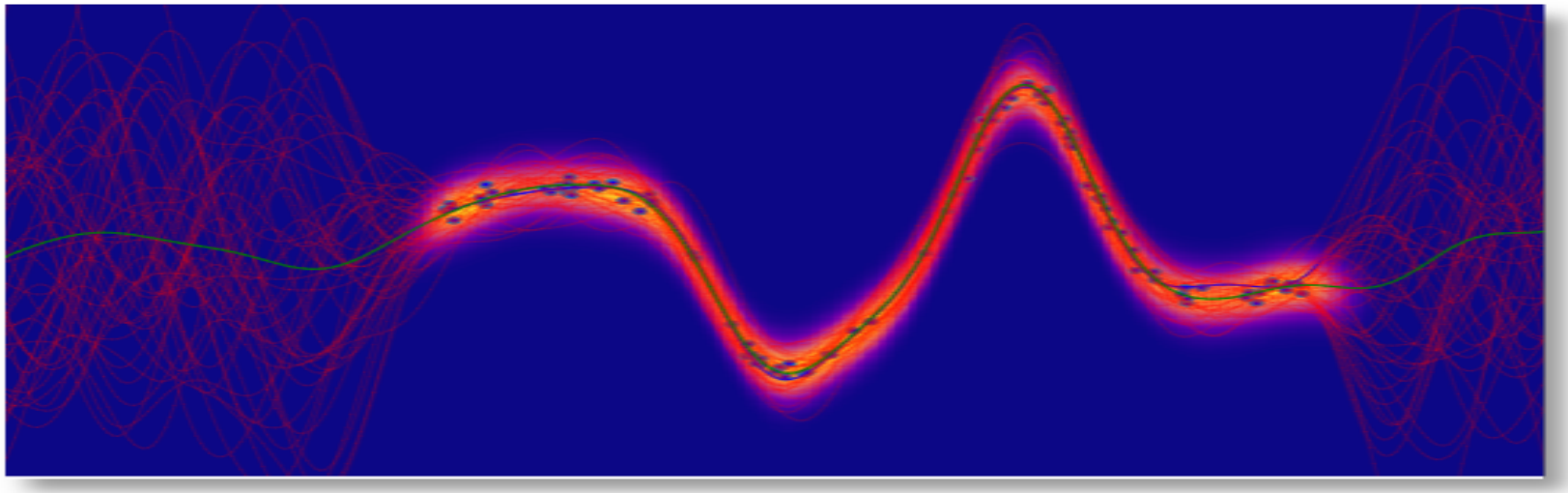
Stochastische Regressionsmodelle für heterogene Messnetzwerke

Gaußprozesse, Kernel, Hyperparameteroptimierung.

