DS-GA 3001.001 Special Topics in Data Science: Probabilistic Time Series Analysis
Homework 4

Due date: Nov 22
YG390

Problem 1. (15pt) Which of these objects are a Gaussian process?

- linear combination of 2 GPs: $f(x) = af_1(x) + bf_2(x)$ where $f_i \sim \mathcal{GP}(\mu_i(x); k_i(x, y))$ (independent) and a, b are fixed parameters. The distribution of the sum of the functions due to linearity of Gaussian distributions is simply another $\mathcal{GP}$: $f(x) \sim \mathcal{GP}(\mu_1(x) + \mu_2(x); k_1(x, y) + k_2(x, y))$

- random linear: $f(x) = ax + w$ where $a \sim \mathcal{N}(0, \sigma_a^2)$, $w \sim \mathcal{N}(0, \sigma_w^2)$. We consider two fixed x_1 and x_2 and note that the two random variables $f(x_1)$ and $f(x_2)$ are jointly Gaussian.

$$\begin{aligned}
\mathrm{E}[f(x_i)] &= \mathrm{E}[ax_i + w] \\
&= x_i \,\mathrm{E}[a] + \mathrm{E}[w] \\
&= 0 \\
\mathrm{Var}[f(x)] &= \mathrm{Var}[ax + w] \\
&= x^2\mathrm{Var}[a] + \mathrm{Var}[w] \\
&= x^2\sigma_a^2 + \sigma_w^2 \\
\mathrm{Covar}[f(x_1), f(x_2)] &= \mathrm{E}(ax_1 + w)(ax_2 + w)] \\
&= x_1x_2 \,\mathrm{E}[a, a] + \mathrm{E}[w, w] \\
&= x_1x_2\sigma_a^2 + \sigma_w^2
\end{aligned}$$

Let $y_1 = f(x_1)$ and $y_2 = f(x_2)$, the multivariate gaussian distribution $p(y_1, y_2)$ is:

$$p(y_1, y_2) = \mathcal{N}\Big(\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} x_1^2\sigma_a^2 + \sigma_w^2 & x_1x_2\sigma_a^2 + \sigma_w^2 \\ x_1x_2\sigma_a^2 + \sigma_w^2 & x_2^2\sigma_a^2 + \sigma_w^2 \end{bmatrix}\Big)$$

Thus $f(x) \sim \mathcal{GP}(\mu(x_1, x_2); \Sigma(x_1, x_2))$.

- random periodic: $f(x) = a\cos(wx) + b\sin(wx)$ with $a \sim \mathcal{N}(0, \sigma^2)$, $b \sim \mathcal{N}(0, \sigma^2)$, w fixed parameter.

"From Professor Savin: for all pairs of variables: if nothing about their dependency structure is explicitly specified, they are independent." Similarly, we have now:

$$\begin{aligned}
\mathrm{E}[f(x_i)] &= \mathrm{E}[a\cos(wx_i) + b\sin(wx_i)] \\
&= \cos(wx_i)\,\mathrm{E}[a] + \sin(wx_i)\,\mathrm{E}[b] \\
&= 0 \\
\mathrm{Var}[f(x)] &= \mathrm{Var}[a\cos(wx) + b\sin(wx)] \\
&= \cos(wx)^2\mathrm{Var}[a] + \sin(wx)^2\mathrm{Var}[w] \\
&= \sigma^2 \\
\mathrm{Covar}[f(x_1), f(x_2)] &= \mathrm{E}[(a\cos(wx_1) + b\sin(wx_1))(a\cos(wx_2) + b\sin(wx_2))] \\
&= \cos(wx_1)\cos(wx_2)\,\mathrm{E}[a, a] + \sin(wx_1)\sin(wx_2)\,\mathrm{E}[w, w] \\
&= \cos(wx_1)\cos(wx_2)\sigma^2 + \sin(wx_1)\sin(wx_2)\sigma^2 \\
&= \sigma^2\cos(w(x_1 - x_2))
\end{aligned}$$

$f(x)$ is a Gaussian process: $f(x) \sim \mathcal{GP}(\mu(x_1, x_2); \Sigma(x_1, x_2))$, where $\mu(x_1, x_2) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \sigma^2 & \sigma^2\cos(w(x_1 - x_2)) \\ \sigma^2\cos(w(x_1 - x_2)) & \sigma^2 \end{bmatrix}\Big)$

If yes, then write down the corresponding mean and covariance functions.

