Application of Gaussian Process regression to the Estimation of time-in-therapeutic-range (TTR) and prediction of the international normalized ratio (INR) in patients under therapy with Vitamin K antagonists

Franz Ruderich

February 5, 2017

1 Data and methods

1.1 Gaussian Processes for Regression

For noisy measurements of (laboratory) values at certain time points it is difficult to fix what the best estimates of values between the time points or at future time points are. The assumption of an underlying linear function makes it easy to fit a straight line by least-squares method. But this is often a oversimplification of the problem. The model will provide poor predictions, if the relationship between dependent and independent variable can not be approximated by a linear function. Even a polynomial regression models the changes of laboratory measurements over time only unsatisfactorily with the problem of over-fitting Regression with Gaussian processes is a smarter method to model the relationship between successive measured values. A Gaussian Process does not assume some specific model underlying the data but calculates the relationship between single data points. In principle are Gaussian Processes a generalization of multivariate Gaussian distributions toward infinite dimensionality. The n observations of a (laboratory) data set $y = \{y_1, y_2, \ldots, y_n\}$ measured at the time points $t = \{t_1, t_2, \ldots, t_n\}$ could be imagined as a sample of a multivariate, n-variate Gaussian distribution.

A Gaussian process of a real process $f(t)$ is characterized by a mean function $m(t)$ and covariance function $k(t, t')$:

$$m(t) = E[f(t)]$$
$$k(t, t') = E[(f(t) - m(t))(f(t') - m(t'))]$$

In the case of repeated laboratory measurements the function values $f(t)$ represent the laboratory value measured at the time point t.

A typical covariance function is the squared exponential covariance function:

$$cov(f(t_p), f(t_q)) = k(t_p, t_q) = e^{-\frac{(t_p - t_q)^2}{2*l^2}}$$

1.2 Estimation of hyperparameters

Maximum likelihood estimator:

- unbiased: $E(\hat{\theta}) = \theta$
- consistent: $\hat{\theta} \to \theta$, as $n \to \infty$
- efficient: small $SE(\hat{\theta})$, as $n \to \infty$
- asymptotically normal: $\frac{(\hat{\theta} - \theta)}{SE(\hat{\theta})} \sim N(0, 1)$

1.2.1 Accuracy of Maximum Likelihood Estimation - The Fisher Information

The estimation of parameters $\boldsymbol{\theta}$ by maximizing the log-likelihood function means to find local maxima of this function. The first derivation of the log-likelihood function, the score, is equated to 0:

$$\frac{\partial \ell(\theta; x)}{\partial \theta} = \frac{\partial \log p(x; \theta)}{\partial \theta} = 0$$

The maxima are found by solving this equation. But it is necessary to know how accurate this estiamtion is. This information is given by calculating the curvature of the likelihood function around the maxima. A sharp curvature around the maximum means a high certainty, while a flat course gives a signal for a quite uncertain estimation. A measure for the curvature

of the score function is the variance of the score, called Fisher Information $I(\theta)$. The variance of the score is calculated by this formula:

$$I(\theta) = V[\frac{\partial \ell(\theta, x)}{\partial \theta}] = E[(\frac{\partial}{\partial \theta}\ell(\theta; x))^2] = -E[\frac{\partial^2 \ell(\theta; x)}{\partial \theta^2}]$$

The MLE is asymptotically normal distributed, so a asymptotic normal approximation and calculation of a confidence interval is possible:

$$\hat{\theta} \pm 1.96\frac{1}{\sqrt{I(\hat{\theta})}}$$