

Vorlesung 13a

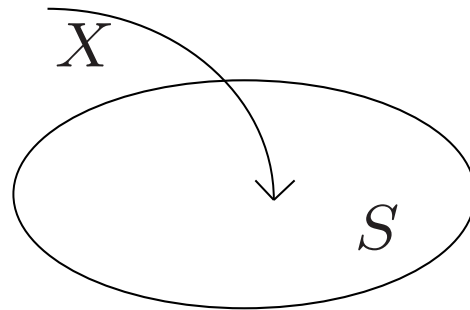
Maximum-Likelihood-Schätzung

1. Goldene Idee der Statistik:
Daten aufgefasst als
Realisierungen von Zufallsvariablen

Statistisches Modell:

eine Zufallsvariable X , bei deren Verteilung
ein Parameter ϑ frei bleibt:

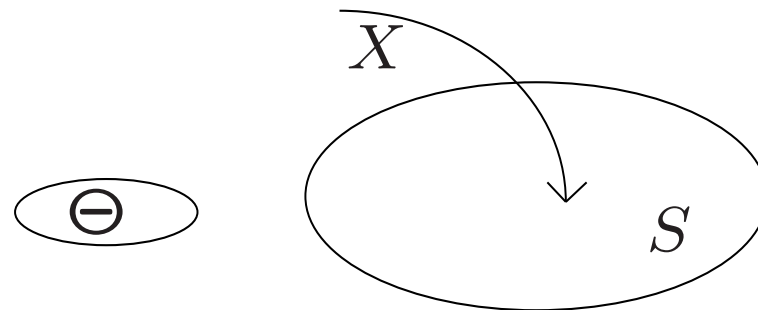
$$\mathbf{P}_{\vartheta}(X \in da) = \rho_{\vartheta}(da), \quad \vartheta \in \Theta$$



z.B. $\vartheta = (\mu, \sigma^2)$ bei der Normalverteilung,
oder $\vartheta = p$ beim Münzwurf.

Statistisches Modell:
eine Zufallsvariable X , bei deren Verteilung
ein Parameter ϑ frei bleibt:

$$\mathbf{P}_{\vartheta}(X \in da) = \rho_{\vartheta}(da), \quad \vartheta \in \Theta$$



Θ ... Menge der Parameter

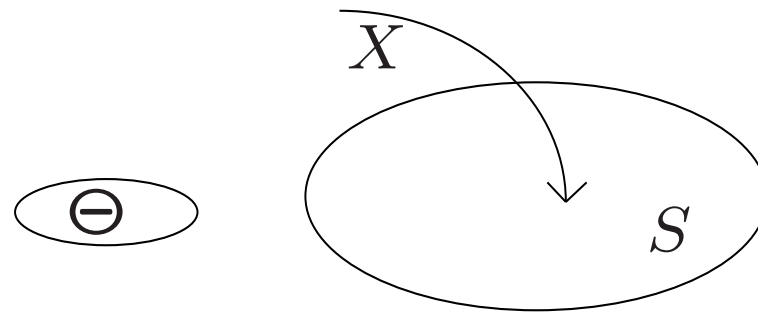
S ... Beobachtungsraum

Statistisches Modell:

eine Zufallsvariable X , bei deren Verteilung

ein Parameter ϑ frei bleibt:

$$\mathbf{P}_{\vartheta}(X \in da) = \rho_{\vartheta}(da), \quad \vartheta \in \Theta$$



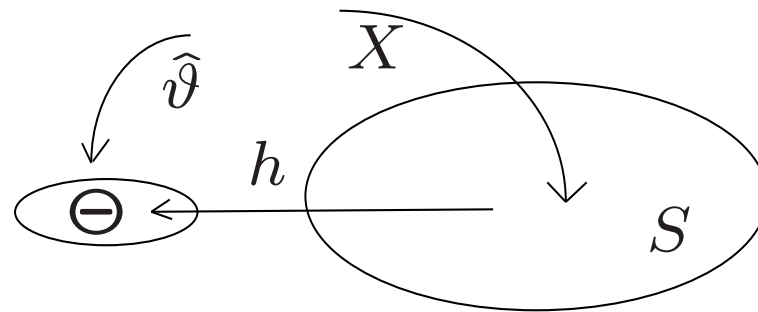
Der Parameter ϑ soll aus den Daten geschätzt werden.

