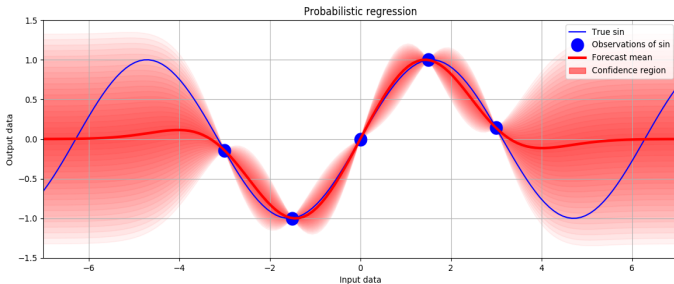


# Gauß-Prozess-Regression

## Bayessche Regression und Gaußprozesse

Dr. rer. nat. Johannes Riesterer



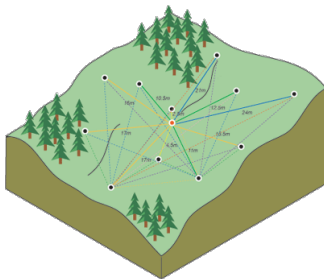


Figure: Quelle: esri.com

## Kriging

Der südafrikanische Bergbauingenieur Danie Krige versuchte 1951, eine optimale Interpolationsmethode für den Bergbau zu entwickeln, basierend auf der räumlichen Abhängigkeit von Messpunkten.

## Daten

Gegeben Daten  $S := \{(x_i, y_i)\}_i$  mit Features  $x_i \in \mathbb{R}^n$  und Werten  $y_i \in \mathbb{R}$ .

## Modell

$y_i = x_i^t \cdot \omega + \epsilon$  mit Gewichten  $\omega \in \mathbb{R}^n$  und Offset  $\epsilon \in \mathbb{R}$ .

## Training

$(\omega, \epsilon) = \min L_S(\omega, \epsilon)$  mit Verlustfunktion  
 $L_S(\omega, \epsilon) := \sum_i (y_i - (x_i^t \cdot \omega + \epsilon))^2$ .

## Bemerkung

Optimierungsproblem. Für  $n = 1$  Gauss'sche Methode der kleinsten Fehlerquadrate.

## Vorhersage

Für Feature  $\tilde{x} \in \mathbb{R}^n$  definiere Vorhersage durch

$$\tilde{y} = \tilde{x}^t \cdot \omega + \epsilon$$

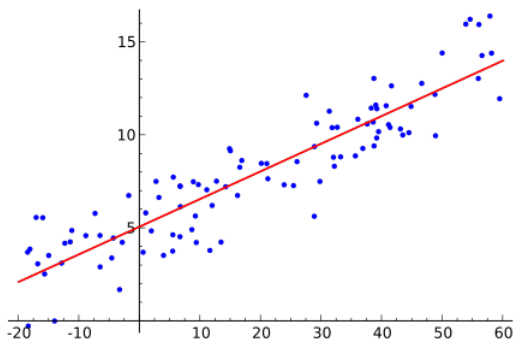


Figure: Quelle: Wikipedia

## Stichproben

Gegeben Stichproben  $S := \{(x_i, y_i)\}_i$  der Zufallsvariablen  $(x^{(i)}, y^{(i)})$  bzw.  $X = (x^{(i)})_i$  und  $Y = (y^{(i)})_i$ .

## Modell

$y^{(i)} = (x^{(i)})^t \cdot \omega + \epsilon$  mit Gewichten  $\omega \sim \mathcal{N}(0, \Sigma)$ ,  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ .  
Hierbei wird  $x^{(i)} \equiv x_i$  als konstant angenommen und nicht als Zufallsvariable modelliert.

$$(y^i | x^i, \omega) \sim \mathcal{N}(\omega^t \cdot x^i, \sigma^2) \quad (1)$$

$$\Leftrightarrow p(y_i | x_i, \omega) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - \omega^t x_i)^2}{2\sigma}\right) \text{(Dichte)} \quad (2)$$

$$y_i \text{ u.i.v.} \Rightarrow p(Y|X, \omega) = \prod_i \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - \omega^t x_i)^2}{2\sigma}\right) \quad (3)$$

## Posterior distribution

Satz von Bayes

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{marginal likelihood}}$$
$$p(\omega|S) = \frac{p(S|\omega) \cdot p(\omega)}{p(S)}$$

mit der Marginalisierung

$$p(S) = \int_{\omega'} p(S, \omega') d\omega' = \int_{\omega'} p(S|\omega') \cdot p(\omega') d\omega'.$$

## Posterior distribution

Einsetzen mit  $p(S|\omega) = p(Y|X, \omega)$  da  $x^{(i)} \equiv x_i$ :

$$p(\omega|S) = \frac{(\prod_i p(y^i|x_i, \omega)) \cdot p(\omega)}{\int_{\omega'} (\prod_i p(y^i|x_i, \omega')) \cdot p(\omega') d\omega'} \quad (4)$$

## Posterior distribution

Mit (3),  $\omega \sim \mathcal{N}(0, \tau^2 \cdot I)$  und Rechenregeln für Normalverteilungen (längere Übungsaufgabe):

$$\omega|S \sim \mathcal{N}\left(\frac{1}{\sigma^2}A^{-1}X^tY, A^{-1}\right)$$

mit  $A = \frac{1}{\sigma^2}X^tX + \Sigma^{-1}$ .