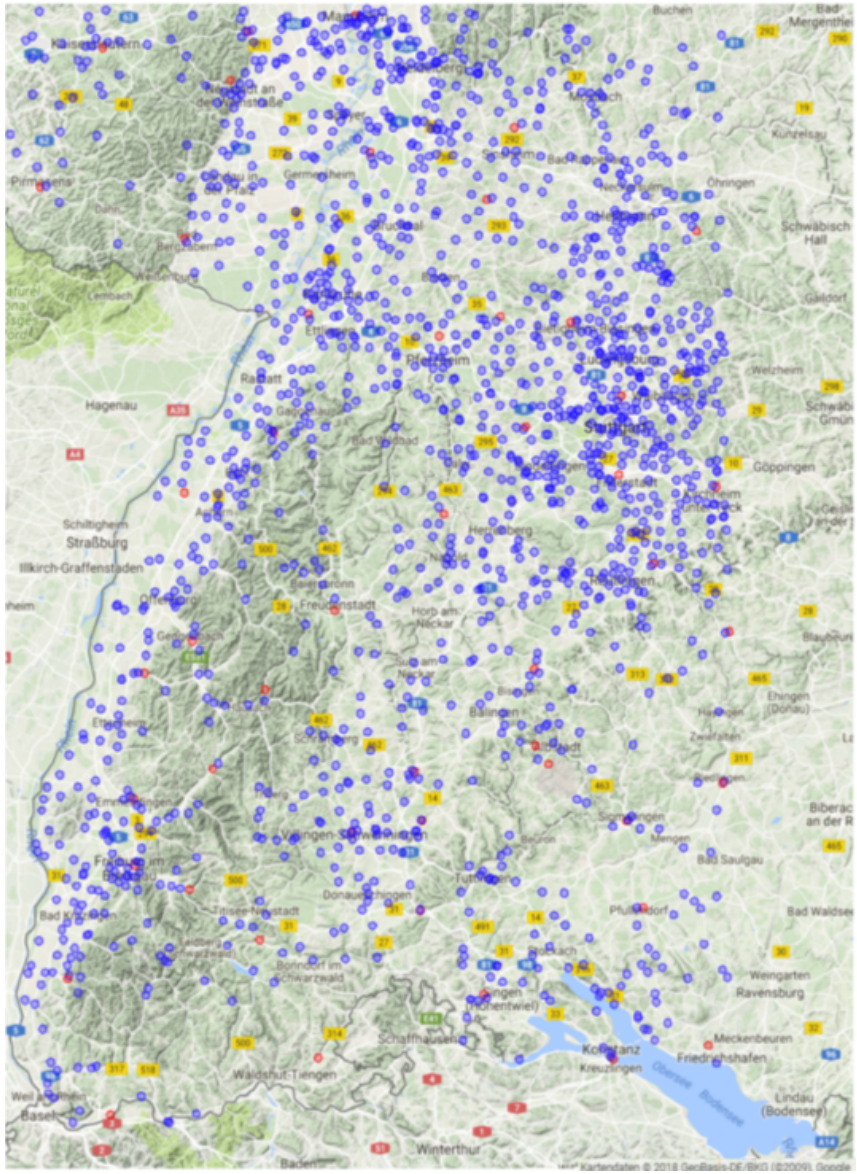
Dr. rer. nat. Johannes Riesterer - KIT/TECO



Motivation

Automated Quality Assessment of (Citizen) Weather Stations; Julian Bruns, Johannes Riesterer, Bowen Wang, Till Riedel, Michael Beigl ; GI-Forum (2018).

Motivation


Straßburg
Hagenau
Rastatt
Freudenstadt
Horb am Neckar
Balingen
Göppingen
Schwäbisch Hall
Ravensburg
Weingarten
Friedrichshafen
Konstanz
Meckenbeuren
Lindau (Bodensee)
Waldshut-Tiengen
Schaffhausen
Winterthur
Basel
Ehingen (Donau)
Bad Saulgau

Statistisches Lernen

Gegeben

Gegeben sehr grosse Stichproben (Big Data) $D = \{(y_i, x_i)_i\}$ der Zufallsvariablen $Y : \mathbb{R}^n \to \mathbb{R}$, $X : \mathbb{R}^n \to \mathbb{R}^m$ mit unbekannten Verteilungen.

Gesucht

Funktion $f : \mathbb{R}^m \to \mathbb{R}$ mit $f = argmin_h L(Y, h(X))$, wobei L eine sogenannte Fehler-Funktion ist.

Bedinger Erwartungswert

$L(Y, h(X)) := \int (Y - h(X))^2 d\mu \Rightarrow f = \mathbb{E}(Y|X)$

Conclusio

Bestimme "posterior predictive distribution" $p(y|x, D)$

Algorithmisches Problem

$$Y = f(X) =: f_X$$

$$f = argmin_{\{ f_\theta | \ \theta \in \Theta \}} L(Y, f_\theta(X))$$

$$L(Y, h(X)) \doteq \sum_D (y - h(x))^2$$

$$D = D_{\text{Train}} \cup D_{\text{Test}} \cup D_{\text{Validation}}$$

Bayessche Interferenz

Posterior predictive distribution

$$p(y^*|x^*, D) = p(f_{x^*}|x^*, D) = \int_\Omega p(f_{y^*}, \omega|x^*, D)d\omega =$$
$$\int_\Omega \underbrace{p(f_{x^*}|\omega, x^*, D)}_{=p(f_{x^*}|\omega,x^*)} \cdot p(\omega|D)d\omega$$

Satz von Bayes

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

Bayessche Interferenz

Conclusio

Wähle problemspezifische likelihood und prior Verteilungen $p(D|\omega)$ und $p(\omega)$.

