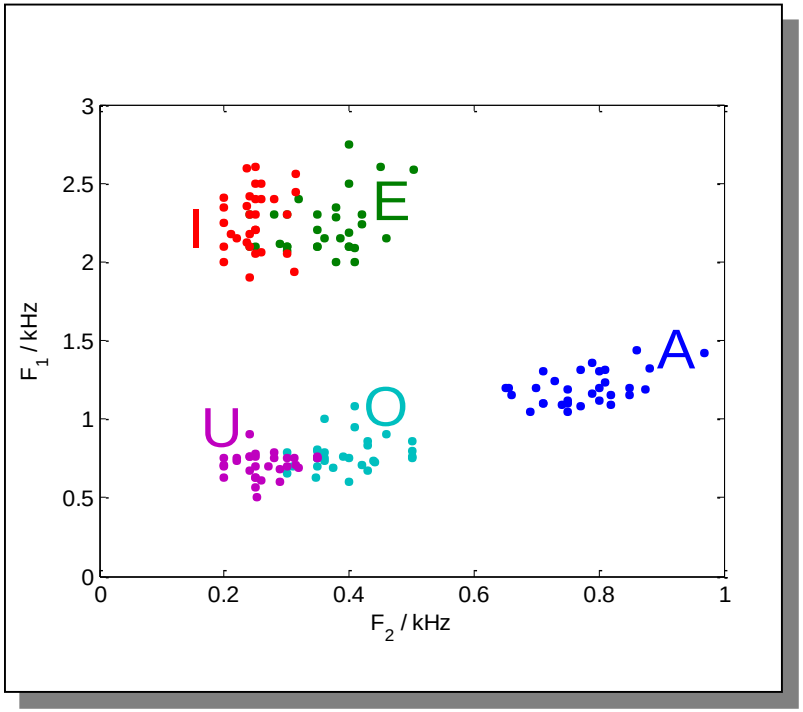
Beispiel: $d = 2$, $c = 2$, Daten und daran angepasste Normalverteilungen

