Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Bayessches Lernen (II)

Christoph Sawade/Niels Landwehr
Jules Rasetaharison
Tobias Scheffer

Überblick

- Wahrscheinlichkeiten, Erwartungswerte, Varianz
- Grundkonzepte des Bayesschen Lernens
- (Bayessche) Parameterschätzung für Wahrscheinlichkeitsverteilungen
- Bayessche Lineare Regression, Naive Bayes

Parameter von Verteilungen schätzen

- Oft können wir annehmen, dass Daten einer bestimmten Verteilung folgen
  - Z.B. Binomialverteilung für *N* Münzwürfe
  - Z.B. Gaußverteilung für Körpergröße, IQ, …
- Diese Verteilungen sind parametrisiert
  - Binomialverteilung: Parameter μ ist Wahrscheinlichkeit für „Kopf“
  - Gaußverteilung: Parameter μ, σ für Mittelwert und Standardabweichung
- „Echte“ Wahrscheinlichkeiten/Parameter kennen wir nie.
- Welche Aussagen über echte Wahrscheinlichkeiten können wir machen, gegeben Daten?

Parameter von Verteilungen schätzen

- Problemstellung Parameter von Verteilungen schätzen:
  - Gegeben parametrisierte Familie von Verteilungen (z.B. Binomial, Gauß) mit Parametervektor θ
  - Gegeben Daten L
  - Gesucht: a-posteriori Verteilung $P(\theta \mid L)$ bzw. maximum a-posteriori Schätzung

$$\theta^* = \arg\max_\theta P(\theta \mid L)$$

- Verwende Bayessche Regel:

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

Binomialverteilte Daten Schätzen

- Beispiel: Münzwurf, schätze Parameter $\mu = \theta$
  - N Mal Münze werfen.
  - Daten L: N_k mal Kopf, N_z mal Zahl.

- Beste Schätzung θ gegeben L? Bayes‘ Gleichung:

Likelihood der Daten gegeben Parameter, wie gut erklärt Parameter die Beobachtungen?

A-priori Verteilung über Parameter, repräsentiert Vorwissen

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

A-posteriori Verteilung über Parameter, charakterisiert wahrscheinliche Parameterwerte und verbleibende Ungewissheit

Binomialverteilte Daten Schätzen

- Likelihood der Daten:

$$P(L \mid \theta)$$

($\theta = \mu$ Wahrscheinlichkeit für „Kopf“)

- Likelihood ist binomialverteilt:

$$P(L \mid \theta) = P(N_k, N_z \mid \theta) = \mathrm{Bin}(N_k \mid N, \theta) \qquad N = N_k + N_z$$

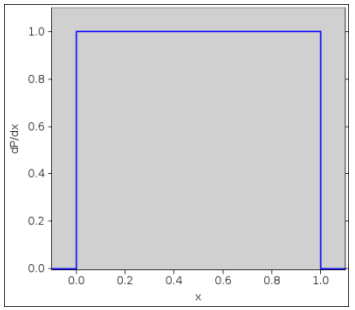
$$= \binom{N_k + N_z}{N_k} \theta^{N_k} (1-\theta)^{N_z}$$

Wahrscheinlichkeit, bei N Münzwürfen N_k-mal Kopf und N_Z-mal Zahl zu sehen, für Münzparameter θ

Binomialverteilte Daten Schätzen

- Was ist der Prior $P(\theta)$ im Münzwurfbeispiel?
- 1) Versuch: Kein Vorwissen

$$P(\theta) = \begin{cases} 1 : 0 \le \theta \le 1 \\ 0 : sonst \end{cases} \qquad \text{Dichte}$$



- Beispiel:
  - Daten L = {Zahl,Zahl,Zahl}
  - MAP Modell:

$$\theta^* = \arg\max_{\theta\in[0,1]} P(\theta \mid L) = \arg\max_{\theta\in[0,1]} \frac{P(L \mid \theta)P(\theta)}{P(L)}$$

$$= \arg\max_{\theta\in[0,1]} P(L \mid \theta) = \arg\max_{\theta\in[0,1]} \binom{3}{0} \theta^0 (1-\theta)^3 = 0$$

- Schlussfolgerung: Münze wird **niemals** „Kopf“ zeigen
  - Schlecht, Überanpassung an Daten („Overfitting“)

Binomialverteilte Daten Schätzen

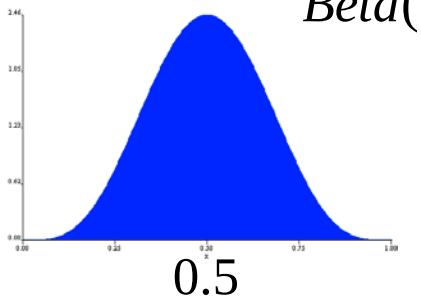
- Was ist der Prior $P(\theta)$ im Münzwurfbeispiel?
- Besser mit Vorwissen: Unwahrscheinlich, dass Münze immer Kopf oder immer Zahl zeigt
- Gutes Modell für Vorwissen über θ: Beta-Verteilung.

$$P(\theta) = Beta(\theta \mid \alpha_k, \alpha_z)$$

$$= \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1}(1-\theta)^{\alpha_z - 1}$$

$(\theta \in [0,1])$

$Beta(\theta \mid 5,5)$



0.5

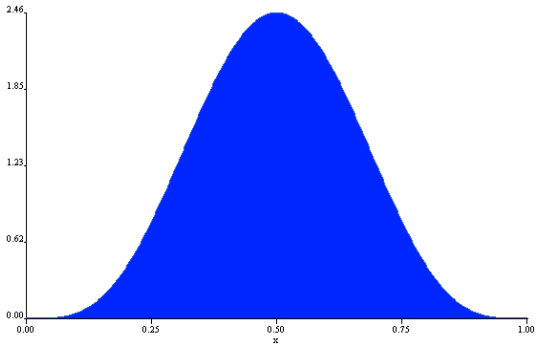
- Gamma-Funktion $\Gamma(\alpha)$ kontinuierliche Fortsetzung der Fakultätsfunktion

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \qquad \forall n \in \mathbb{N} : \Gamma(n) = (n-1)!$$

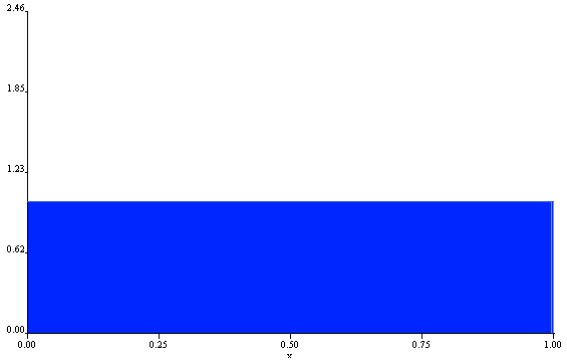
Binomialverteilte Daten Schätzen

- α_k und α_z sind Parameter der Beta-Verteilung („Hyperparameter“)
- Beta-Verteilung ist Verteilung über Verteilungen

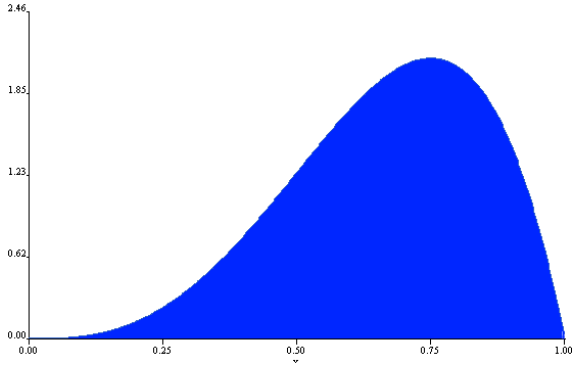
$\alpha_K = 5, \quad \alpha_Z = 5$



$\alpha_K = 1, \quad \alpha_Z = 1$



$\alpha_K = 4, \quad \alpha_Z = 2$



- Normalisierte Dichte $\int_0^1 Beta(\theta \mid \alpha_K, \alpha_Z) d\theta = 1$

Binomialverteilte Daten Schätzen

- Warum gerade diese a-priori-Verteilung?
- Strukturelle Ähnlichkeit mit Likelihood:

Prior $$P(\theta) = Beta(\theta \mid \alpha_k, \alpha_z) = \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1}(1-\theta)^{\alpha_z - 1}$$

Likelihood $$P(L \mid \theta) = P(N_k, N_z \mid \theta) = \binom{N_k + N_Z}{N_k} \theta^{N_k}(1-\theta)^{N_z}$$

- Einfach, Beobachtungen zu berücksichtigen: Produkt aus Likelihood und Prior hat wieder dieselbe Form wie Prior

$$P(\theta \mid L) \propto P(L \mid \theta) P(\theta)$$

Binomialverteilte Daten Schätzen

- Wenn wir den Beta-Prior in Bayes‘ Gleichung einsetzen, dann:

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

$$= \frac{1}{Z} \mathrm{Bin}(N_K \mid N, \theta) Beta(\theta \mid \alpha_k, \alpha_z)$$

$$= \frac{1}{Z} \binom{N_k + N_z}{N_k} \theta^{N_k} (1-\theta)^{N_z} \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1} (1-\theta)^{\alpha_z - 1}$$

$$= \frac{1}{Z'} \theta^{\alpha_k + N_k - 1} (1-\theta)^{\alpha_z + N_z - 1}$$

$$= ?$$

Binomialverteilte Daten Schätzen

- Wenn wir den Beta-Prior in Bayes‘ Gleichung einsetzen, dann:

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

$$= \frac{1}{Z} \mathrm{Bin}(N_K \mid N, \theta) Beta(\theta \mid \alpha_k, \alpha_z)$$

$$= \frac{1}{Z} \binom{N_k + N_z}{N_k} \theta^{N_k} (1-\theta)^{N_z} \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1} (1-\theta)^{\alpha_z - 1}$$

$$= \frac{1}{Z'} \theta^{\alpha_k + N_k - 1} (1-\theta)^{\alpha_z + N_z - 1}$$

$$= \frac{\Gamma(\alpha_k + N_k + \alpha_z + N_z)}{\Gamma(\alpha_k + N_k)\Gamma(\alpha_z + N_z)} \theta^{\alpha_k + N_k - 1} (1-\theta)^{\alpha_z + N_z - 1}$$

$$= Beta(\theta \mid \alpha_k + N_k, \alpha_z + N_z)$$

- **Beta-Verteilung ist „konjugierter“ Prior: Posterior ist wieder Beta-verteilt**

Zusammenfassung Bayessche Parameterschätzung Binomialverteilung

- Bayessche Regel

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

- Posterior $P(\theta \mid L)$: Wie wahrscheinlich ist Modell θ, nachdem wir Daten L gesehen haben?