Konjugierte Paare

Selten analytisch lösbar - wichtige Ausnahme konjugierte Paare.

Rejection sampling/Markov-Chain-Monte-Carlo

Integrale lösen durch sampling:

$$\sum_{\theta_i} \frac{g(\theta_i)}{p(\theta_i)} \rightarrow \int_{\Theta} g d\theta$$

Normalisierung

Mit der Marginalisierung ist
$p(D) = \int_{\omega'} p(D, \omega') d\omega' = \int_{\omega'} p(D|\omega') \cdot p(\omega') d\omega'$ nun berechenbar.

Lineare Bayessche Regression

Modell

$Y = f(X) = X^t \cdot \omega + \epsilon$ mit Gewichten $\omega \sim \mathcal{N}(0, \Sigma)$,
$\epsilon \sim \mathcal{N}(0, \sigma^2)$ und X konstant.

Likelihood

$$
\begin{aligned}
 & (y_i|x_i, \omega) && \sim \mathcal{N}(\omega^t \cdot x_i, \sigma^2) \\
\Leftrightarrow \quad & p(y_i|x_i, \omega) && = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \omega^t x_i)^2}{2\sigma}\right) \\
y_i \underset{\Rightarrow}{\text{ u.i.v.}} \quad & p(Y|X, \omega) && = \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \omega^t x_i)^2}{2\sigma}\right)
\end{aligned}
$$

Lineare Bayessche Regression

Posterior distribution

$$p(\omega|S) = \frac{(\prod_i p(y_i|x_i,\omega)) \cdot p(\omega)}{\int_{\omega'}(\prod_i p(y_i|x_i,\omega')) \cdot p(\omega')d\omega'}$$

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

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

Posterior predictive distribution

$$y^*|x^*, D \sim \mathcal{N}\left(\frac{1}{\sigma^2}(x^*)^tA^{-1}X^tY,\ (x^*)^tA^{-1}x^*\right)$$

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

Vorhersage

Für Feature x^* wird die Vorhersage durch

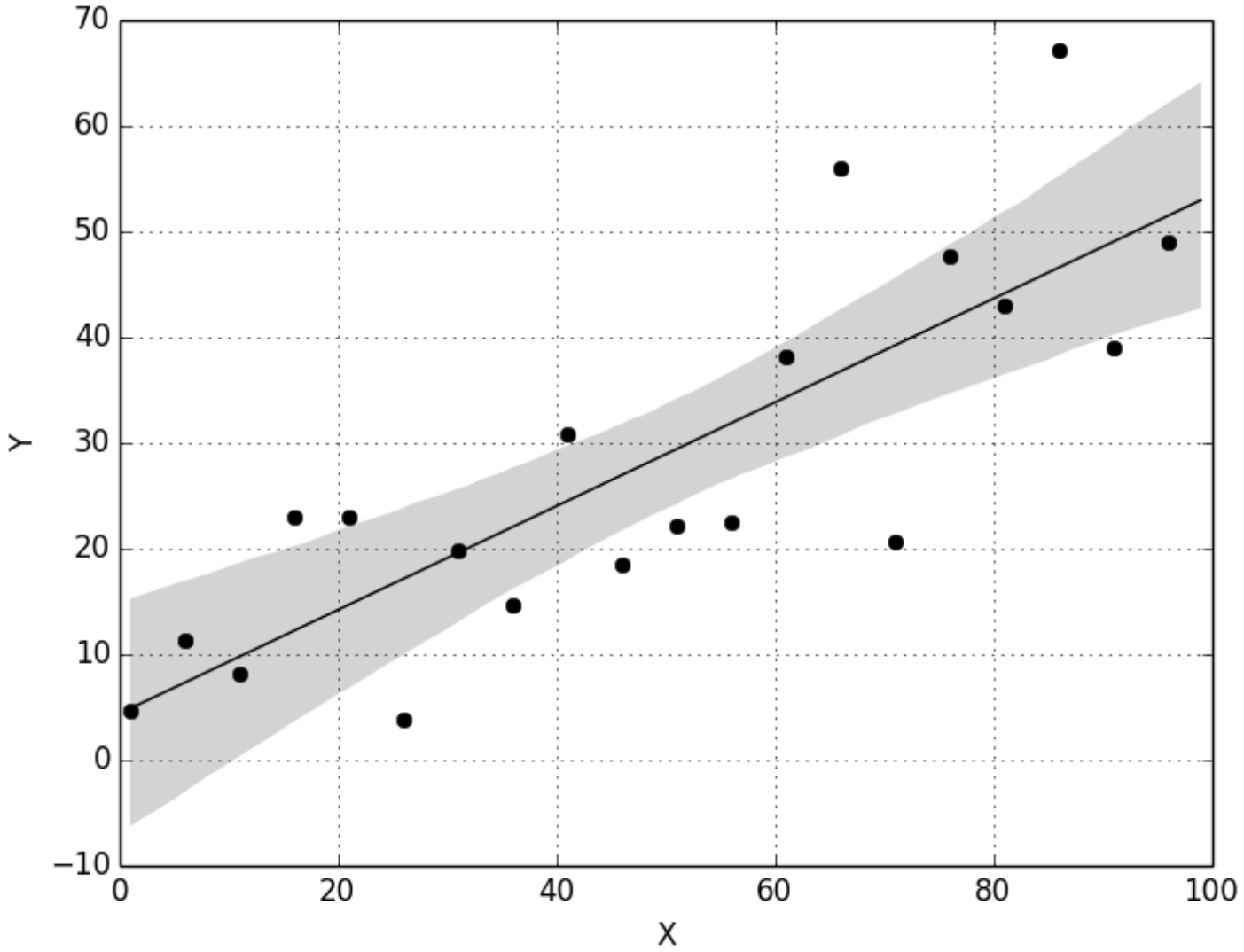
$$\mathbb{E}((y^*|x^*, D)) = \frac{1}{\sigma^2}(x^*)^t A^{-1} X^t Y$$