Problem 2. (10pt) How would you construct a GP-equivalent of an ARIMA (1,2,1) model?

An ARIMA (1,2,1) model x_t could be defined as the combination of a trend and some noise: $x_t = \mu_t + y_t$ where $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$ and y_t is an ARMA(1,1). Differencing such process leads to a stationary process: $\Delta^2 x_t = \beta_2 + \Delta^2 y_t$ which is stationary (constant mean), the corresponding GP-equivalent has for mean function $m(.) = \beta_0 + \beta_1 t + \beta_2 t^2$. The ARMA y_t process to be sensible as to be causal and suppose that, $y_t = \phi y_{t-1} + w_t + \theta w_{t-1}$ where $|\phi| < 1$ and $w_t \sim \mathcal{N}(0, \sigma_w)$. The autocovariance function satisfies:

$$\gamma(h) - \phi\gamma(h-1) = 0, h = 2, 3, \cdots$$

And the general solution is:

$$\gamma(h) = c\phi^h, h = 1, 2, \cdots$$

The initial conditions are

$$\gamma(0) = \phi\gamma(1) + \sigma_w^2[1 + \theta\phi + \theta^2]$$
$$\gamma(1) = \phi\gamma(0) + \sigma_w^2\theta$$

Solving for $\gamma(0)$ and $\gamma(1)$, we obtain:

$$\gamma(0) = \sigma_w^2 \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2} \text{ and } \gamma(1) = \sigma_w^2 \frac{(1 + \theta\phi)(\theta + \phi)}{1 - \phi^2}$$

Dividing by $\gamma(0)$ yields:

$$\rho(h) = \frac{(1 + \theta\phi)(\theta + \phi)}{1 + 2\theta\phi + \theta^2}\phi^{h-1}, h \geq 1$$

The squared exponential kernel with covariance function is defined as:

$$k(x_i, x_j) = \sigma^2 \exp\left(\frac{-(x_1 - x_2)^2}{2l^2}\right)$$

where

- The length scale l determines the length of the "wiggles". For x_t, using ρ as l, the GP process cannot extrapolate more than ρ time steps away: as x_{t1} and x_{t2} are less correlated, $|\rho|$ tends to zero, the exponential and $k(x_{t1}, x_{t2})$ tend to zero.
- The output variance σ^2 determines the average distance of your function away from its mean, we use $\sigma^2 = \gamma(0)$.

The Matérn covariance function is the generalization of the squared exponential kernel and used to define the statistical covariance between measurements made at two points that are d units distant from each other:

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\Big(\sqrt{2\nu}\frac{d}{\rho}\Big)^\nu K_\nu\Big(\sqrt{2\nu}\frac{d}{\rho}\Big)$$

Then the process x_t is the discrete time equivalents of Gaussian process models with Matérn covariance function with $\nu = \frac{1}{2}$ and $p = 0$. The GP-equivalent of an ARIMA (1,2,1) model is: $\mathcal{GP}(\beta_0 + \beta_1 t + \beta_2 t^2; \sigma^2 \exp -(\frac{|x_1 - x_2|}{\rho}))$ where $\sigma^2 = \gamma(0)$ and ρ are given by the expressions above.

Problem 3. (15pt) Derive the mean and covariance of $\mathrm{P}(y|\theta)$ for the FITC approximation described in the lecture (this is obtained by marginalizing out $\mathbf{u}$ and $\mathbf{f}$).

Hint: one can think of the approximate model as a sequence of linear gaussian steps and use the usual simple gaussid.pdf properties.

Given the two factor graphs in the slides of the lecture, $\{y_i\}$ are conditionally independent given $\{f_i\}$ and depend directly on the $\{f_i\}$ which are conditionally independent given $\{u_i\}$. The $\boldsymbol{u}$ variables summarize the dependencies in $\boldsymbol{f}$. Setting $p(\boldsymbol{u}) = \mathcal{N}(0, \boldsymbol{K_{uu}})$, the conditional distribution $p(f_t|\boldsymbol{u}) = \mathcal{N}(f_t; K_{f_tu}K_{uu}^{-1}\boldsymbol{u}, K_{f_tf_t} - K_{f_tu}K_{uu}^{-1}K_{uf_t})$.