Dazu verarbeitet man X zu einem *Schätzer* $\hat{\vartheta}$ für ϑ .

Statistisches Modell:

eine Zufallsvariable X , bei deren Verteilung
ein Parameter ϑ frei bleibt:

$$\mathbf{P}_{\vartheta}(X \in da) = \rho_{\vartheta}(da), \quad \vartheta \in \Theta$$



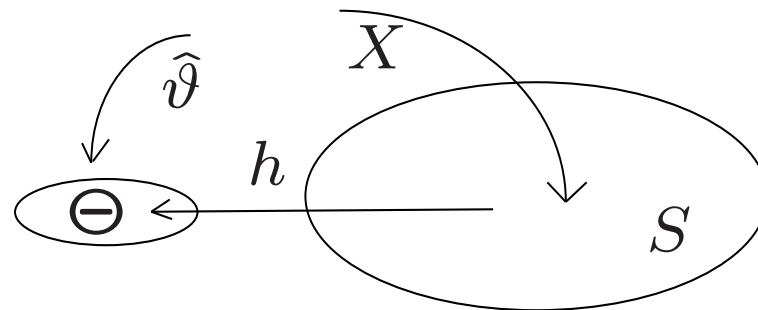
Der Parameter ϑ soll aus den Daten geschätzt werden.

Dazu verarbeitet man X zu einem *Schätzer* $\hat{\vartheta}$ für ϑ .

Statistisches Modell:

eine Zufallsvariable X , bei deren Verteilung
ein Parameter ϑ frei bleibt:

$$\mathbf{P}_{\vartheta}(X \in da) = \rho_{\vartheta}(da), \quad \vartheta \in \Theta$$

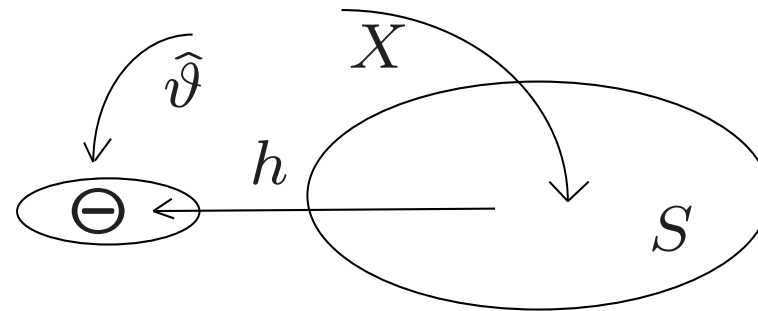


Θ ... Menge der Parameter

S ... Beobachtungsraum

$\hat{\vartheta} := h(X)$... Schätzer für den Parameter ϑ

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \Theta$$



Θ ... Menge der Parameter

S ... Beobachtungsraum

$\hat{\vartheta} := h(X)$... Schätzer für den Parameter ϑ

Naheliegende (“naive”) Schätzer:

für p aus einem n -fachen p -Münzwurf (X_1, \dots, X_n) :

$$\hat{p} = \frac{1}{n}(X_1 + \dots + X_n).$$

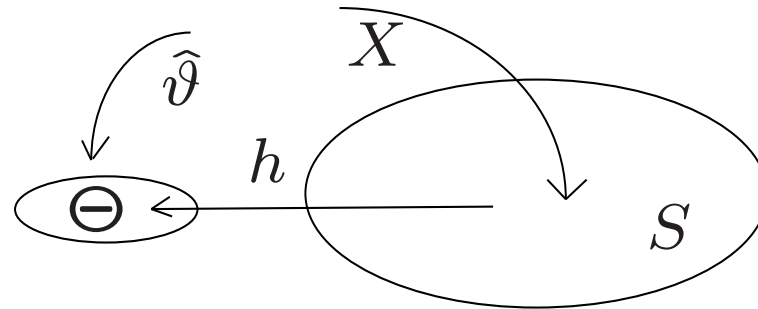
Und allgemeiner:

für μ aus unabhängigen reellwertigen
Zufallsvariablen (X_1, \dots, X_n) mit $\mathbf{E}[X_i] = \mu$:

$$\hat{\mu} = \frac{1}{n}(X_1 + \dots + X_n).$$

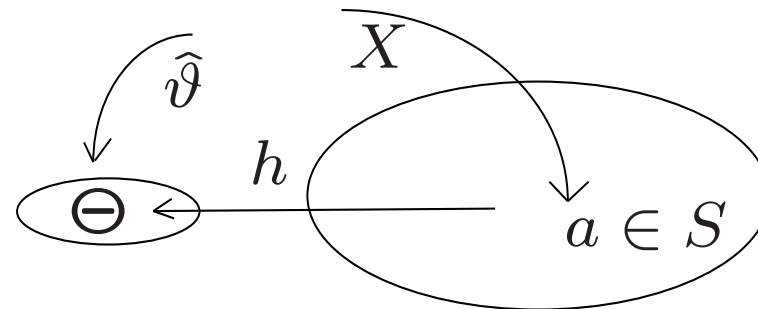
2. Das Maximum-Likelihood-Prinzip

$$\mathbf{P}_{\vartheta}(X \in da) = \rho_{\vartheta}(da), \quad \vartheta \in \Theta$$



Ein tragfähiges Prinzip zur Wahl der Abbildung h :
 Sei $h(a)$ dasjenige ϑ , für das die Wahrscheinlichkeit,
 den Ausgang a zu erhalten, maximal wird.

$$\mathbf{P}_{\vartheta}(X \in da) = \rho_{\vartheta}(da), \quad \vartheta \in \Theta$$



Ein tragfähiges Prinzip zur Wahl der Abbildung h :

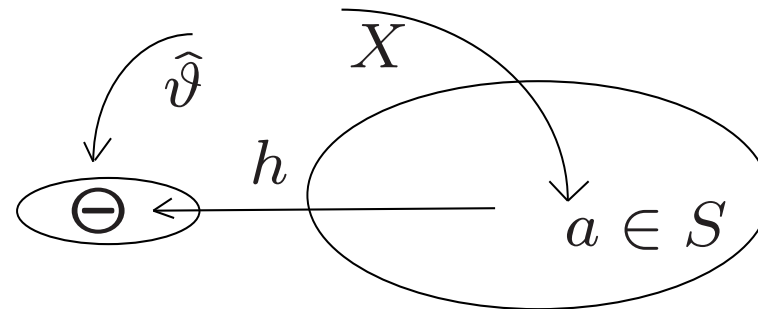
Für diskretes X : Wähle $h(a)$ so, dass

$$\mathbf{P}_{h(a)}(X = a) = \max_{\vartheta \in \Theta} \mathbf{P}_{\vartheta}(X = a).$$

Die Zufallsvariable $\hat{\vartheta} := h(X)$ nennt man dann

Maximum-Likelihood-Schätzer für ϑ auf der Basis von X :

$$\mathbf{P}_\vartheta(X \in da) = \rho_\vartheta(da), \quad \vartheta \in \Theta$$



Für eine Familie mit Dichten, $\rho_\vartheta(da) = f_\vartheta(a) da$, $\vartheta \in \Theta$,
geht man analog vor: Wähle $h(a)$ so, dass

$$f_{h(a)}(a) = \max_{\vartheta \in \Theta} f_\vartheta(a).$$

Die Zufallsvariable $\hat{\vartheta} := h(X)$ nennt man dann
Maximum-Likelihood-Schätzer für ϑ auf der Basis von X .

3. Beispiel: Münzwurf.

(X_1, \dots, X_n) sei n -facher p -Münzwurf mit unbekanntem p ,

$K_n := X_1 + \dots + X_n$ die Anzahl der Erfolge.