definiert. Die Varianz

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

dient als Mass der Güte der Vorhersage.

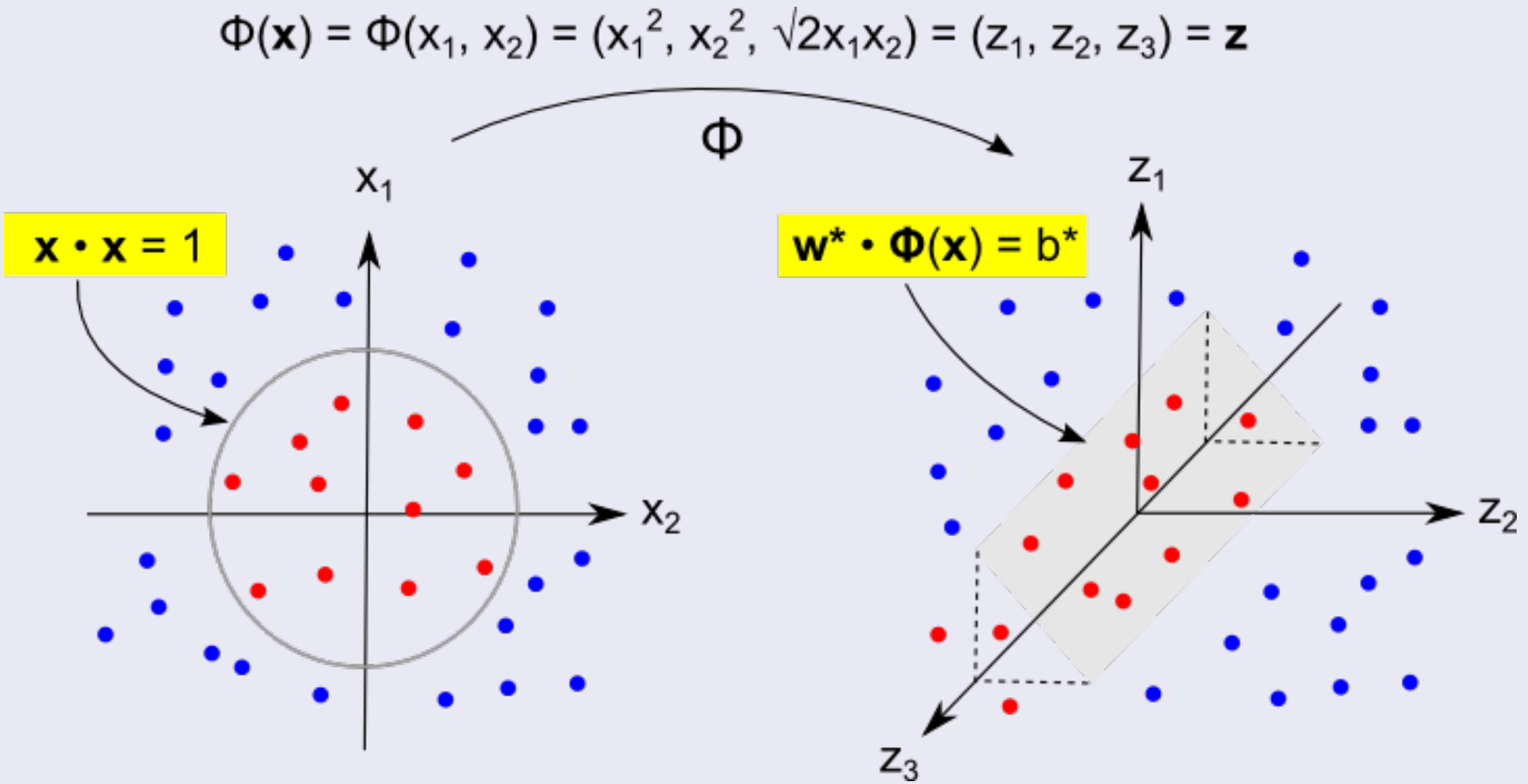
Lineare Bayessche Regression



Kernel Trick

$f(X) = \phi(X)^t \cdot \omega$ mit $\phi : \mathbb{R}^n \to \mathbb{R}^N$

Beispiel



Bayessche Regression

Analoge Rechnung zu linearer Bayesscher Regression:

$$f_{x*}|x*, D \sim \mathcal{N}\left(\frac{1}{\sigma^2}\underbrace{\phi(x^*)^t A^{-1}\phi(x)}_{:=k(x^*,x)} Y,\ \phi(x^*)^t A^{-1}\phi(x^*)\right)$$

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

Kernel

Funktionen der From

$$k(x^*, x) = \psi(x^*)^t\psi(x)$$

werden Kernel oder Kovarianzfunktionen genannt.

Mercers Theorem

Ist $k(s,t) = k(t,s)$ und $\int_{X\times X} k(s,t)f(s)f(t) > 0$ für alle $f \in L^2(X)$, dann ist

$$k(s,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s)\phi_i(t)$$

wobei ϕ_i die Eigenfunktionen zum Eigenwert λ_i des linearen Operators $T_k f := \int_X k(\cdot,t)f(t)$ sind.

Gauß-Prozess-Regression

Stochastischer Prozess

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

Gauß-Prozess

Wir bezeichnen einen stochastischen Prozess als Gauß-Prozess $f_x \sim \mathcal{GP}\left(m(x), k(x, x')\right)$, 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 $X = (x_1, \cdots, x_n) \in \mathcal{X}$. 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 $\mathcal{GP}\left(m(x), k(x, x')\right)$ mit