- Vorwissen $P(\theta)$ und Evidenz der Trainingsdaten L werden zu neuem Gesamtwissen $P(\theta \mid L)$ integriert.

- Beispiel Münzwurf: Vorwissen Beta($\theta \mid \alpha_k, \alpha_z$) und Beobachtungen N_k, N_z werden zu Posterior Beta($\theta \mid \alpha_k + N_k, \alpha_z + N_z$).

Münzwurf: Wahrscheinlichste Wahrscheinlichkeit

- Wahrscheinlichster Parameter θ.

$$\arg\max_\theta P(\theta \mid L) = \arg\max_\theta Beta(\theta \mid \alpha_k + N_k, \alpha_z + N_z)$$

$$= \arg\max_\theta \frac{\Gamma(\alpha_k + \alpha_z + N_k + N_z)}{\Gamma(\alpha_k + N_k)\Gamma(\alpha_z + N_z)} \theta^{\alpha_k + N_k - 1}(1-\theta)^{\alpha_z + N_z - 1}$$

Ableiten, Ableitung null setzen ($\alpha_z \geq 1,\ \alpha_k \geq 1$)

$$= \frac{N_k + \alpha_k - 1}{N_k + N_z + \alpha_k + \alpha_z - 2}$$

Normalisierer, unabhängig von θ

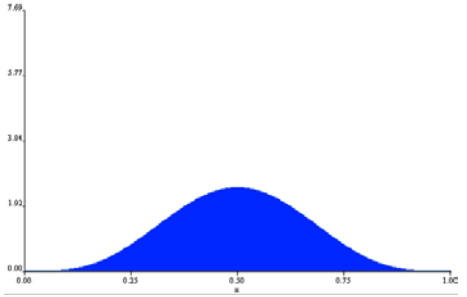
  - Für $\alpha_z = \alpha_k = 1$ ergibt sich ML Schätzung

- Interpretation der Hyperparameter $\alpha_z - 1 / \alpha_k - 1$:
  - $\alpha_z - 1 / \alpha_k - 1$ „Pseudocounts“ , die auf beobachtete „Counts“ N_z / N_k aufgeschlagen werden
  - wie oft im Leben Münzwurf mit „Kopf“/“Zahl“ gesehen?

Münzwurf: Wahrscheinlichste Wahrscheinlichkeit

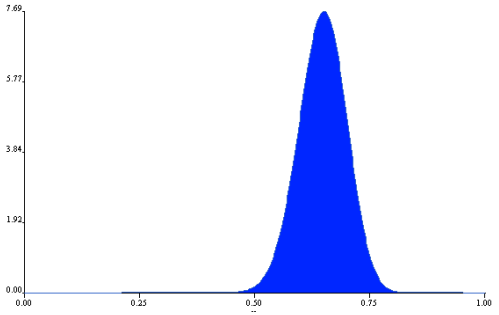
- Beispiel MAP Schätzung Parameter

Prior $P(\theta) = Beta(\theta \mid 5,5)$



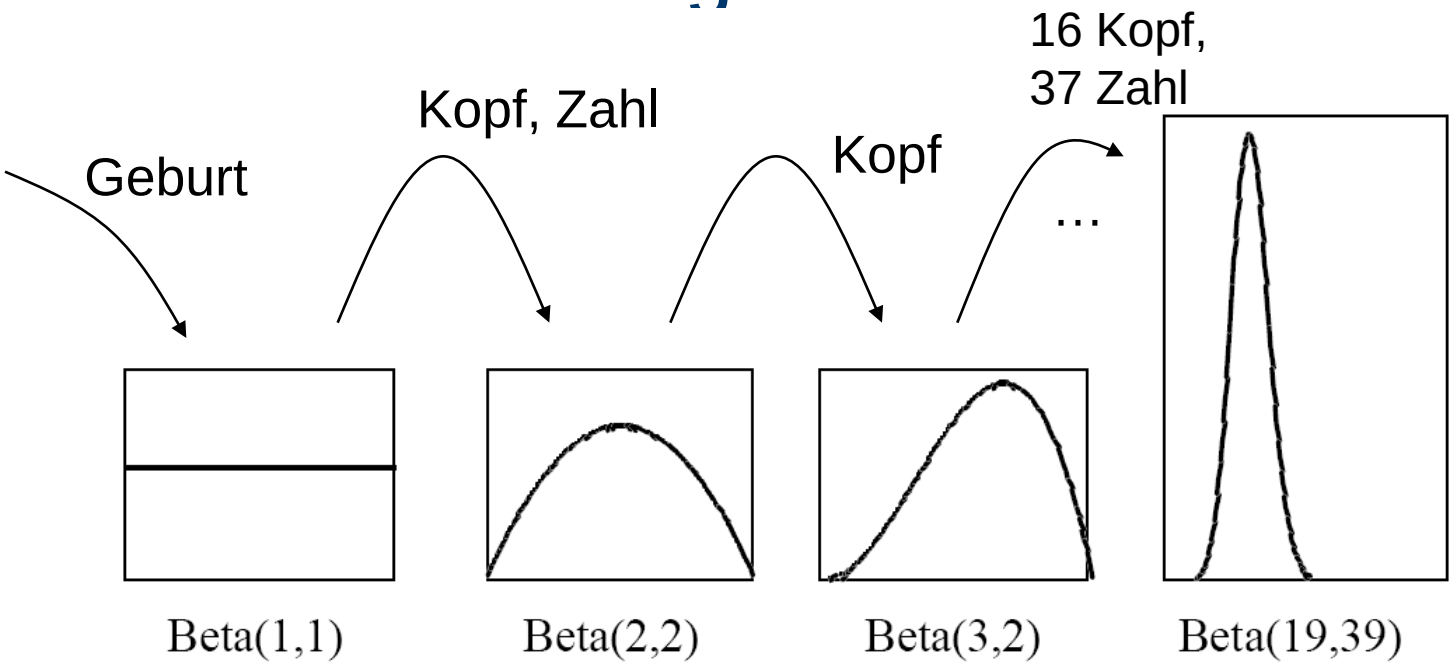
Posterior nach 50x Kopf, 25x Zahl:

$P(\theta \mid N_K = 50, N_Z = 25) = Beta(\theta \mid 55,30)$



MAP Schätzung: $\theta^* = \arg\max_\theta P(\theta \mid N_K = 50, N_Z = 25) = \frac{54}{54+29} \approx 0.65$

Bayessche Schätzung als Sequentielles Update der Verteilung



$$\underbrace{Beta(\theta \mid \alpha_k + N_k, \alpha_z + N_z)}_{\text{Posterior}} = \frac{1}{Z} \underbrace{\binom{N_K + N_Z}{N_K} \theta^{N_k} (1-\theta)^{N_z}}_{\text{Likelihood}} \underbrace{Beta(\theta \mid \alpha_k, \alpha_z)}_{\text{Prior}}$$

Verallgemeinerung: Würfelwurf statt Münzwurf

- Münzwurf: 2 Ausgänge.
  - Prior Beta-verteilt,
  - Binomiale Likelihood,
  - Posterior wieder Beta-verteilt.
  - Modell für Prozesse mit binärem Ergebnis.
- Verallgemeinerung Würfelwurf: *k* Ausgänge.
  - Prior Dirichlet-verteilt,
  - Likelihood Multinomial,
  - Posterior wieder Dirichlet-verteilt.
  - Modell für diskrete Prozesse mit mehreren möglichen Ergebnissen

Einschub: Begriff „Schätzer“

- Wir haben uns mit der Schätzung von Parametern von Verteilungen aus Daten beschäftigt

- Formalisierung: ein **Schätzer** ist ein Verfahren, das Beobachtungen L auf einen Schätzwert abbildet.
  - z.B. Münzwurf: Beobachtung N_k, N_z, schätze Münzparameter
  - Schätzer für (unbekannten) Wert θ wird mit $\hat{\theta}$ bezeichnet

- Schätzer ist Zufallsvariable, Verteilung bestimmt durch die Verteilung $p(L \mid \theta)$ der Daten gegeben den echten Parameter