Beobachtet wird die Realisierung (a_1, \dots, a_n)

mit $k = a_1 + \dots + a_n$.

Behauptung: Unter allen p ist k/n derjenige Parameter,

für den $\mathbf{P}_p(X_1 = a_1, \dots, X_n = a_n) = p^k (1 - p)^{n-k}$

maximal ist.

Wir betrachten zuerst die Fälle $k = n$ und $k = 0$.

Hierfür stimmt die Behauptung wegen

$$\mathbf{P}_1(K_n = n) = \mathbf{P}_0(K_n = 0) = 1.$$

Behauptung: Unter allen p ist k/n derjenige Parameter, für den $\mathbf{P}_p(X_1 = a_1, \dots, X_n = a_n) = p^k(1 - p)^{n-k}$ maximal ist.

Denn: Der Logarithmus der rechten Seite ist

$$k \ln p + (n - k) \ln(1 - p).$$

Die Fälle $k \in \{0, 1\}$ hatten wir schon betrachtet.

Für $0 < k < n$ hat dieser Ausdruck sein Maximum in $p = k/n$, wie man durch Differenzieren feststellt.

Also ist

$$\frac{1}{n}(X_1 + \dots + X_n)$$

der Maximum-Likelihood-Schätzer für p .

Für $k = n$ (kein Misserfolg in n Versuchen) ergibt sich 1 als Maximum-Likelihood-Schätzung von p .

Das ist möglicherweise zu optimistisch.

Eine Alternative bietet der sogenannte *Bayes-Schätzer* (vgl Buch S. 127).

Hier denkt man an ein zweistufiges Experiment:

1. eine auf $[0, 1]$ uniform verteilte Zufallsvariable U
2. gegeben $\{U = u\}$ einen Münzwurf mit Erfolgswahrscheinlichkeit U .

$$\tilde{p} := \mathbf{E}[U|K] .$$

In der nächsten Vorlesung (13b) werden wir sehen:

Z_1, Z_2, \dots sei ein Münzwurf mit uniform auf $[0, 1]$ verteiltem zufälligem Erfolgsparameter U ,

K_n sei die Anzahl der Erfolge in den ersten n Versuchen.

Dann ist

$$\mathbf{E}[U \mid K_n = k] = \frac{k + 1}{n + 2}.$$

Man nennt dies auch den

Bayes-Schätzer für die Erfolgswahrscheinlichkeit

(bei a priori uniform verteilter Erfolgswahrscheinlichkeit).

4. Beispiel:

Unabhängige, identisch normalverteilte
Zufallsvariable.

X_1, \dots, X_n seien unabhängig und $N(\mu, \sigma^2)$ -verteilt, $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$.

Behauptung:

Der ML-Schätzer für $\vartheta = (\mu, \sigma)$ ist dann $(\hat{\mu}, \hat{\sigma})$

mit

$$\hat{\mu} := \frac{1}{n}(X_1 + \dots + X_n),$$
$$\hat{\sigma}^2 := \frac{1}{n}((X_1 - \hat{\mu})^2 + \dots + (X_n - \hat{\mu})^2) .$$

Denn: Die gemeinsame Dichtefunktion ist

$$\varphi_{\mu, \sigma^2}(a_1) \cdots \varphi_{\mu, \sigma^2}(a_n)$$

mit $\varphi_{\mu, \sigma^2}(x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)}$, $x \in \mathbb{R}$.

Damit ist die Aufgabe: Finde für gegebenes (a_1, \dots, a_n)
das maximierende (μ, σ^2) .

Im ersten Schritt betrachten wir für festes σ^2 die Abbildung

$$\begin{aligned} \mu &\mapsto L(\mu, \sigma) := \ln \varphi_{\mu, \sigma^2}(a_1) + \dots + \ln \varphi_{\mu, \sigma^2}(a_n) \\ &= \text{const} - n \ln \sigma - \frac{1}{2\sigma^2} \left((a_1 - \mu)^2 + \dots + (a_n - \mu)^2 \right). \end{aligned}$$

Diese wird maximiert bei $m = \frac{1}{n}(a_1 + \dots + a_n)$ (warum?)