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

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

Gauß-Prozess-Regression

Prior distribution

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

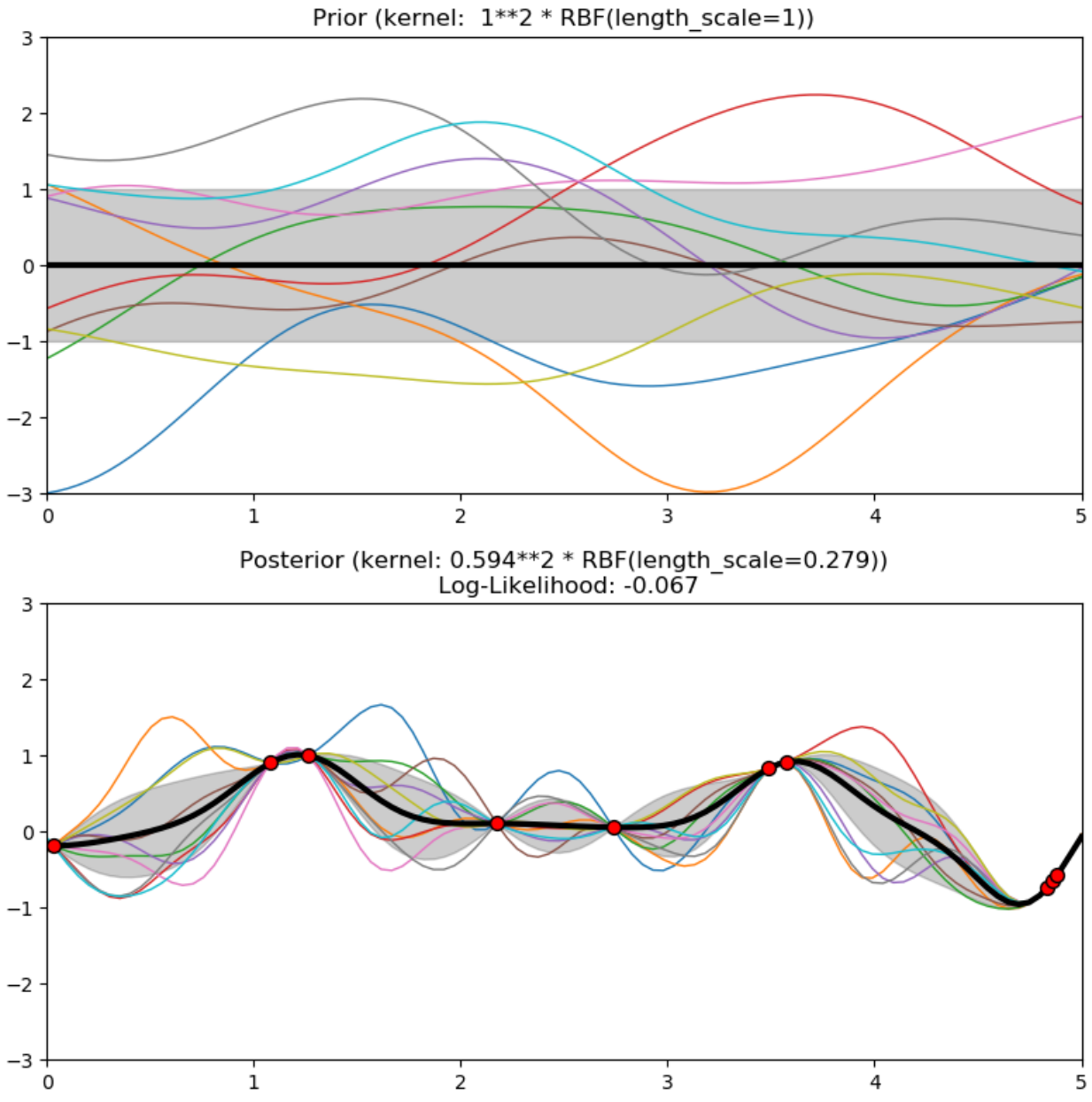
$$\begin{pmatrix} f_X \\ f_{X^*} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(X, X) & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{pmatrix} \right)$$

Posterior predictive distribution

Mit den Rechenregeln für multivariate Normalverteilungen folgt:

$$f_{X^*}|X^*, f_X, X \sim \mathcal{N}(\mu, \Sigma)$$
$$\mu := K(X^*, X)K(X, X)^{-1} \cdot f$$
$$\Sigma := K(X^*, X^*) - K(X^*, X)K(X, X)^{-1}K(X, X^*)$$

Gauß-Prozess-Regression



Vorhersage

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

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

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	σ_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	$\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{l} r\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{l} r\right)$
expotenziell	$exp\left(-\frac{r}{l}\right)$
γ-expotenziell	$exp\left(-\left(\frac{r}{l}\right)^{\gamma}\right)$
rational quadratisch	$\left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha}$

Qualitätsmodell

Qualitätsmodell

Gegeben Menge von Sensoren $\{(p_i, y_i)\}$ mit Position p_i und Messwert y_i unbekannter Qualität. Weise jedem Sensor einen Parameter $q_i \in (0, \infty)$ zu.