- Schätzer heißt **erwartungstreu**, wenn $E[\hat{\theta}] = \theta$

Schätzer

- Beispiel: Münzwurf, Beobachtung N_k, N_z.
- MAP-Schätzer Münzwurf:
  - $$\hat{\theta}_{MAP} = \arg\max_\theta P(\theta \mid L)$$
    $$= \arg\max_\theta Beta(\theta \mid \alpha_k + N_k, \alpha_z + N_z)$$
    $$= \frac{\alpha_k + N_k - 1}{\alpha_k + N_k + \alpha_z + N_z - 2}$$
- ML-Schätzer Münzwurf:
  - $$\hat{\theta}_{ML} = \arg\max_\theta P(L \mid \theta)$$
    $$= \arg\max_\theta \theta^{N_k}(1-\theta)^{N_z}$$
    $$= \frac{N_k}{N_k + N_z}$$

Schätzer

- Maximum Likelihood Schätzer erwartungstreu:
  - Angenommen echter Münzparameter ist θ
  - Dann

$$\mathbb{E}\left[\hat{\theta}_{ML}\right] = \mathbb{E}\left[\frac{N_K}{N}\right] = \frac{1}{N}\mathbb{E}\left[N_K\right] = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[X_i\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\theta$$

$$= \theta$$

Erwartungswert additiv

Erwartungswert über mögliche beobachtete Münzwürfe

„Kopf“ Indikator für einzelnen Münzwurf

- MAP Schätzer nicht erwartungstreu:

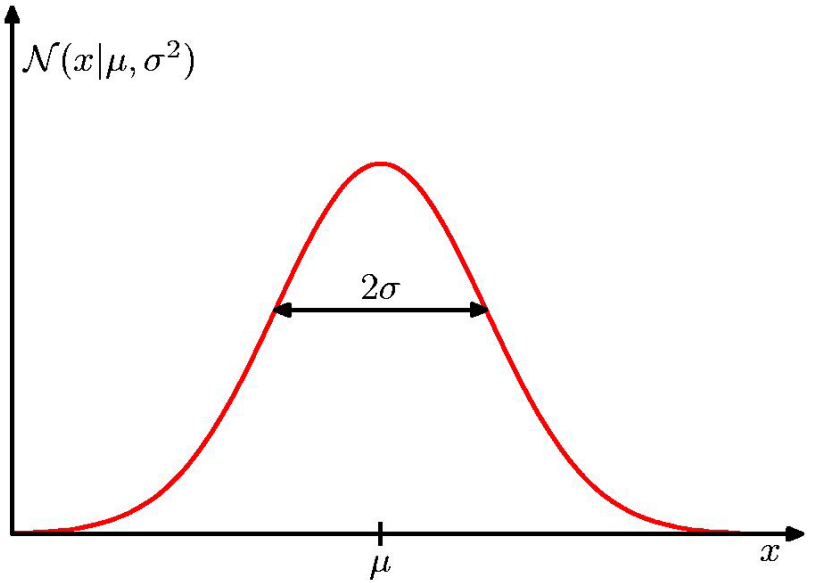
$$\mathbb{E}\left[\hat{\theta}_{MAP}\right] = \frac{N\mu + \alpha_K - 1}{N + \alpha_K + \alpha_Z - 2}$$

Schätzen Kontinuierlicher Daten: Normalverteilung

- Normalverteilung häufige Wahl zur Modellierung kontinuierlicher ZV
- Hier: eindimensionale Daten, univariate Normalverteilung
  - Mittelwert-Parameter μ
  - Varianz-Parameter σ^2

Dichtefunktion:

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$



Normalverteilte Daten Schätzen: ML

- Schätzen einer Normalverteilung aus Daten
  - Annahme: Daten folgen Normalverteilung
  - Aber Mittelwert μ und Standardabweichung σ unbekannt

- Gegeben: Daten *L* bestehend aus *n* unabhängigen Datenpunkten

$$x_1, ..., x_n \qquad x_i \sim \mathcal{N}(x \mid \mu, \sigma^2) \qquad \text{unabhängig gezogen}$$

- Gesucht: Schätzungen $\hat{\mu}$, $\hat{\sigma}$ für die unbekannten Parameter μ, σ

Normalverteilte Daten Schätzen: ML

- Einfachster Ansatz: Maximum Likelihood, finde

$$(\hat{\mu},\hat{\sigma}) = \arg\max_{\mu,\sigma} p(L|\mu,\sigma)$$

- Berechnen der Likelihood

$$p(L \mid \mu,\sigma) = p(x_1,...,x_n|\mu,\sigma)$$

$$= \prod_{i=1}^{n} p(x_i|\mu,\sigma)$$ Datenpunkte unabhängig

$$= \prod_{i=1}^{n} \mathcal{N}(x_i|\mu,\sigma^2)$$ Verteilungsannahme einsetzen

$$= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Normalverteilte Daten Schätzen: ML

- Logarithmieren:

$$\arg\max_{\mu,\sigma} p(L|\mu,\sigma) = \arg\max_{\mu,\sigma} \log p(L|\mu,\sigma)$$

- Log-Likelihood:

$$\begin{aligned}
\log p(L|\mu,\sigma) &= \log\left(\prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right) \\
&= \log\left((2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

Normalverteilte Daten Schätzen: ML

- Log-Likelihood

$$\log p(L|\mu,\sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

- Maximierung über μ : betrachte partielle Ableitung

$$\frac{\partial}{\partial\mu}\log p(L|\mu,\sigma) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu)$$

$$= \frac{1}{\sigma^2}\underbrace{\left(\sum_{i=1}^{n}x_i - n\mu\right)}_{\text{Null setzen}}$$

- Null setzen:

$$\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

Intuitiv: geschätzter Mittelwert = Durchschnitt

Normalverteilte Daten Schätzen: ML

- Log-Likelihood

$$\log p(L|\mu,\sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

- Maximierung über σ^2: betrachte partielle Ableitung

$$\frac{\partial}{\partial\sigma^2}\log p(x_1,...,x_n|\hat{\mu},\sigma) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

$$= \frac{1}{2\sigma^2}\underbrace{\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 - n\right)}_{\text{Null setzen}}$$

- Null setzen:

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

Intuitiv: geschätzte Varianz = durchschnittliche Abweichung vom Mittelwert

Normalverteilte Daten Schätzen: ML

- Mittelwert-Schätzer für Normalverteilung erwartungstreu?

$$\begin{aligned}\mathbb{E}\left[\hat{\mu}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[x_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu\end{aligned}$$

- Schätzer $\hat{\mu}$ erwartungstreu

Normalverteilte Daten Schätzen: ML

- Varianz-Schätzer für Normalverteilung erwartungstreu?

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2\right]$$

$$= \ldots$$

$$= \frac{n-1}{n}\sigma^2$$

- Schätzer $\hat{\sigma}^2$ nicht erwartungstreu – Varianz wird systematisch unterschätzt

- Schätzer ist aber *konsistent* – der systematische Fehler verschwindet für $n \to \infty$

Normalverteilte Daten Schätzen: Beispiel ML Schätzung

- Wir wollen IQ einer Population schätzen
  - IQ typischerweise normalverteilt mit $\mu_0 = 100, \quad \sigma_0 = 15$
  - Wir wollen IQ-Verteilung schätzen für Subpopulation
  - Wohl auch normalverteilt, aber evtl andere Parameter