## Posterior predictive distribution

Für Feature  $\tilde{x}$  erhalten wir durch Marginalisierung:

$$p(\tilde{y}|\tilde{x}, S) = \int_{\omega} p(\tilde{y}, \omega|\tilde{x}, S) d\omega = \int_{\omega} \underbrace{p(\tilde{y}|\omega, \tilde{x}, S)}_{=p(\tilde{y}|\omega, \tilde{x})} \cdot p(\omega|S) d\omega$$

( $\tilde{y}$  unabh. von  $S$ ).

## Posterior predictive distribution

Mit (2) und (4) und Rechenregeln für Normalverteilungen (wieder längere Übungsaufgabe):

$$\tilde{y}|\tilde{x}, S \sim \mathcal{N}\left(\frac{1}{\sigma^2}\tilde{x}^t A^{-1} X^t Y, \tilde{x}^t A^{-1} \tilde{x}\right)$$

mit  $A = \frac{1}{\sigma^2} X^t X + \Sigma^{-1}$



## Vorhersage

Für Feature  $\tilde{x}$  wird die Vorhersage durch

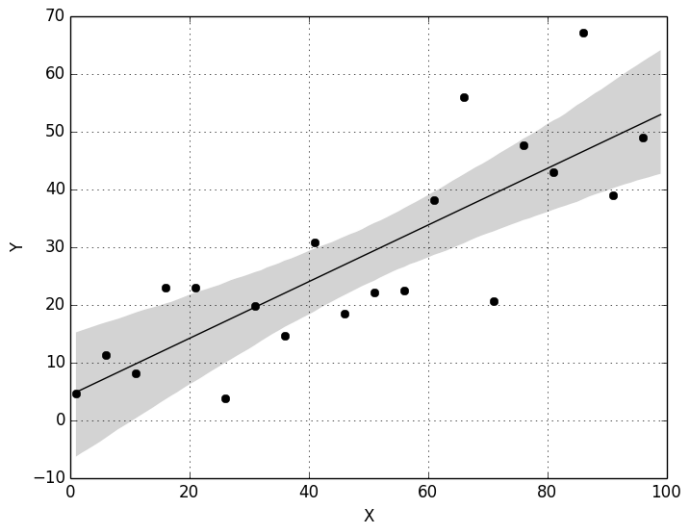
$$\tilde{y} := \mathbb{E}((\tilde{y}|\tilde{x}, S)) = \frac{1}{\sigma^2} \tilde{x}^t A^{-1} X^t Y$$

definiert. Die Varianz

$$\mathbb{V}((\tilde{y}|\tilde{x}, S)) = \tilde{x}^t A^{-1} \tilde{x}$$

dient als Mass der Güte der Vorhersage.

# Lineare Bayessche Regression



## Stichproben

Stichproben  $S := \{(x_i, y_i)\}_i$  der Zufallsvariablen  $(x^{(i)}, y^{(i)})$ .

## Modell

$$y^{(i)} = f(x^{(i)}) + \epsilon.$$

## Vorhersage

Im Allgemeinen sind die Integrale, welche in der posterior und posterior predictive distribution vorkommen, nicht geschlossen lösbar. In diesem Fall wird häufig Markov-Chain-Monte-Carlo Integration (MCMC) verwendet.

## Kernel Trick

Der Spezialfall  $f(x^{(i)}) = \phi(x^{(i)})^t \cdot \omega$  mit  $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^N$  (Bsp:  $\phi(x_1, x_2) = (x_1, x_1^2, x_2, x_2^2)$ ) ist geschlossen lösbar, da linear. Mit nahezu analoger Rechnung erhält man

$$\tilde{f}|\tilde{x}, S \sim \mathcal{N} \left( \frac{1}{\sigma^2} \phi(\tilde{x})^t A^{-1} X^t Y, \phi(\tilde{x})^t A^{-1} \phi(\tilde{x}) \right)$$

mit  $A = \frac{1}{\sigma^2} \phi(X)^t \phi(X) + \Sigma^{-1}$

## Stochastischer Prozess

Ein stochastischer Prozess ist eine indizierte Menge von Zufallsvariablen  $\{f_x \mid x \in \mathcal{X}\}$ .