Kernel

$$K((p_i, q_i), (p_j, q_j)) := K_{Matern}(p_i, p_j) + K_q(q_i, q_j)$$

$$K_q(q_i, q_j) := \begin{cases} \frac{1}{q_i^2} \text{ i =j} \\ 0 \text{ else} \end{cases}$$

Modell mit Hyperparametern

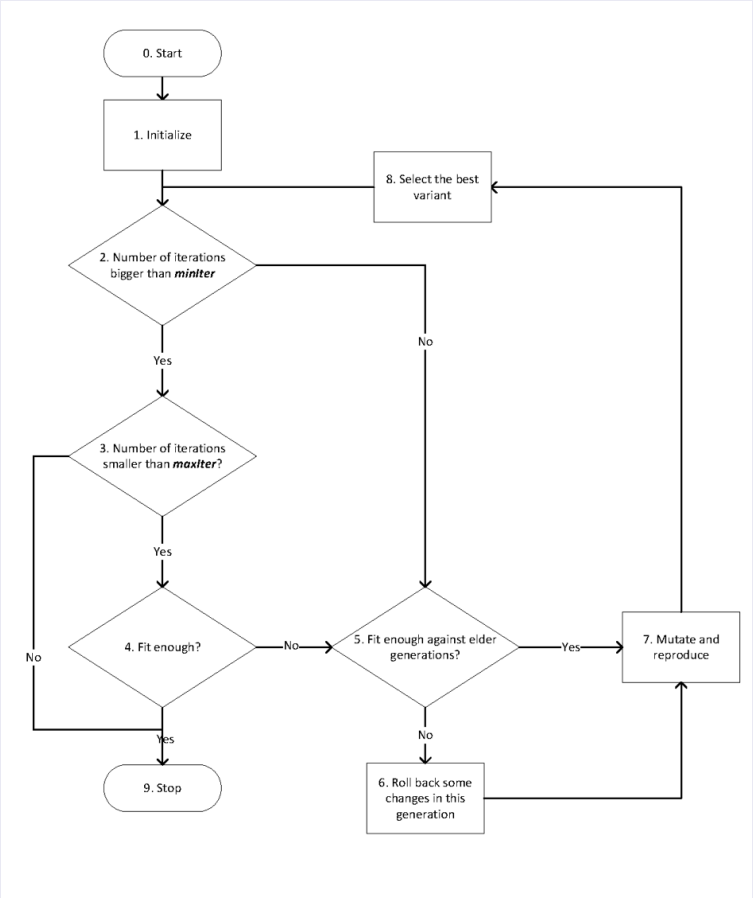
Modell

Erhalten damit Regressionmodell $GP(Q)$ mit Hyperparametern $Q = \{(q_i)\}$