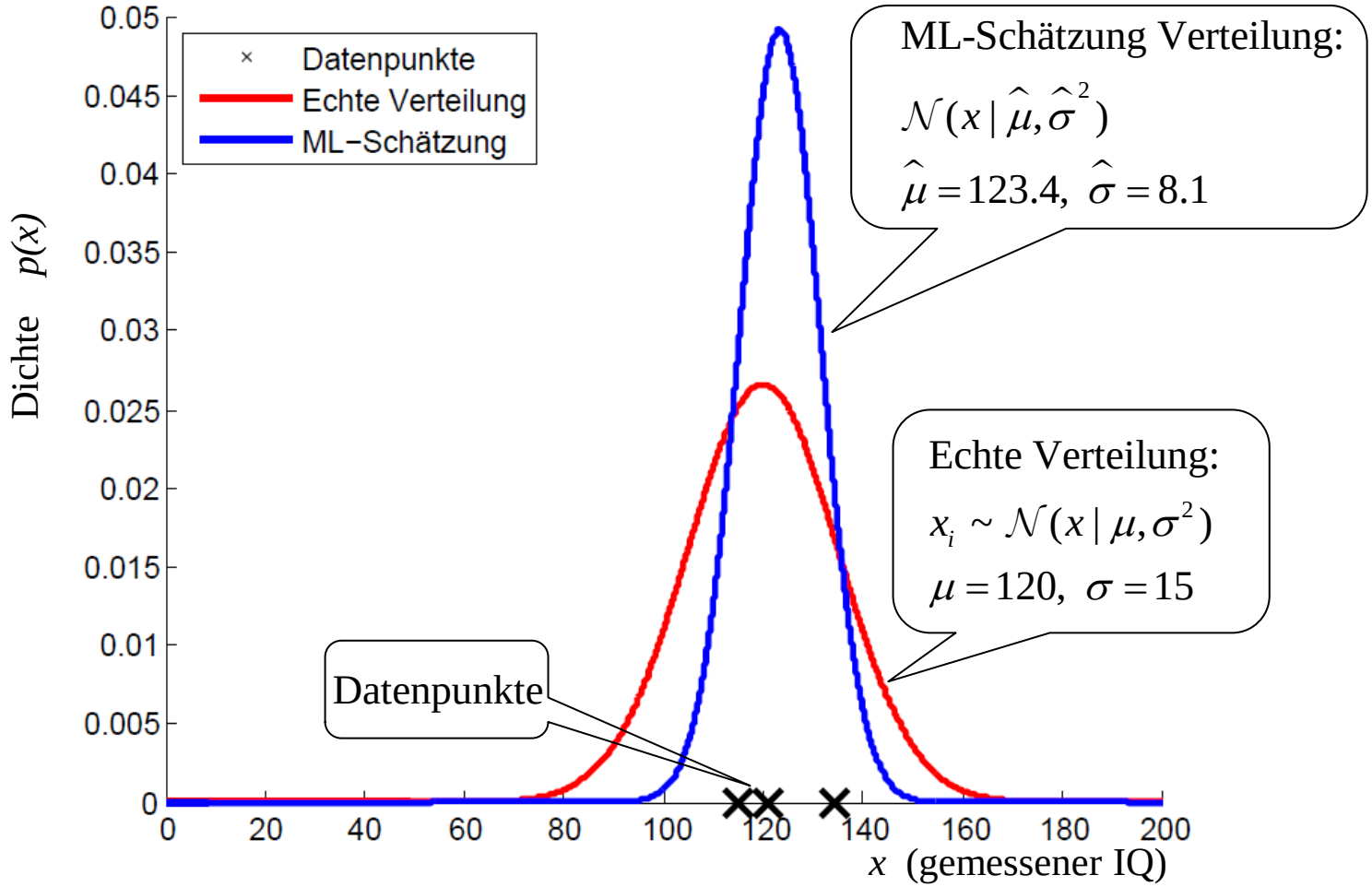
- Intelligenztest mit *n* Probanden: ergibt *n* unabhängige Datenpunkte $x_1, ..., x_n$
- Annahme: Normalverteilung mit unbekanntem Mittelwert und unbekannter Varianz

$$x_i \sim \mathcal{N}(x \mid \mu, \sigma^2)$$

- Maximum-Likelihood Schätzung $\hat{\mu}, \hat{\sigma}$

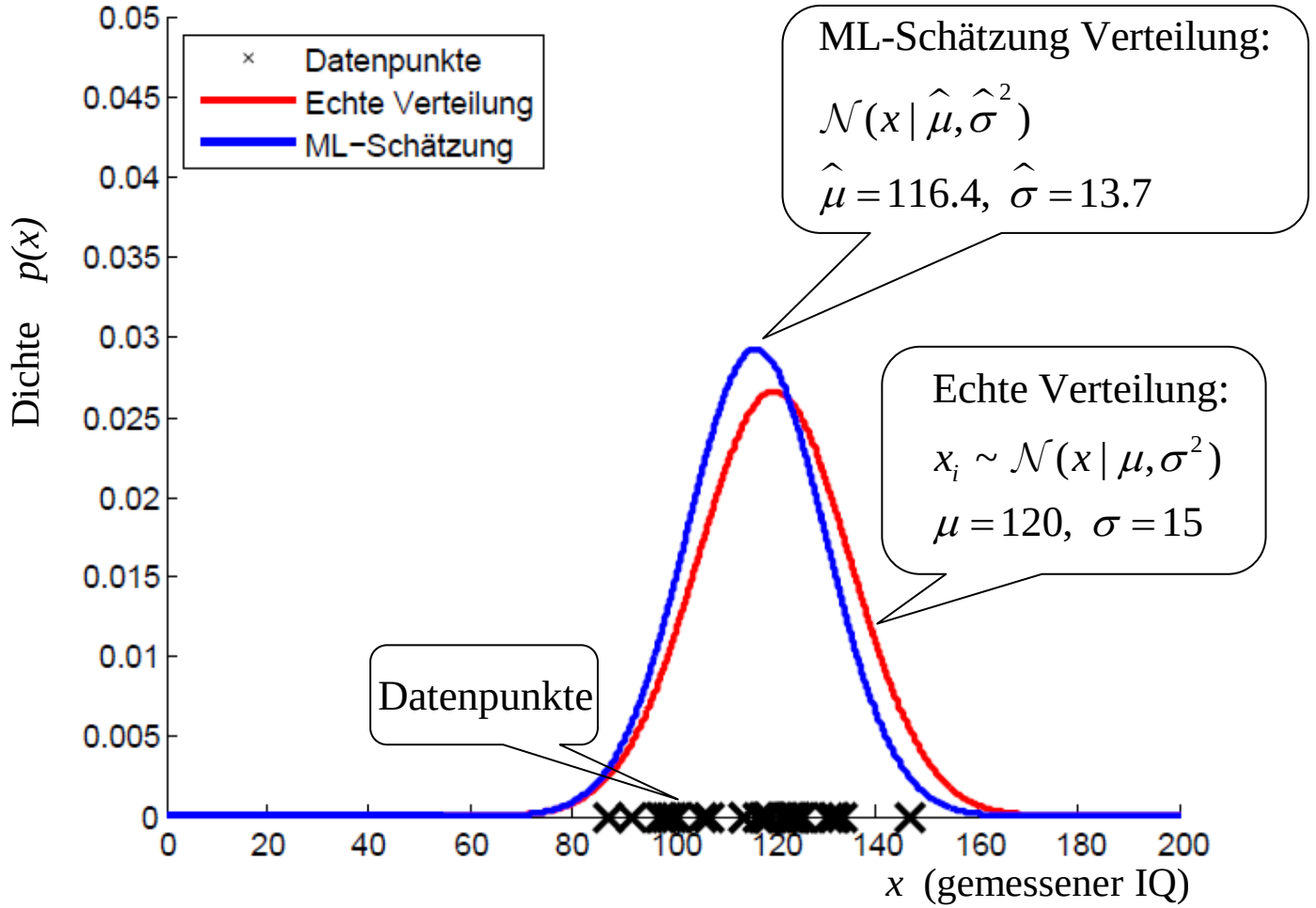
Normalverteilte Daten Schätzen: Beispiel ML Schätzung

- Simulation: *n=3* Punkte ziehen aus echter Verteilung mit $\mu = 120, \sigma = 15$, ML Parameter schätzen



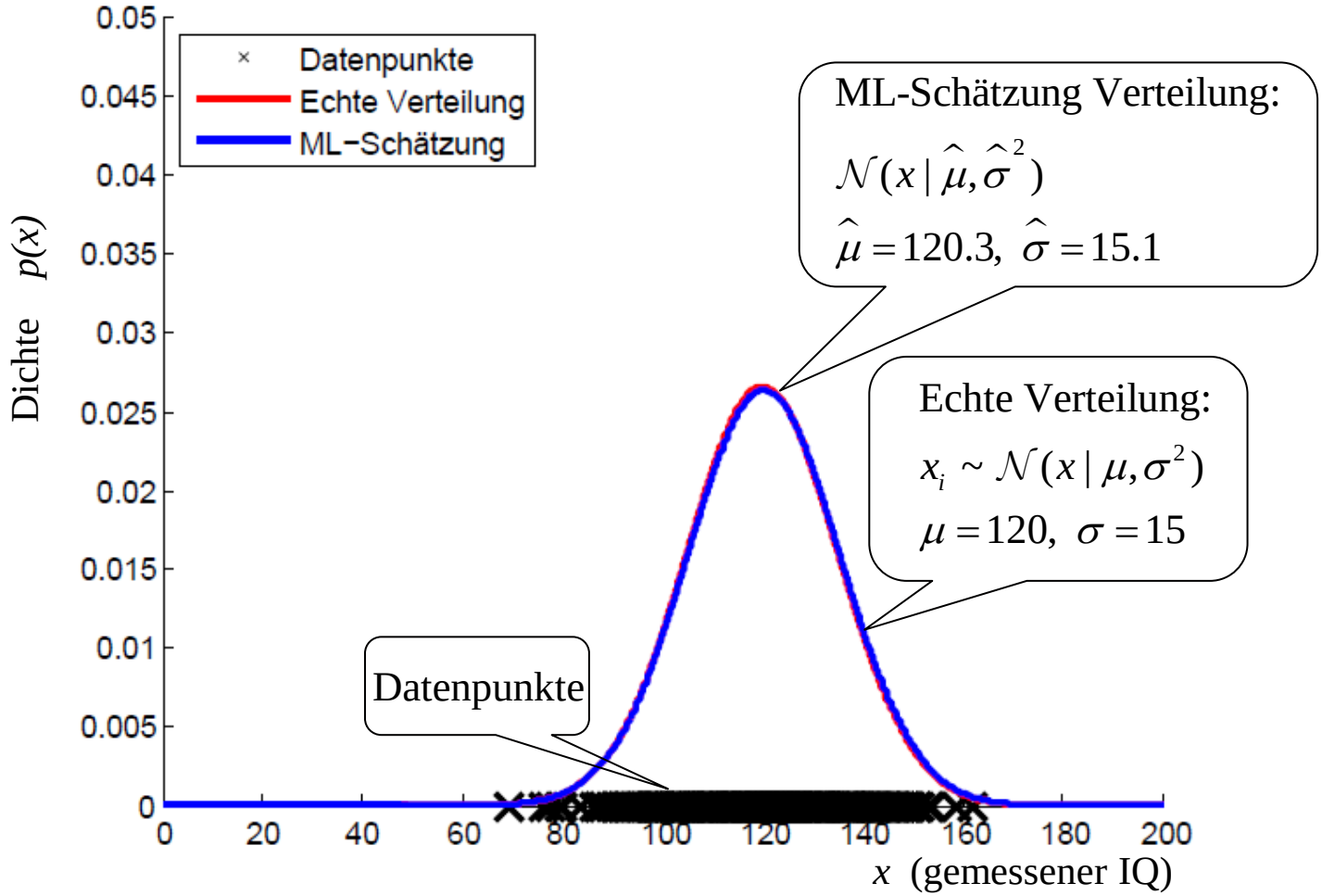
Normalverteilte Daten Schätzen: Beispiel ML Schätzung