## Gauß-Prozess

Wir bezeichnen einen stochastischen Prozess als Gauß-Prozess  $f_x \sim \mathcal{GP}(m(x), k(x, x'))$ , falls

$$\begin{pmatrix} f_{x_1} \\ \vdots \\ f_{x_n} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{pmatrix} \right)$$

für jede endliche Teilmenge  $X = (x_1, \dots, x_n) \in \mathcal{X}$ . Man nennt  $k(x, x')$  Kovarianz-Funktion oder auch Kernel. Zulässig sind nur Funktionen, bei denen die Matrix positiv definit und symmetrisch ist für jede endliche Teilmenge  $X$ .

## Beispiel

$f_x = \phi(x)^t \cdot \omega$  mit  $\omega \sim \mathcal{N}(0, \Sigma)$  ist ein  $\mathcal{GP}(m(x), k(x, x'))$  mit

$$\mathbb{E}(f_x) = \phi(x)^t \mathbb{E}(\omega) = \underbrace{0}_{:=m(x)}$$

$$\mathbb{E}(f_x f_{x'}) = \phi(x)^t \mathbb{E}(\omega \omega^t) \phi(x') = \underbrace{\phi(x)^t \Sigma \phi(x')}_{:=k(x, x')}$$

## Prior distribution

Sei  $f \sim \mathcal{GP}(0, k(x, x'))$  ein Gauß-Prozess. Angenommen man kennt  $f = (f_{x_1} \dots f_{x_n})$  an den Punkten  $X = (x_1, \dots, x_n)$  und möchte  $\tilde{f} = (f_{\tilde{x}_1} \dots f_{\tilde{x}_n})$  an den Punkten  $\tilde{X} = (\tilde{x}_1, \dots, \tilde{x}_n)$  vorhersagen. Aus der GP-Eigenschaft folgt

$$\begin{pmatrix} f_x \\ \tilde{f}_{\tilde{x}} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(X, X) & K(X, \tilde{X}) \\ K(\tilde{X}, X) & K(\tilde{X}, \tilde{X}) \end{pmatrix} \right)$$

## Posterior predictive distribution

Mit den Rechenregeln für multivariate Normalverteilungen folgt:

$$\tilde{f} | \tilde{X}, X, f \sim \mathcal{N}(\mu_{\tilde{f}}, \Sigma_{\tilde{f}})$$

$$\mu_{\tilde{f}} := K(\tilde{X}, X)K(X, X)^{-1} \cdot f$$

$$\Sigma_{\tilde{f}} := K(\tilde{X}, \tilde{X}) - K(\tilde{X}, X)K(X, X)^{-1}K(X, \tilde{X})$$



# Gauß-Prozess-Regression

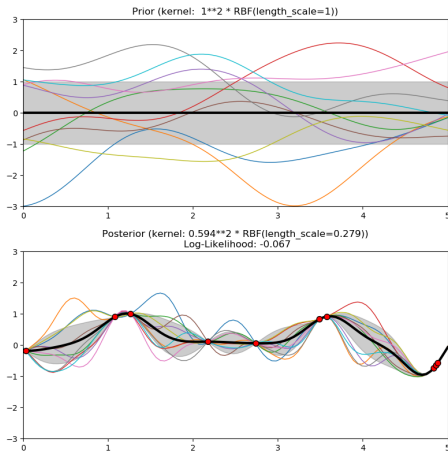


Figure: Quelle: Wikipedia

## Vorhersage

Für Prior  $\tilde{f}$  wird die Vorhersage durch

$$R(\tilde{f}) := \mathbb{E}((\tilde{f}|\tilde{X}, X, f)) = K(\tilde{X}, X)K(X, X)^{-1} \cdot f$$

definiert. Die Varianz

$$\mathbb{V}((\tilde{f}|\tilde{X}, X, f)) = K(\tilde{X}, \tilde{X}) - K(\tilde{X}, X)K(X, X)^{-1}K(X, \tilde{X})$$

dient als Mass der Güte der Vorhersage.

## Kernel

Kernel	Funktion
konstant	$\sigma_0^2$
linear	$\sum_{d=1}^D \sigma_d^2 x_d x'_d$
polynomial	$(x \cdot x' + \sigma_0^2)^p$
squared exponential	$\exp(-\frac{r^2}{2l^2})$
Matérn	$\frac{1}{2^{\nu-1} \Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{l} r\right) K_\nu \left(\frac{\sqrt{2\nu}}{l} r\right)$
exponentziell	$\exp\left(-\frac{r}{l}\right)$
$\gamma$ -exponentziell	$\exp\left(-\left(\frac{r}{l}\right)^\gamma\right)$
rational quadratisch	$\left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha}$

Figure: Quelle: Wikipedia