Im zweiten Schritt

differenzieren wir die Abbildung $\sigma \mapsto L(m, \sigma)$:

$$\frac{d}{d\sigma} L(m, \sigma) = -\frac{n}{\sigma} - \frac{1}{\sigma^3} \left((a_1 - m)^2 + \dots + (a_n - m)^2 \right).$$

Damit bekommen wir für das Quadrat der Maximalstelle

$$\hat{\sigma}^2 := \frac{1}{n} \left((a_1 - m)^2 + \dots + (a_n - m)^2 \right). \quad \square$$

Fazit: Für unabhängige, $N(\mu, \sigma^2)$ -verteilte X_1, \dots, X_n
ist der ML-Schätzer für σ^2

die uns aus Übungsaufgabe 29 bekannte Zufallsvariable

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - M)^2.$$

Dort haben wir gesehen:

$$\mathbf{E}_{(\mu, \sigma^2)} [\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2.$$

$$\mathbf{E}_{(\mu, \sigma^2)} [\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2.$$

Der Schätzer $\hat{\sigma}^2$ ist also nicht “erwartungstreu”,
wohl aber seine Modifikation

$$S^2 := \frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} ((X_1 - M)^2 + \dots + (X_n - \hat{M})^2).$$

Und es gilt sogar (siehe Buch S. 138) der Satz von Gosset-Fisher:

M und S^2 sind unabhängig, und $\frac{S^2}{\sigma^2}$ ist so verteilt wie die Summe aus $n - 1$ Quadraten von unabhängigen standard-normalverteilten Z_i .

5. Beispiel: Einfache lineare Regression.

x_1, \dots, x_n seien feste reelle Zahlen,

$$Y_i = \beta_0 + \beta_1 x_i + \sigma Z_i, \quad i = 1, \dots, n,$$

mit Z_1, \dots, Z_n unabhängig, $N(0, 1)$ -verteilt.

Die Dichtefunktion von $Y = (Y_1, \dots, Y_n)$ hat

am Ausgang $a = (a_1, \dots, a_n)$ den Wert

$$(*) \quad \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{|a - \mu|^2}{2\sigma^2}\right),$$

mit $\mu_i := \beta_0 + \beta_1 x_i$, $\mu := (\mu_1, \dots, \mu_n)$.

Wieder maximieren wir zuerst bei festgehaltenem σ .
Damit ergeben sich die uns aus Vorlesung 8a, Abschnitt 8,
wohlbekannten Koeffizienten der Regressionsgeraden:

$$b_1 = \frac{\sum_{i=1}^n (a_i - \bar{a})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad b_0 = \bar{a} - b_1 \bar{x} .$$

Betrachten wir jetzt (*) als Funktion von σ ,

mit $\hat{\mu}_i := b_0 + b_1 x_i$ statt μ_i ,

so finden wir durch Logarithmieren und Differenzieren
dessen Maximalstelle bei

$$\frac{1}{n} \sum_{i=1}^n (a_i - \hat{\mu}_i)^2 .$$

Der ML-Schätzer für $(\beta_0, \beta_1, \sigma^2)$ ergibt sich durch Einsetzen von Y anstelle von a :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

Ein erwartungstreuer Schätzer mit schönen Verteilungseigenschaften (siehe Buch S. 138) ist