- Simulation: *n=30* Punkte ziehen aus echter Verteilung mit $\mu = 120, \sigma = 15$, ML Parameter schätzen



Normalverteilte Daten Schätzen: Beispiel ML Schätzung

- Simulation: *n=500* Punkte ziehen aus echter Verteilung mit $\mu = 120, \sigma = 15$, ML Parameter schätzen



Normalverteilte Daten Schätzen: Bayessche Schätzungen

- Bisher nur ML-Schätzung
- Bayessche Schätzungen für Parameter μ, σ ?
  - Brauchen geeignete a-priori Verteilung
  - Im Allgemeinen gemeinsame a-priori Verteilung $p(\mu, \sigma)$

- Zunächst einfacher Fall:
  - Varianz σ bekannt
  - Schätzung des Mittelwertes $\hat{\mu}$ mit Prior $p(\mu)$

Normalverteilte Daten Schätzen: Bayessche Schätzungen

- Konjugierter Prior zur Normalverteilung mit bekannter Varianz ist Normalverteilung

Prior: $p(\mu) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2)$

Vermuteter Mittelwert

Wie stark ist Vorwissen?

Likelihood: $p(x_1, ..., x_n \mid \mu) = \prod_{i=1}^{n} \mathcal{N}(x_i \mid \mu, \sigma^2)$

Posterior: $p(\mu \mid L) = \frac{p(L \mid \mu)\, p(\mu)}{p(L)} = \mathcal{N}(\mu \mid \mu_n, \sigma_n^2)$

Posterior wieder normalverteilt!

mit $\mu_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \hat{\mu}_{ML}$, $\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}$

ML-Schätzung

Normalverteilte Daten Schätzen: Bayessche Schätzungen

- Weder Mittelwert noch Varianz ist bekannt: geeigneter konjugierter Prior ist Normal-Gamma
  - Definiere

$$\lambda = \frac{1}{\sigma} \qquad \text{"Precision"}$$

  - Konjugierter Prior ist Produkt aus Normalverteilung und Gamma-Verteilung:

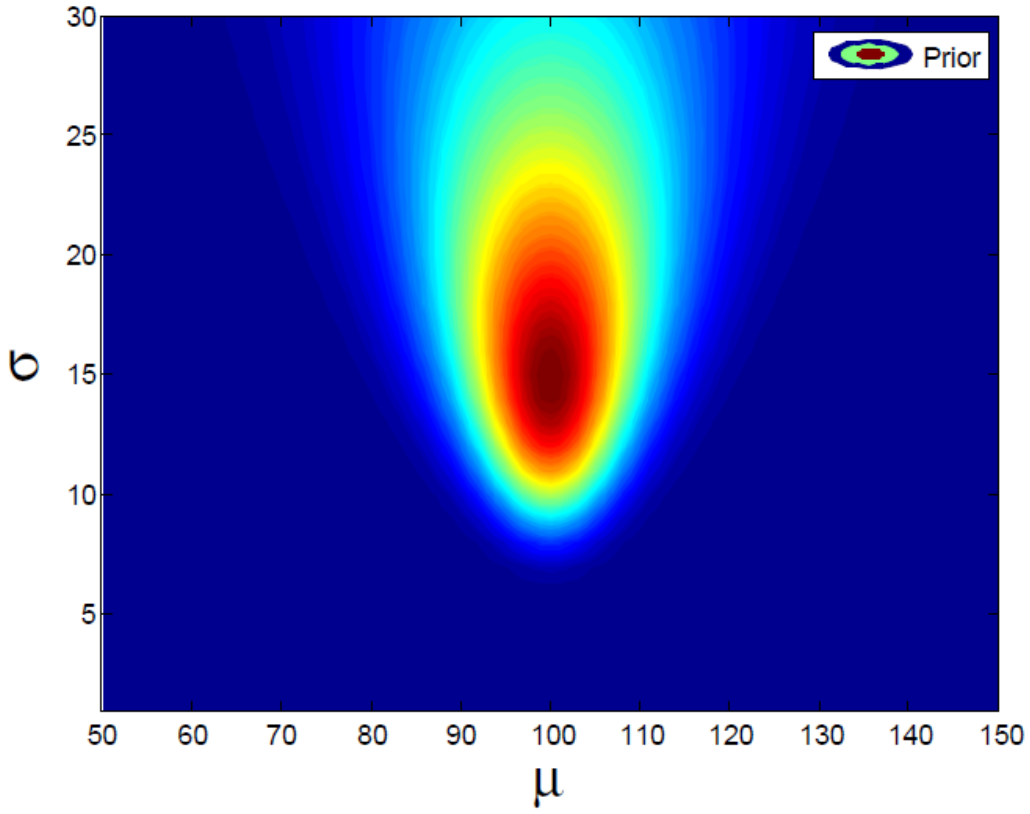
$$p(\mu, \lambda) = \mathcal{N}(\mu \mid \mu_0, (\beta\lambda)^{-1}) \, \mathrm{Gam}(\lambda \mid a, b)$$

$$\text{mit} \quad \mathrm{Gam}(\lambda \mid a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

  - Posterior $p(\mu, \lambda \mid x_1, ..., x_n)$ ist wieder Normal-Gamma

Normalverteilte Daten Schätzen: Beispiel Bayessche Schätzung

- Zurück zum Beispiel: schätzen der IQ-Verteilung anhand von *n* unabhängigen Datenpunkten
- Normal-Gamma Prior: erwarte $\mu \approx 100, \quad \sigma \approx 15$



Farbkodierung

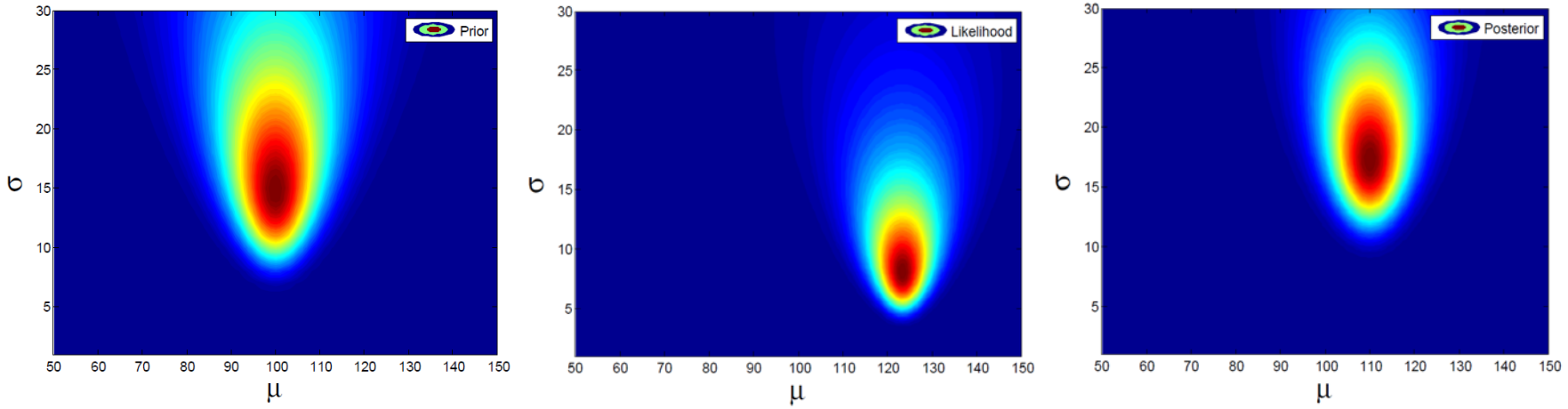
Dichte $p(\mu, \sigma)$

Erwartung:

$\mu \approx 100, \quad \sigma \approx 15$

Normalverteilte Daten Schätzen: Beispiel Bayessche Schätzung

- Simulation: *n=3* Punkte ziehen aus echter Verteilung mit $\mu = 120, \sigma = 15$, statt ML-Schätzung berechnen wir Posterior



Prior:

$$\mathcal{N}(\mu \mid \mu_0, (\beta\lambda)^{-1})\mathrm{Gam}(\lambda \mid a, b)$$

Likelihood:

$$\prod_{i=1}^{n} \mathcal{N}(x_i \mid \mu, \sigma^2)$$

Posterior:

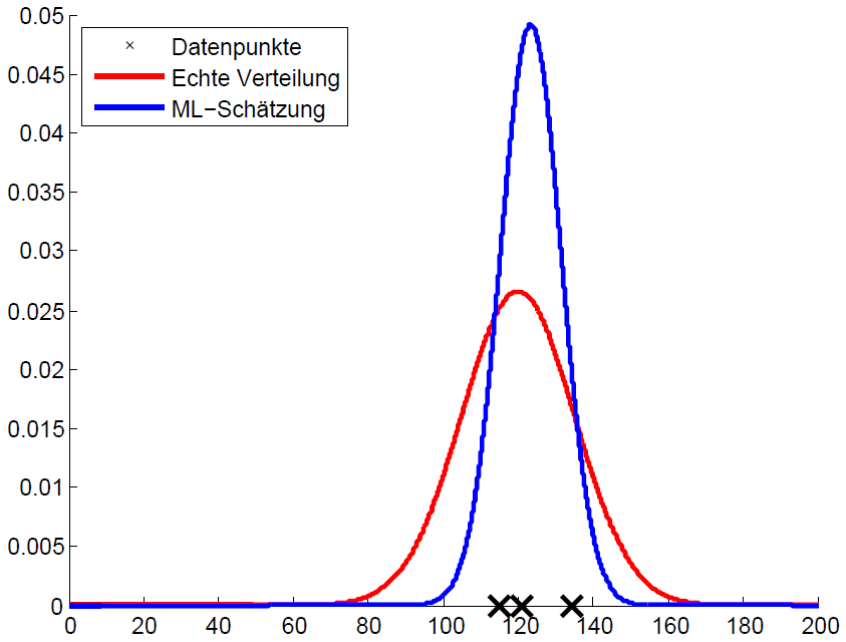
$$\mathcal{N}(\mu \mid \mu_0^*, (\beta^*\lambda)^{-1})\mathrm{Gam}(\lambda \mid a^*, b^*)$$

- Prior bewirkt Korrektur der ML-Schätzung in Richtung des Vorwissens

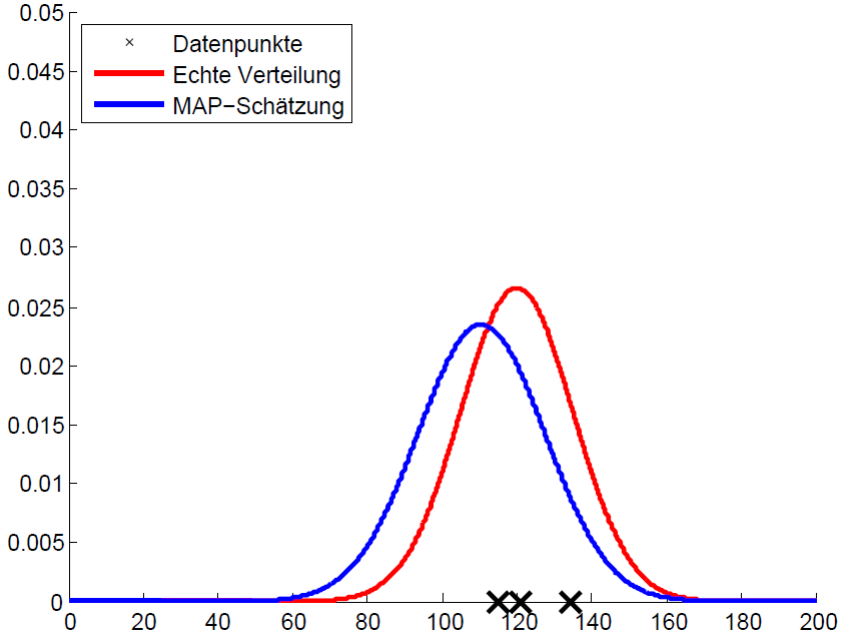
Normalverteilte Daten Schätzen: Beispiel MAP Parameter

- Simulation für *n=3:* Vergleich ML und MAP Lösung

ML

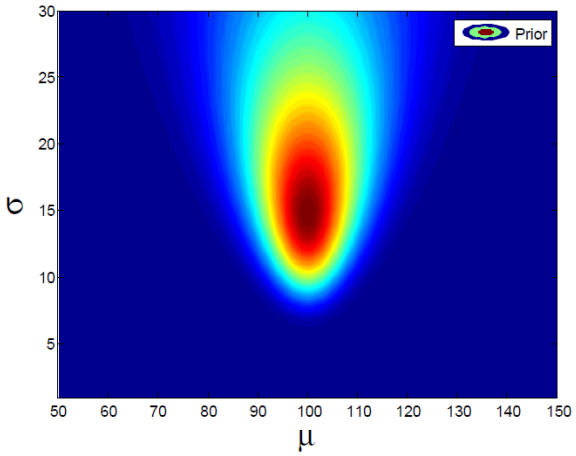


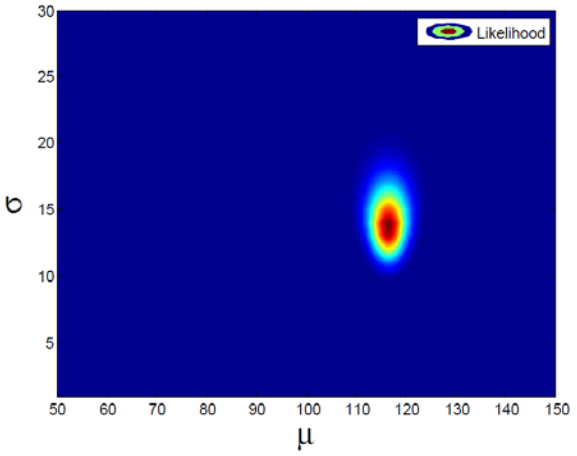
MAP

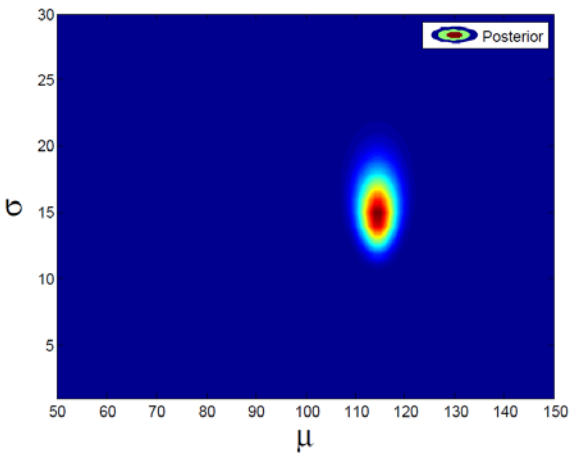


Normalverteilte Daten Schätzen: Beispiel Bayessche Schätzung

- Simulation: *n=30* Punkte ziehen aus echter Verteilung, statt ML-Schätzung berechnen wir a posteriori Verteilung







Prior:

$$\mathcal{N}(\mu \mid \mu_0, (\beta\lambda)^{-1})\,\mathrm{Gam}(\lambda \mid a, b)$$

Likelihood:

$$\prod_{i=1}^{n} \mathcal{N}(x_i \mid \mu, \sigma^2)$$

Posterior:

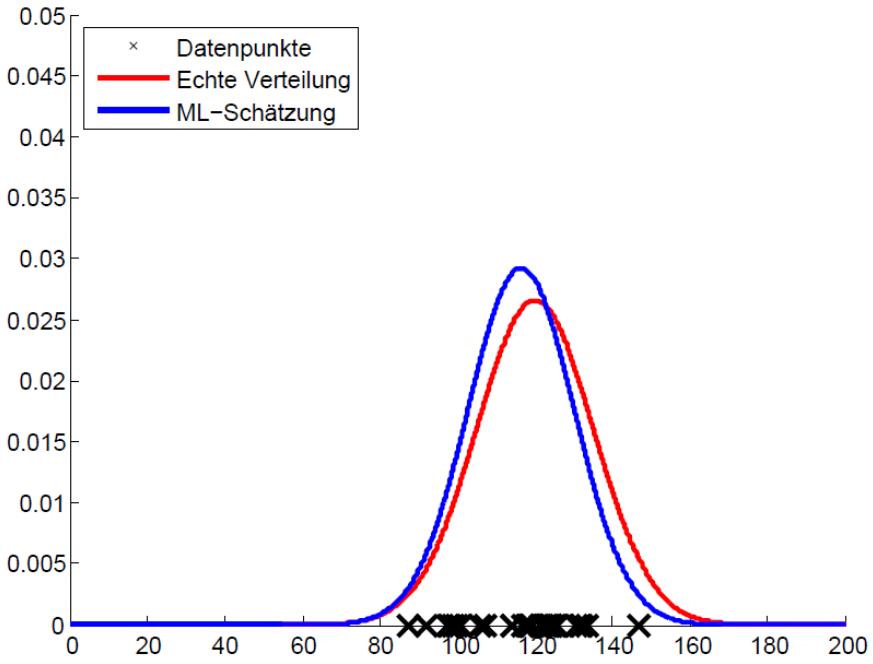
$$\mathcal{N}(\mu \mid \mu_0^*, (\beta^*\lambda)^{-1})\,\mathrm{Gam}(\lambda \mid a^*, b^*)$$