We have then:

$$\begin{aligned}
\boldsymbol{y_t} &= \boldsymbol{f_t} + \sigma_y \boldsymbol{\epsilon} \text{ with } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\
\text{let } D &= K_{f_tf_t} - K_{f_tu}K_{uu}^{-1}K_{uf_t} \\
\boldsymbol{f_t} &= K_{f_tu}K_{uu}^{-1}\boldsymbol{u_t} + D\boldsymbol{\epsilon'} \text{ with } \boldsymbol{\epsilon'} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})
\end{aligned}$$

$\boldsymbol{f_t}$ has 0 mean as it only depends on $\boldsymbol{u_t}$ which has 0 mean, the same for $\boldsymbol{y_t}$ which depends only on $\boldsymbol{f_t}$ with some noise. Now:

$$\begin{aligned}
\mathrm{Covar}[\boldsymbol{y_t}, \boldsymbol{y_s}] &= \mathop{\mathrm{E}}_{\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{\epsilon}}[(\boldsymbol{f_t} + \sigma_y\boldsymbol{\epsilon})(\boldsymbol{f_s} + \sigma_y\boldsymbol{\epsilon})^T] \\
&= \mathrm{E}[\boldsymbol{f_t}, \boldsymbol{f_s}] + \sigma_y^2 \, \mathrm{E}[\boldsymbol{\epsilon}, \boldsymbol{\epsilon}] \\
&= \mathrm{E}[\boldsymbol{f_t}, \boldsymbol{f_s}] + \sigma_y^2 \boldsymbol{I} \\
\mathrm{E}[\boldsymbol{f_t}, \boldsymbol{f_s}] &= \mathrm{E}[K_{f_tu}K_{uu}^{-1}\boldsymbol{u_t}\boldsymbol{u_s}K_{uu}^{-T}K_{uf_s}] + D \, \mathrm{E}[\boldsymbol{\epsilon'}, \boldsymbol{\epsilon'}] \\
&= K_{f_tu}K_{uu}^{-1} \, \mathrm{E}[\boldsymbol{u_t}, \boldsymbol{u_s}]K_{uu}^{-T}K_{uf_s} + D\boldsymbol{I} \\
&= K_{f_tu}K_{uu}^{-1}K_{uu}K_{uu}^{-1}K_{uf_s} + D \\
&= K_{f_tu}K_{uu}^{-1}K_{uf_s} + D
\end{aligned}$$

Where the cross-terms in the second and fourth equalities disappear as these terms are mutually independent. Thus

$$\mathrm{P}(\boldsymbol{y}|\theta) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf} + \mathbf{D} + \sigma_y^2\mathbf{I}\right).$$

Problem 4. (10pt) What GP-based model would you use for the Johnson&Johnson quarterly earnings database? Explain your choices. Would it matter if the goal of your analysis is to interpolate to account for missing data in the middle of the recorded time interval vs. extrapolating a decade into the future?

I used a linear combination of squared exponential kernel, linear, polynomial and periodic kernels. I tested the effect of the number of observations (sampling of the data) for the interpolation and forecasting (see 1). Experimenting with different weighting of the kernels, I found that, a linear combination of an SE and a periodic kernels captured the best the increasing trend, and some of the wiggles of the time series (see 2). The error bars showed that, when the observations were sparse the GP model confidence interval increased very rapidly confirming that, the GP model had larger confidence intervals for its predictions (see 2: third and fourth points and forecasting beyond the last observation 3). The same combination of kernels for the forecasting task, for different horizons, had a harder time to predict the wiggles of the process in the future (see 3). Note the difference between interpolation in the middle when the GP uses neighboring points (see 2: third and fourth points), vs. extrapolating in the future when the only available points are the last observations before the forecasting (see 3). So for interpolating to account for missing data, the best choice could be an SE + a linear or polynomial model, which takes into account clusters of points and for forecasting an Exp-Sine-Squared or a spectral mixture kernels might be more appropriate in an effort to gives weight to clusters of observations, but at the same time, to capture the dynamics of the process. Lastly, in order to improve the overall model performance, I used a spectral mixture kernel which seemed to predict the swings of the data including the ones in the future (see 4).