$$s^2 := \frac{n}{n-2} \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

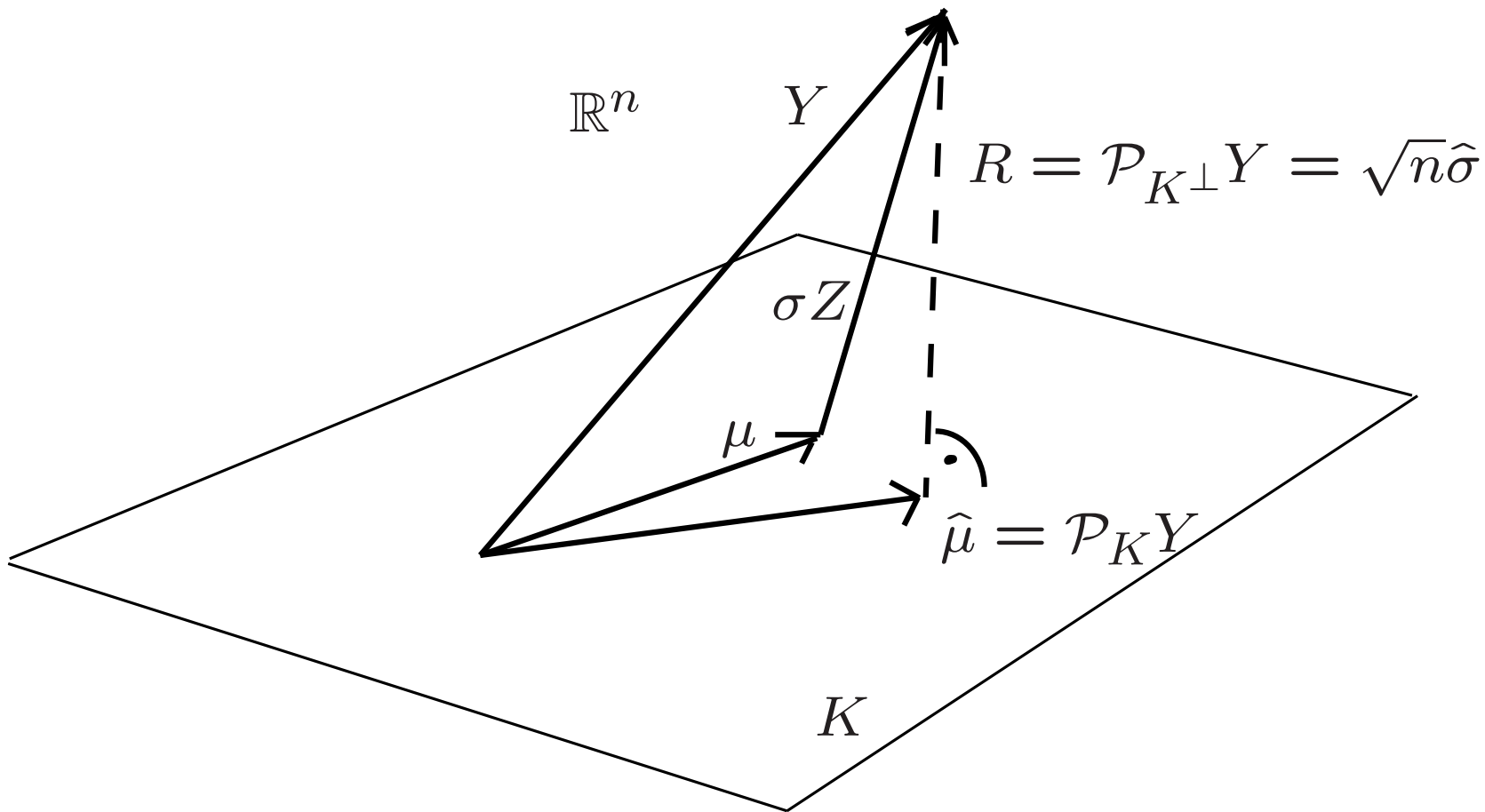
Für alle, die Geometrie mögen:

Der n -dimensionale Datenvektor a soll dargestellt werden durch einen 2-dimensionalen “systematischen Beitrag” plus ein (möglichst kleines) Residuum (“Rest”):

$$a = b_0 \mathbf{1} + b_1 x + r, \quad \text{mit } |r|^2 = \min!.$$

Dabei ist $\mathbf{1} = (1, \dots, 1)$ und $x := (x_1, \dots, x_n)$.

Im folgenden Bild (vgl Buch S. 136) ist K der von $\mathbf{1}$ und x aufgespannte 2-dimensionale Teilraum des Datenraumes \mathbb{R}^n .