- Prior bewirkt Korrektur der ML-Schätzung in Richtung des Vorwissens

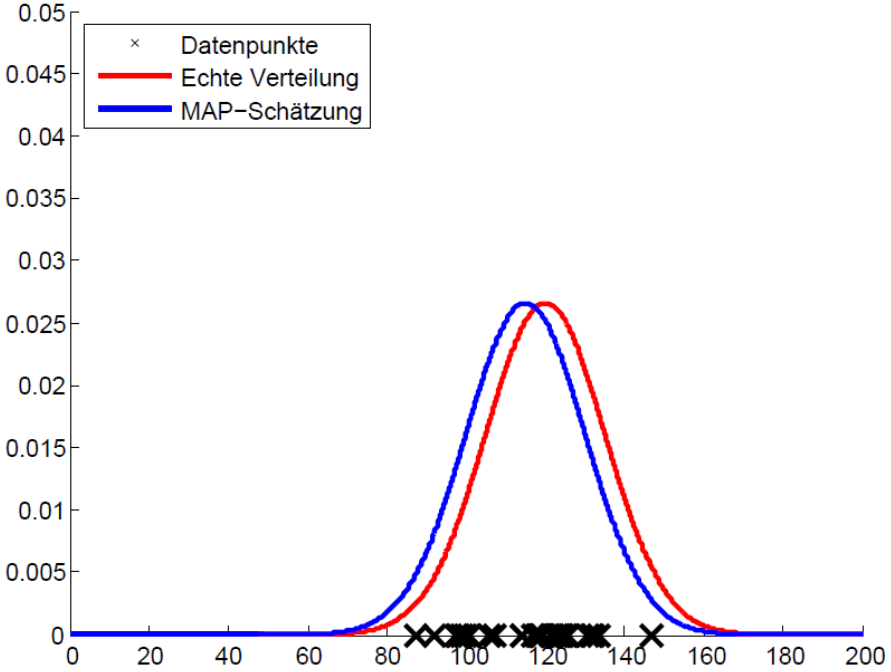
Normalverteilte Daten Schätzen: Beispiel MAP Parameter

- Simulation für *n=30:* Vergleich ML und MAP Lösung

ML

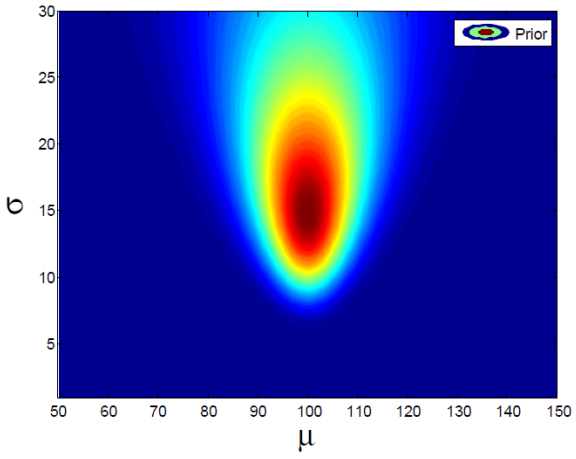


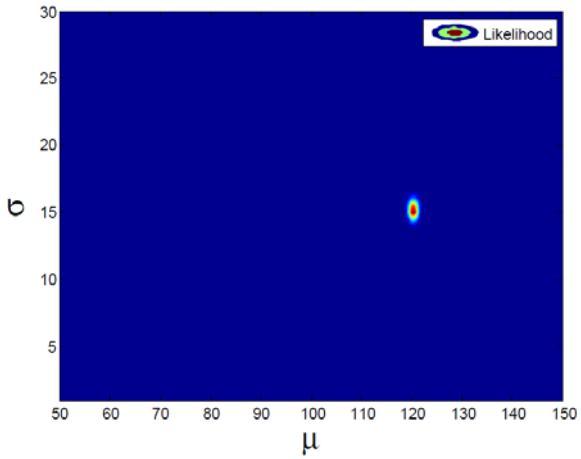
MAP

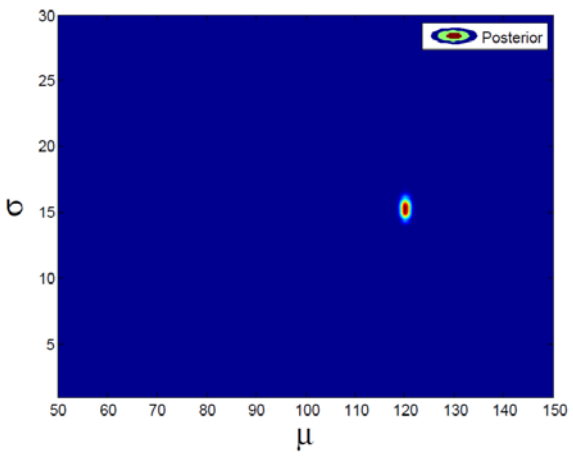


Normalverteilte Daten Schätzen: Beispiel Bayessche Schätzung

- Simulation: *n=500* Punkte ziehen aus echter Verteilung, statt ML-Schätzung berechnen wir a posteriori Verteilung







Prior:

$$\mathcal{N}(\mu \mid \mu_0, (\beta\lambda)^{-1})\,\mathrm{Gam}(\lambda \mid a, b)$$

Likelihood:

$$\prod_{i=1}^{n} \mathcal{N}(x_i \mid \mu, \sigma^2)$$

Posterior:

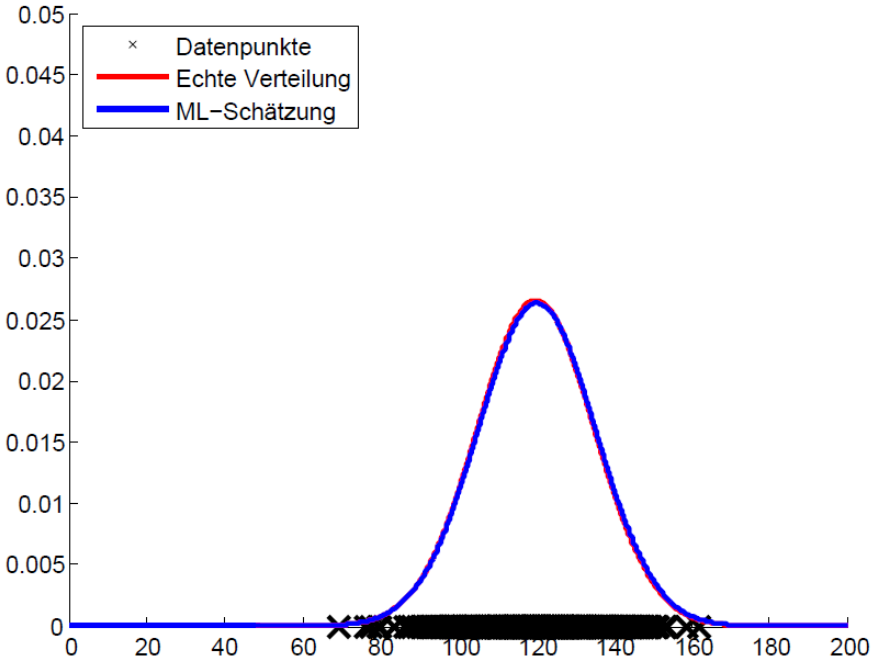
$$\mathcal{N}(\mu \mid \mu_0^*, (\beta^*\lambda)^{-1})\,\mathrm{Gam}(\lambda \mid a^*, b^*)$$

- Für grosse *n* nähert sich MAP Schätzung der ML Schätzung an

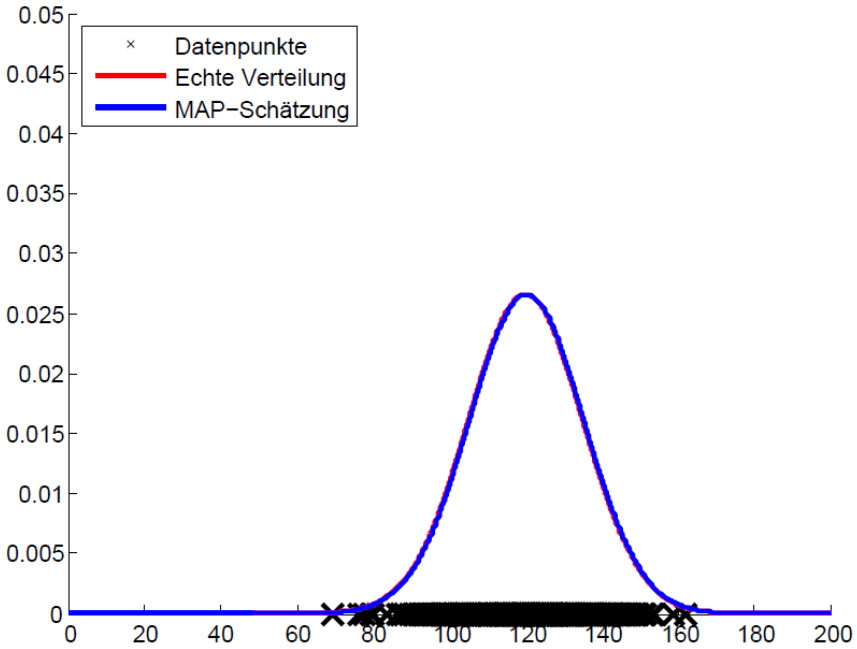
Normalverteilte Daten Schätzen: Beispiel MAP Parameter