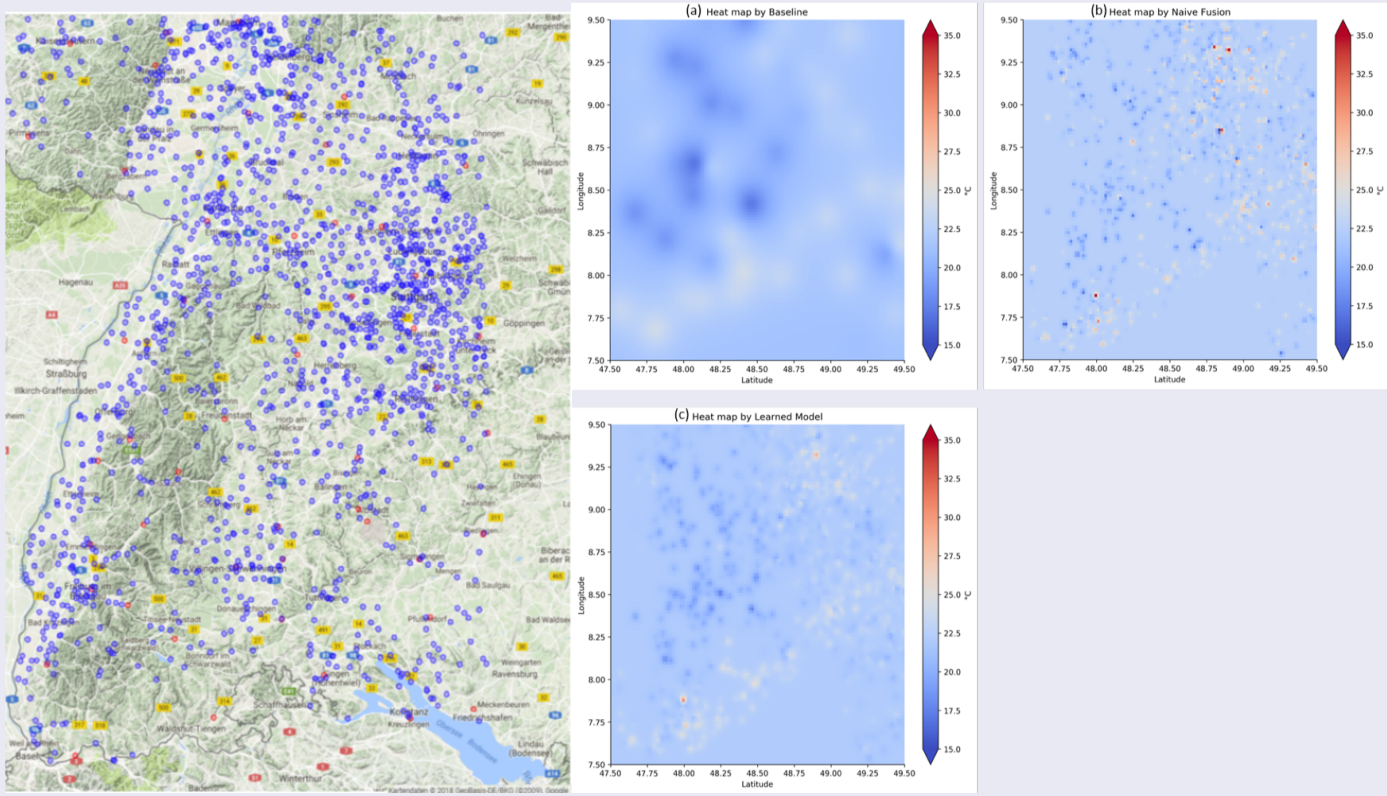
Optimierung

Optimiere $GP(Q)$ bzgl Q. Zum Beispiel mit Hilfe eines genetischen Algorithmus.

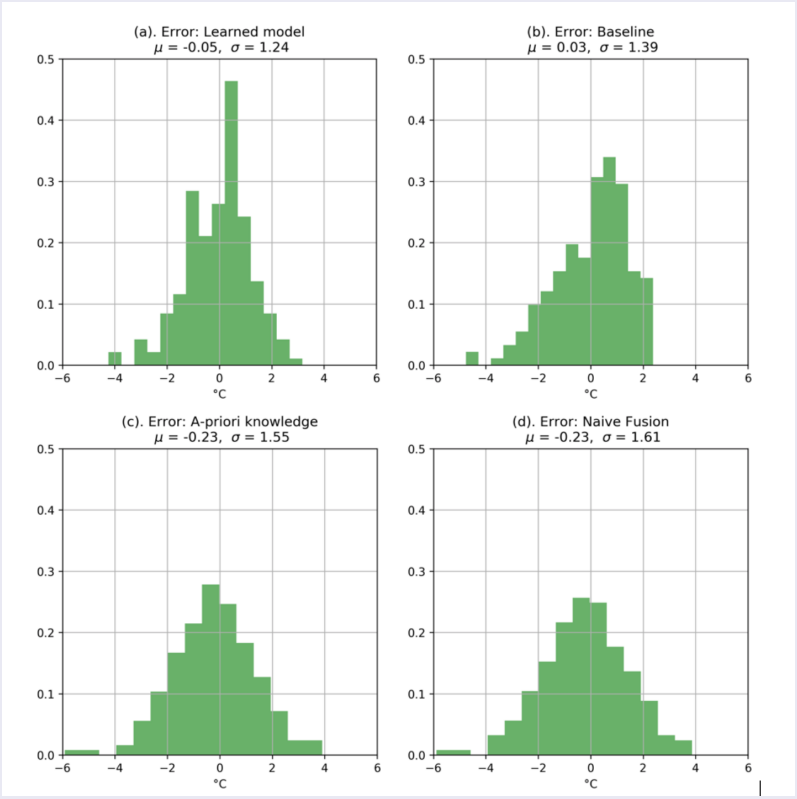
Ablaufdiagramm



Fehleranalyse



Hypothesentest



Signifikant zum Signifikanzniveau $\alpha = 0.05$