6. Beispiel:
Zweiseitige Exponentialverteilung.

Für $\vartheta \in \mathbb{R}$ seien X_1, \dots, X_n
unabhängig und identisch verteilt mit Dichte

$$g_{\vartheta}(x) := \frac{1}{2}e^{-|x-\vartheta|}, x \in \mathbb{R}.$$

Was ist der ML-Schätzer für ϑ ?

$$f_{\vartheta}(a_1, \dots, a_n) := \left(\frac{1}{2}\right)^n \prod_{i=1}^n e^{-|a_i - \vartheta|}, \quad a_i \in \mathbb{R}$$

Betrachten wir erst einmal den Fall $n = 2$,

und für festes $a_1, a_2 \in \mathbb{R}$ die Funktion

$$\ell : \vartheta \mapsto |a_1 - \vartheta| + |a_2 - \vartheta|.$$

Angenommen, $a_1 < a_2$.

Für $\vartheta < a_1$ hat ℓ die Steigung -2 ,

für $\vartheta > a_2$ hat ℓ die Steigung $+2$,

für $\vartheta \in (a_1, a_2)$ hat ℓ die Steigung 0 .

Also ist jedes $\vartheta \in [a_1, a_2]$ Minimalstelle von ℓ .

$$f_{\vartheta}(a_1, \dots, a_n) := \left(\frac{1}{2}\right)^n \prod_{i=1}^n e^{-|a_i - \vartheta|}, \quad a_i \in \mathbb{R}$$

Betrachten wir jetzt den Fall $n = 3$,
und für festes $a_1, a_2, a_3 \in \mathbb{R}$ die Funktion
 $\ell : \vartheta \mapsto |a_1 - \vartheta| + |a_2 - \vartheta| + |a_3 - \vartheta|$.

Angenommen, $a_1 < a_2 < a_3$.

Für $\vartheta < a_2$ hat ℓ negative Steigung,

für $\vartheta > a_2$ hat ℓ positive Steigung.

Also ist a_2 die einzige Minimalstelle von ℓ .

Was 2 und 3 recht ist, soll n billig sein.

Definition. Seien $a_1, \dots, a_n \in \mathbb{R}$.

Eine Zahl heißt **Median** von a_1, \dots, a_n ,

wenn

ebenso viele der a_i links wie rechts von ihr liegen.

Für ungerades n führt die Definition auf einen einzigen Wert,
für gerades n führt sie auf ein Intervall.

Fazit:

Im Beispiel der zweiseitigen Exponentialverteilung ist der ML-Schätzer für den Lageparameter (das “Zentrum”) ϑ von der Form

$$h(X_1, \dots, X_n) := \text{Median von } X_1, \dots, X_n$$

(und nicht, wie man auf die Schnelle vielleicht vermuten würde, der arithmetische Mittelwert der X_i).

7. Beispiel: Uniforme Verteilung

X_1, \dots, X_n seien unabhängig und uniform verteilt auf $[0, \vartheta]$.

Was ist der ML-Schätzer für ϑ ?

Die Dichtefunktion von (X_1, \dots, X_n) ist

$$f_{\vartheta}(a_1, \dots, a_n) = \frac{1}{\vartheta^n}, \quad 0 \leq \min(a_i) \leq \max(a_i) \leq \vartheta.$$

Für festes (a_1, \dots, a_n) wird sie maximiert bei $\vartheta = \max(a_i)$.

Also ist

$$\hat{\vartheta} = \max\{X_i : i = 1, \dots, n\}.$$

Unter \mathbf{P}_ϑ ist $\max\{X_i : i = 1, \dots, n\}$ so verteilt wie

$$\vartheta \max\{U_i : i = 1, \dots, n\},$$

mit U_1, \dots, U_n unabhängig und $\text{Unif}([0, 1])$ -verteilt.

Also ist (vgl. die kommende Vorlesung 13b)

$$\mathbf{E}_\vartheta[\hat{\vartheta}] = \mathbf{E}_\vartheta[\max\{X_i : i = 1, \dots, n\}] = \frac{n}{n+1}\vartheta.$$