- Simulation für *n=500:* Vergleich ML und MAP Lösung

ML



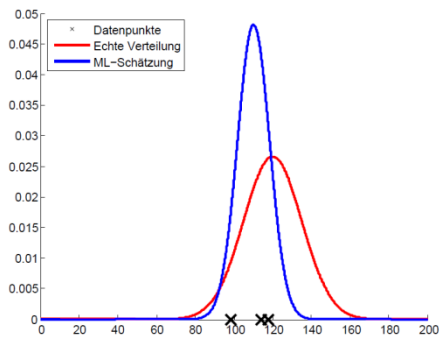
MAP

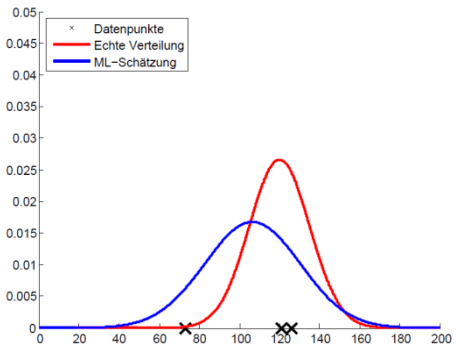


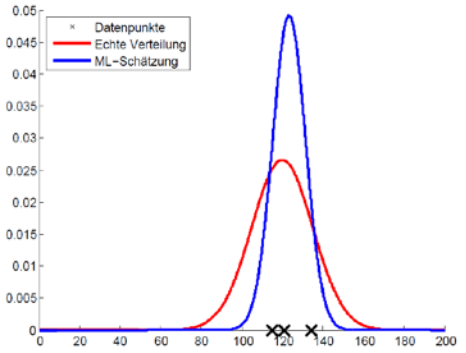
Normalverteilte Daten Schätzen: Beispiel MAP Parameter

- Mehrmalige Wiederholung der Simulation: *n=3* Punkte ziehen aus echter Verteilung, Vergleich ML/MAP Schätzung:

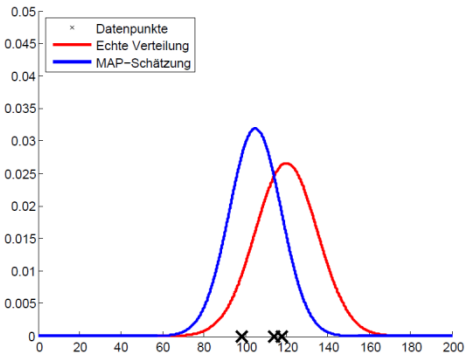
ML

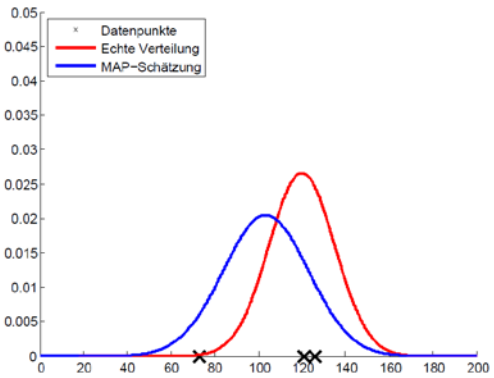


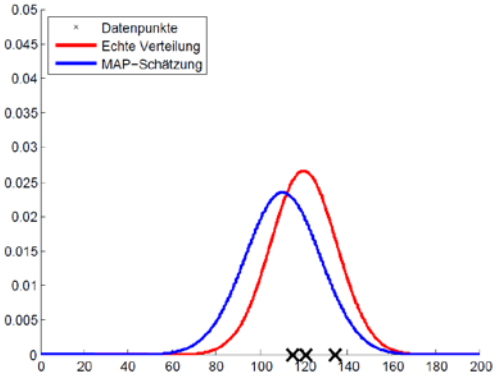




MAP







Beobachtungen ML vs. MAP Schätzung

- MAP Schätzungen Kompromiss zwischen Vorwissen und Evidenz der Daten
- MAP Schätzungen sind stabiler als ML Schätzungen: Schwankungen in den Daten beeinflussen Ergebnis weniger
- Je mehr Daten, desto kleiner die Varianz der Posterior-Verteilung: immer sicherer, was bestes Modell ist
- Für unendlich viele Daten ($n \to \infty$) konvergiert die MAP Lösung gegen die ML Lösung

Normalverteilung: Kumulative Verteilungsfunktion

- Gegeben Normalverteilung: was ist $p(\textit{beobachteter Wert} \geq x)$?
- Beispiel:
  - IQ einer zufällig gezogenen Person Zufallsvariable mit

$$X \sim \mathcal{N}(x \mid \mu, \sigma^2) \qquad \mu = 100, \quad \sigma = 15$$

  - Was ist $p(X \geq 120)$?
  - Normalisierung zur Standardnormalverteilung

$$X \sim \mathcal{N}(x \mid \mu, \sigma^2) \quad \Rightarrow \; Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(x \mid 0,1)$$

  - Wahrscheinlichkeit, IQ von 120 oder größer zu sehen?

$$p(X \geq 120) = P\left(\frac{X - 100}{15} \geq \frac{120 - 100}{15}\right) = p(Z \geq \frac{4}{3}) = 1 - p(Z \leq \frac{4}{3})$$

Kumulative Verteilungsfunktion

Normalverteilung: Kumulative Verteilungsfunktion

- Kumulative Verteilungsfunktion

$$\Phi(z) = p(Z \leq z)$$

$$= \int_{-\infty}^{z} \mathcal{N}(x \mid 0,1) dx$$

$$= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-x^2 / 2\right) dx$$

- Keine geschlossene Lösung, nachschlagen in Tabelle

Verteilungsfunktion der Normalverteilung

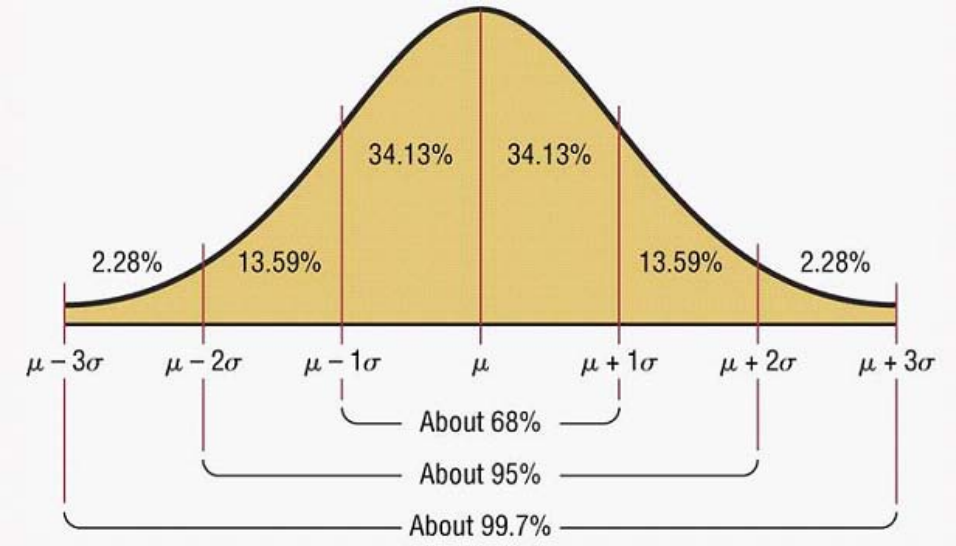
$$\Phi\left(\frac{4}{3}\right) \approx 0.9082$$

$$p(X \geq 120) \approx 0.0918$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

Normalverteilung: Kumulative Verteilungsfunktion

- Normalverteilung konzentriert die meiste Wahrscheinlichkeitsmasse „nahe“ dem Mittelwert
  - $p(\mu - \sigma \leq X \leq \mu - \sigma) \approx 0.68$
  - $p(\mu - 2\sigma \leq X \leq \mu - 2\sigma) \approx 0.95$
  - $p(\mu - 3\sigma \leq X \leq \mu - 3\sigma) \approx 0.997$



Multivariate Normalverteilung

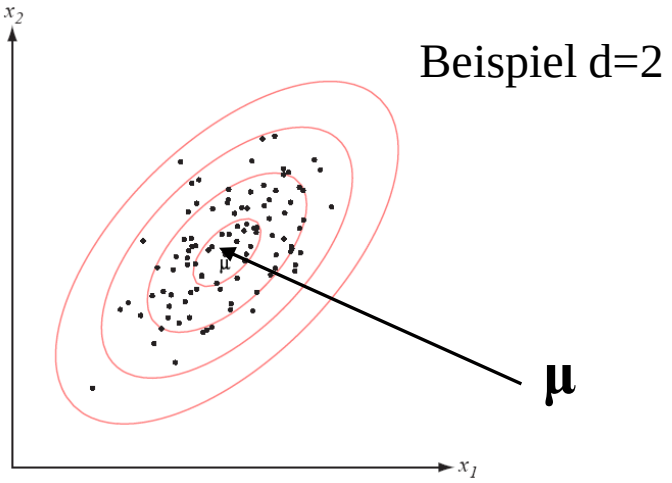
- Zufallsvariable **x** mit *d* Dimensionen.

$\mathbf{x} \in \mathbb{R}^d$ normalverteilt, wenn Verteilung beschrieben wird durch Dichte

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)$$

Determinante

- Mittelwertvektor $\boldsymbol{\mu} \in \mathbb{R}^d$
- Kovarianzmatrix Σ



- Koarianzmatrix entscheidet, wie Punkte streuen

Überblick

- Wahrscheinlichkeiten, Erwartungswerte, Varianz
- Grundkonzepte des Bayesschen Lernens
- (Bayessche) Parameterschätzung für Wahrscheinlichkeitsverteilungen
- Bayessche Lineare Regression, Naive Bayes