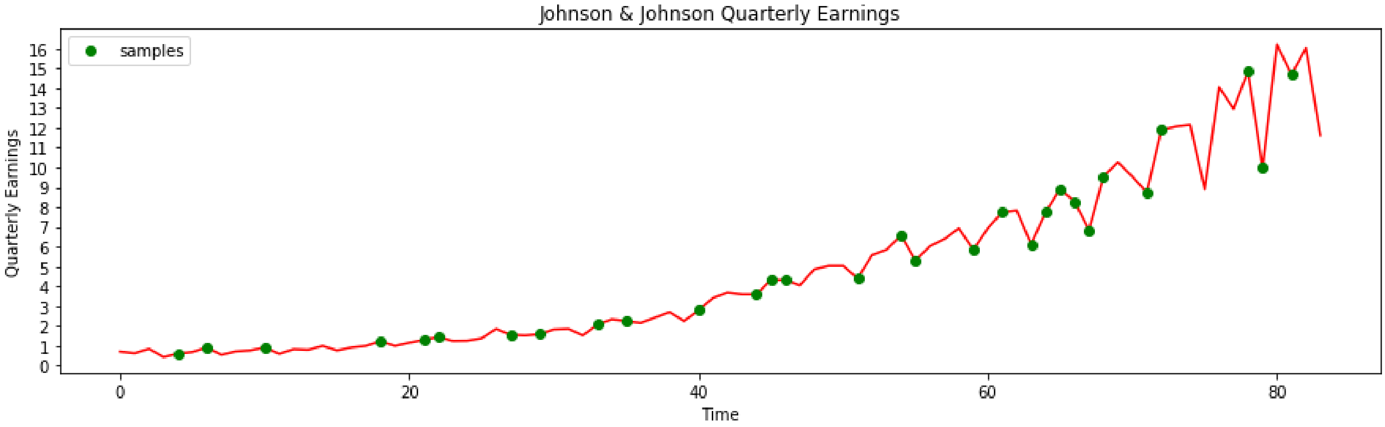
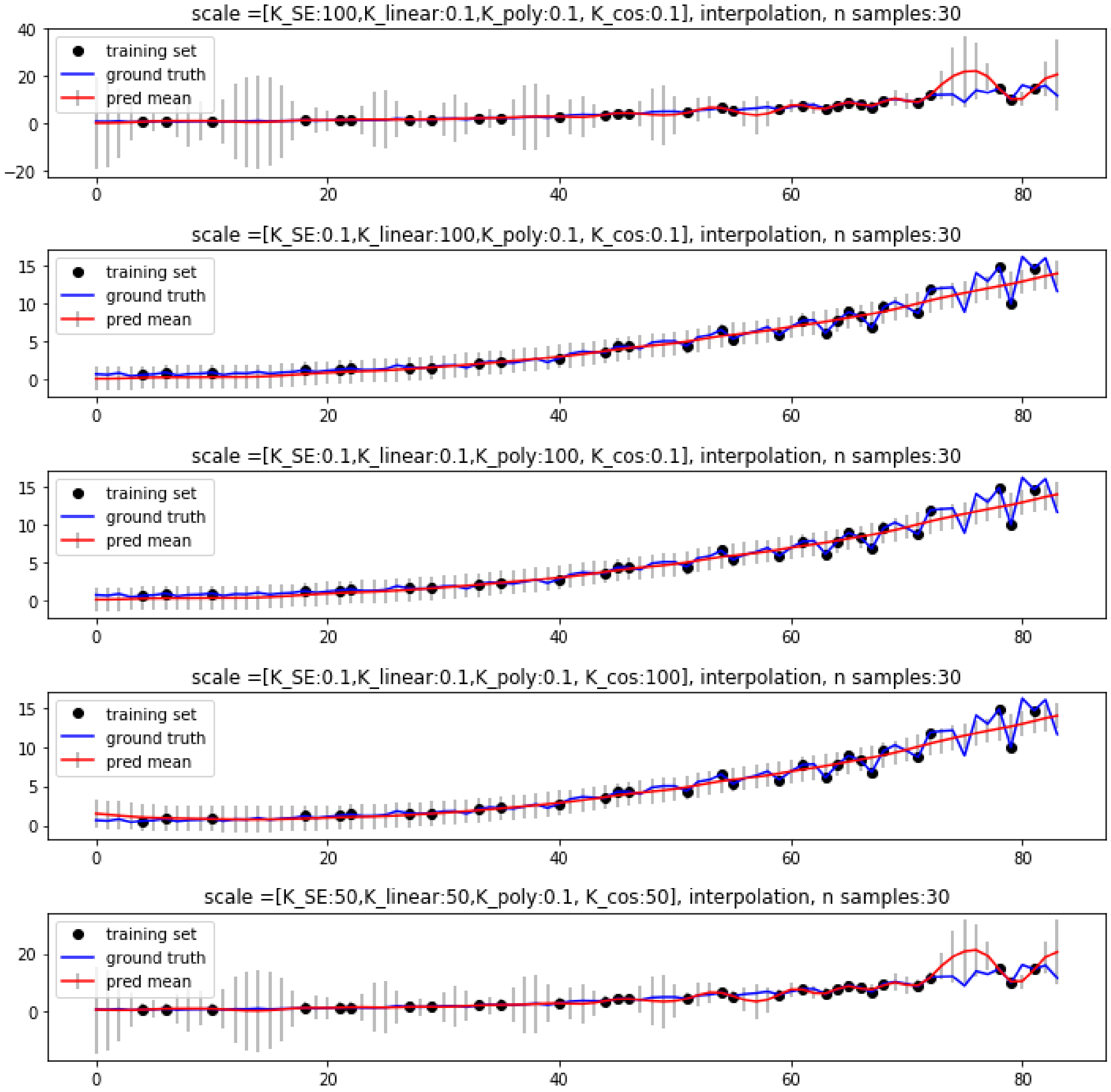