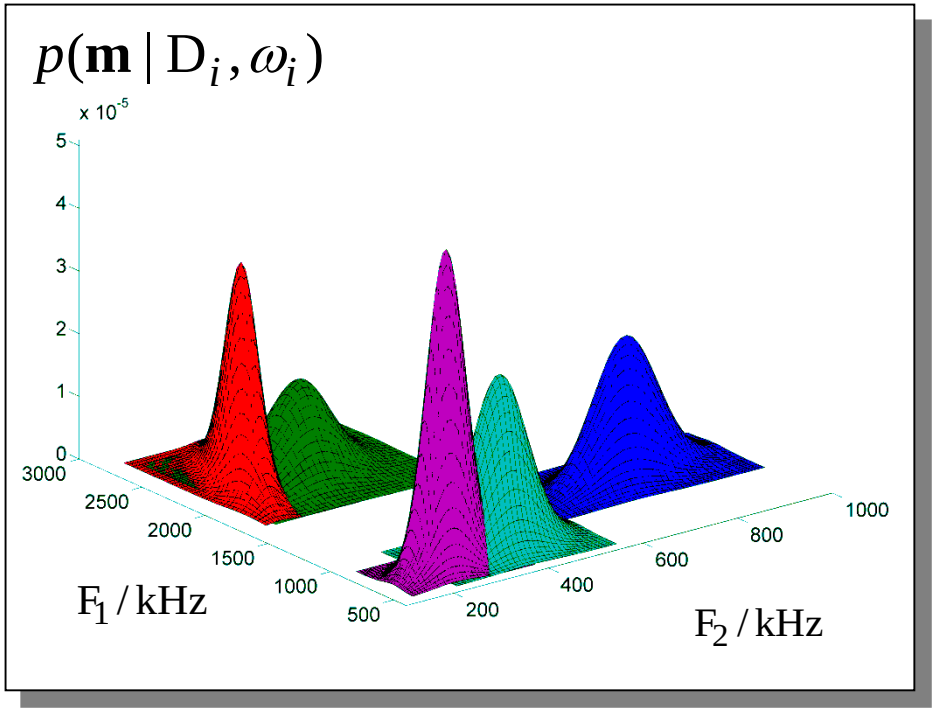


4.2. Bayes'sche Bestimmung der AP-Klassenwahrscheinlichkeiten

Beispiel: Vokalerkennung, $d = 2$, $c = 5$,

Daten und daran angepasste Normalverteilungen





4.3. Bayes'sche Parameterschätzung

Wenn $p(\boldsymbol{\theta} \mid \mathrm{D})$ sich stark um einen Wert $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ konzentriert, dort eine ausgeprägte Spitze ausbildet, dann kann $p(\mathbf{m} \mid \mathrm{D})$ durch $p(\mathbf{m} \mid \hat{\boldsymbol{\theta}})$ approximiert werden.

Mit dieser Approximation wird der Rechenaufwand des Verfahrens aus Abschnitt 4.2. erheblich reduziert, da insbesondere das Integral

$$p(\mathbf{m} \mid \mathrm{D}) = \int p(\mathbf{m} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathrm{D}) \mathrm{d}\boldsymbol{\theta}$$

nicht ausgewertet werden muss.

$$p(\mathbf{m} \mid \mathrm{D}) = \int p(\mathbf{m} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathrm{D}) \mathrm{d}\boldsymbol{\theta} \approx \int p(\mathbf{m} \mid \boldsymbol{\theta}) \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \mathrm{d}\boldsymbol{\theta} = p(\mathbf{m} \mid \hat{\boldsymbol{\theta}})$$

Im Folgenden werden die wichtigsten Techniken zur Bayes'schen Parameterschätzung gezeigt.

θ wird als Zufallsvariable angesehen und über eine WV beschrieben.

Ansatz 1: Minimaler quadratischer Schätzfehler

$$\mathrm{E}\{\|\hat{\boldsymbol{\theta}}(\mathrm{D})-\boldsymbol{\theta}\|^2\} = \int_{\mathbb{M}^N}\int_{\Theta} l(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\, p(\boldsymbol{\theta},\mathrm{D})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\mathrm{D} \overset{!}{\rightarrow} \text{minimal} \qquad \mathrm{d}\mathrm{D} := \prod_{i=1}^{N}\mathrm{d}\mathbf{m}_i$$

Kostenfunktion:

$$l(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) := (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathrm{T}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$$

$$\mathrm{E}\{\|\hat{\boldsymbol{\theta}}(\mathrm{D})-\boldsymbol{\theta}\|^2\} = \int_{\mathbb{M}^N}\int_{\Theta} (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathrm{T}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\, p(\boldsymbol{\theta}\,|\,\mathrm{D})\mathrm{d}\boldsymbol{\theta}\; p(\mathrm{D})\,\mathrm{d}\mathrm{D} \overset{!}{\rightarrow} \text{minimal}$$

$$\Leftrightarrow \quad I(\mathrm{D}) := \int_{\Theta} (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathrm{T}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\, p(\boldsymbol{\theta}\,|\,\mathrm{D})\mathrm{d}\boldsymbol{\theta} \text{ ist minimal.}$$

$$I(\mathrm{D}) = \int_{\Theta} \sum_{j=1}^{K} (\hat{\theta}_j(\mathrm{D})-\theta_j)^2\, p(\boldsymbol{\theta}\,|\,\mathrm{D})\mathrm{d}\boldsymbol{\theta}$$

4.3. Bayes'sche Parameterschätzung

Notwendige Bedingung für Minimum:

$$\frac{\partial I(\mathrm{D})}{\partial \hat{\theta}_i(\mathrm{D})} \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \int_{\Theta} 2(\hat{\theta}_i(\mathrm{D}) - \theta_i)\, p(\boldsymbol{\theta} \mid \mathrm{D})\, \mathrm{d}\boldsymbol{\theta} = 0$$

$$\Leftrightarrow \quad \hat{\theta}_i(\mathrm{D}) = \int_{\Theta} \theta_i\, p(\boldsymbol{\theta} \mid \mathrm{D})\, \mathrm{d}\boldsymbol{\theta}$$