Figure 1: Johnson&Johnson



scale =[K_SE:100,K_linear:0.1,K_poly:0.1, K_cos:0.1], interpolation, n samples:30

scale =[K_SE:0.1,K_linear:100,K_poly:0.1, K_cos:0.1], interpolation, n samples:30

scale =[K_SE:0.1,K_linear:0.1,K_poly:100, K_cos:0.1], interpolation, n samples:30

scale =[K_SE:0.1,K_linear:0.1,K_poly:0.1, K_cos:100], interpolation, n samples:30

scale =[K_SE:50,K_linear:50,K_poly:0.1, K_cos:50], interpolation, n samples:30

Figure 2: Interpolation using a weighted combinations of kernels

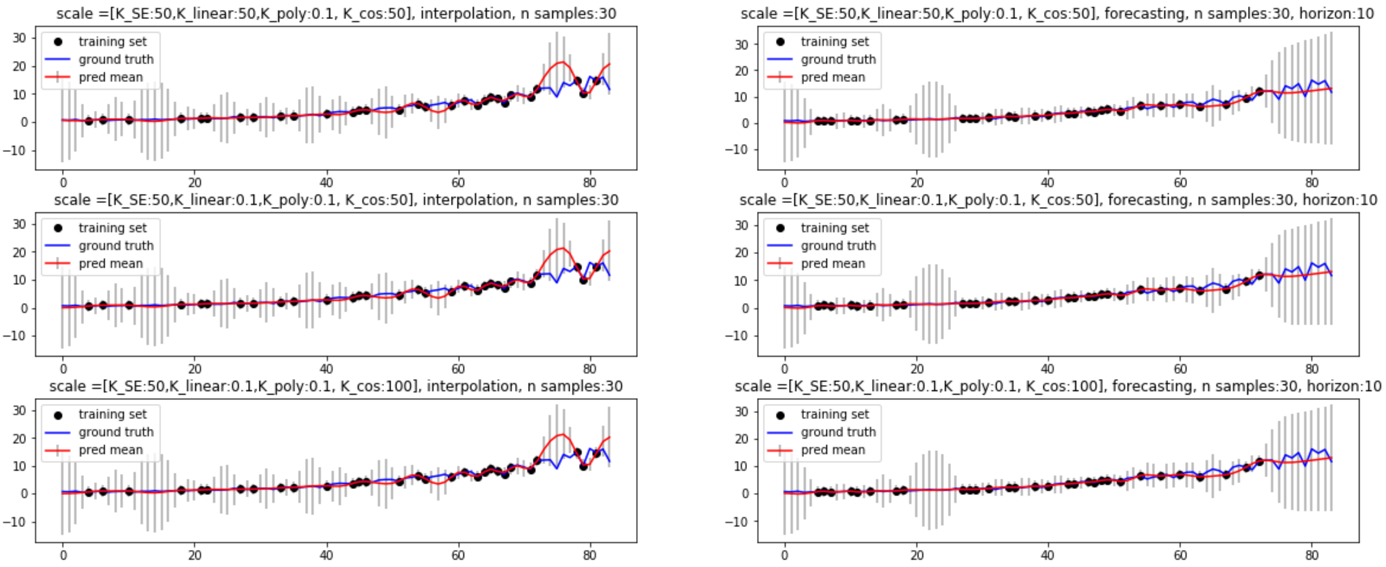


Figure 3: SE + periodic kernels - Interpolation and Forecasting with error bars

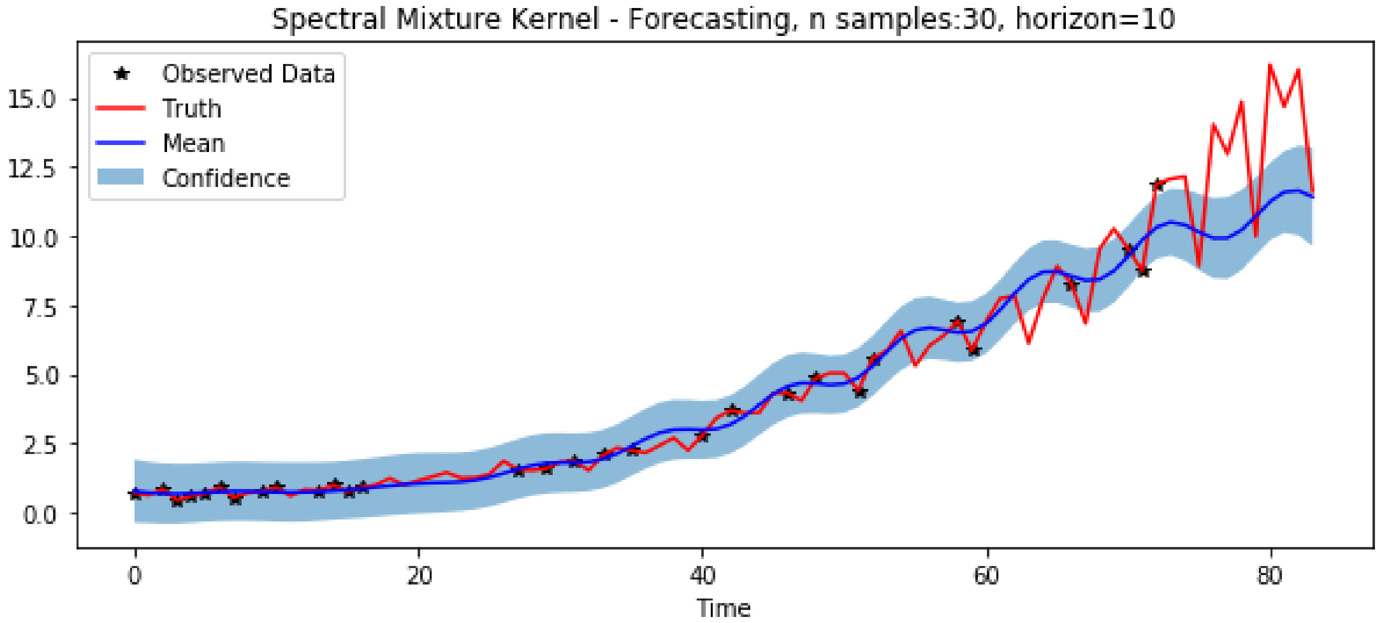


Figure 4: Spectral Mixture Kernel with confidence intervals - two standard deviations