Hinreichende Bedingung für Minimum:

$$\frac{\partial^2 I(\mathrm{D})}{\partial (\hat{\theta}_i(\mathrm{D}))^2} = 2 > 0$$

Schätzer mit minimalem mittleren quadratischen Fehler:

$$\hat{\boldsymbol{\theta}}(\mathrm{D}) = \int_{\Theta} \boldsymbol{\theta} p(\boldsymbol{\theta} \mid \mathrm{D})\, \mathrm{d}\boldsymbol{\theta} = \mathrm{E}\{\boldsymbol{\theta} \mid \mathrm{D}\}$$

mit $$p(\boldsymbol{\theta} \mid \mathrm{D}) = \frac{p(\mathrm{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathrm{D})}$$

A Posteriori Erwartungswert der Zufallsvariablen $\boldsymbol{\theta}$

Ansatz 2: Konstante Gewichtung großer Fehler

Kostenfunktion:

$$l(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) := \begin{cases} 0 & \text{für} \quad \left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right\| < \Delta, \;\; \Delta > 0 \\ 1 & \text{sonst} \end{cases}$$

$$\mathrm{E}\{l(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\} = \int_{\mathbb{M}^N} \int_{\Theta} l(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\, p(\boldsymbol{\theta},\mathrm{D})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{dD} \overset{!}{\rightarrow} \text{minimal}$$

$$\Leftrightarrow \qquad I(\mathrm{D}) = \int_{\Theta} l(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})\, p(\boldsymbol{\theta}\,|\,\mathrm{D})\,\mathrm{d}\boldsymbol{\theta} \qquad \text{ist minimal.}$$

$$\Leftrightarrow \qquad I(\mathrm{D}) = 1 - \int_{\{\boldsymbol{\theta}\,|\,\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|<\Delta\}} p(\boldsymbol{\theta}\,|\,\mathrm{D})\,\mathrm{d}\boldsymbol{\theta} \qquad \text{ist minimal.}$$

Für $\Delta \rightarrow 0$ folgt:

$$\hat{\boldsymbol{\theta}}(\mathrm{D}) = \arg\max_{\boldsymbol{\theta}}\{p(\boldsymbol{\theta}\,|\,\mathrm{D})\}$$

MAP-Schätzer

Ansatz 2: Konstante Gewichtung großer Fehler

Wegen der Monotonie der Logarithmusfunktion, kann man äquivalent die logarithmierte A Posteriori WDF maximieren; das bringt oft analytische Vorteile.

$$\hat{\boldsymbol{\theta}}(\mathrm{D}) = \arg\max_{\boldsymbol{\theta}} \{\ln\, p(\boldsymbol{\theta} \mid \mathrm{D})\}$$

MAP-Schätzer

Falls $\ln\, p(\boldsymbol{\theta}|\mathbf{m})$ differenzierbar ist, lautet die notwendige Bedingung für Maxima:

$$\nabla_{\boldsymbol{\theta}} \ln(\, p(\boldsymbol{\theta} \mid \mathrm{D})) = \sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}} \ln\, p(\mathbf{m}_k \mid \boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \ln\, p(\boldsymbol{\theta}) \stackrel{!}{=} \mathbf{0}$$

In vielen praktisch relevanten Fällen liefern MAP-Schätzung und der A Posteriori Erwartungswert das gleiche Ergebnis.

Details siehe z.B. in: K. Kroschel, „Statistische Informationstechnik“, Springer 2004

Fehler bei der Bayes'schen Klassifikation

- Bayes'scher Fehler: ergibt sich durch die Überlappung der Verteilungsdichten; spiegelt die Diskriminanz der Merkmale wider.

- Modellfehler: ergibt sich durch ein nicht passendes Modell.
  Auswahl eines passenden Modells mit Hilfe eines Anpassungstests für WVen, z.B. Chi-Quadrat-Test; Details in statistischer Grundlagenliteratur.

- Schätzfehler: ergibt sich aufgrund endlicher Datensätze.