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

Due date: April 26th, by midnight

Problem 1. (10pt) Derive the mean and covariance of $\mathrm{P}(y_t|\theta)$ for the FITC approximation.
Hint: think of the approximate model as a sequence of linear gaussian steps.
Solution:

The FITC approximate model can be viewed as a sequence of linear gaussian operations with:

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{uu})$$

$$\mathbf{f} \sim \mathcal{N}(\mathbf{A}\mathbf{u}, \mathbf{D})$$

where $\mathbf{A} = \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}$ and $\mathbf{D} = \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}$.
Finally

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma_y^2\mathbf{I})$$

.

After marginalizing out variables $\mathbf{u}$ and $\mathbf{f}$ the resulting distribution for $\mathbf{y}$ is also multivariate gaussian with mean:

$$\boldsymbol{\mu}_y = \boldsymbol{\mu}_f = \mathbf{A}\boldsymbol{\mu}_u = \mathbf{0}$$

and covariance

$$\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_f + \sigma_y^2\mathbf{I} = \mathbf{A}\mathbf{K}_{uu}\mathbf{A}^{\mathrm{t}} + \mathbf{D} + \sigma_y^2\mathbf{I}$$

where we have used the linear + gaussian noise properties (4a in gaussid.pdf).

The final solution is obtained by plugging in the value of $\mathbf{A}$ and using the fact that $\mathbf{K}_{uu}$ is symmetric[1]:

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

Problem 2. (10pt) Consider a simple GP regression model with gaussian observations: $\mathbf{f} \sim \mathcal{GP}(0, k(\cdot,\cdot))$, $y_i \sim \mathcal{N}(f_i, \sigma^2)$. Let $\mathrm{Var}(f(x^*))_n$ be the variance of the posterior for the value of function f evaluated at a test position x^*, conditioned on n observations $x_{1:n}, y_{1:n}$. If $\mathrm{Var}(f(x^*))_{n-1}$ denotes the same variance conditioned on the first $n-1$ observations, show that the posterior variance decreases by adding the last datapoint, i.e. $\mathrm{Var}(f(x^*))_n \leq \mathrm{Var}(f(x^*))_{n-1}$.
Hint: This is problem 4 from section 2.9 GP textbook (pg.31). See detailed suggestions there.

Solution: We're going to use the formula for the inversion of the partition matrix (A.11) to tackle the problem. First, some notation: I'll use $\mathbf{K}$ to denote the covariance based on points 1 to n-1 (size $n-1 \times n-1$), $\mathbf{k}$ to denote the covariance of the last point with the rest (size $n-1 \times 1$) and scalar k_n for the variance of f_n.

The variance of the prediction based on $n-1$ points is (using multivariate conditioning):

$$\mathrm{Var}(f(x^*))_{n-1} = k_{**} - \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_y^2\mathbf{I}_{n-1}\right)^{-1}\mathbf{k}_*$$

where scalar k_{**} is the prior variance at point x^*, vector $\mathbf{k}_*$ denotes the covariance of f_* and $f_{1:n-1}$ (size $n-1 \times 1$) and $\mathbf{I}_{n-1}$ is the identity matrix of size $n-1$.

Similarly, the variance of the prediction based on n points is :

$$\mathrm{Var}(f(x^*))_n = k_{**} - \begin{bmatrix}\mathbf{k}_* \\ k_{*n}\end{bmatrix}^\top \left(\begin{bmatrix}\mathbf{K} & \mathbf{k} \\ \mathbf{k}^\top & k_n\end{bmatrix} + \sigma_y^2\mathbf{I}_n\right)^{-1}\begin{bmatrix}\mathbf{k}_* \\ k_{*n}\end{bmatrix} \tag{1}$$

where scalar k_{*n} is the covariance between the added n-th point and the test.

[1] To simplify notation we had dropped the dependence on hyperparameters θ in the above, but all matrices $\mathbf{K}$ are determined by a kernel function $k(\cdot,\cdot)$ parametrized by θ

Using the partitioning $P = \mathbf{K} + \sigma_y^2 \mathbf{I}_{n-1}$, $Q = \mathbf{k}$, $R = \mathbf{k}^\top$, $S = k_n + \sigma_y^2$ and applying the formula A.11 for the inversion of a partition matrix we get:

$$M = \left(k_n + \sigma_y^2 - \mathbf{k}^\top \left(\mathbf{K} + \sigma_y^2 \mathbf{I}_{n-1}\right)^{-1} \mathbf{k}\right)^{-1} = \frac{1}{\sigma_n^2} \tag{2}$$

$$\tilde{S} = M = \frac{1}{\sigma_n^2} \tag{3}$$

$$\tilde{Q} = \hat{R}^\top = -\frac{1}{\sigma_n^2} \left(\mathbf{K} + \sigma_y^2 \mathbf{I}_{n-1}\right)^{-1} \mathbf{k} \tag{4}$$

$$\tilde{P} = \left(\mathbf{K} + \sigma_y^2 \mathbf{I}_{n-1}\right)^{-1} + \frac{1}{\sigma_n^2} \left(\mathbf{K} + \sigma_y^2 \mathbf{I}_{n-1}\right)^{-1} \mathbf{k}\mathbf{k}^\top \left(\mathbf{K} + \sigma_y^2 \mathbf{I}_{n-1}\right)^{-1} \tag{5}$$

where have noticed that M is actually the precision for the prediction $f_n | f_{1:n-1}$.

The last thing we need to do is to plug in the inverted matrix above into the expression for $\mathrm{Var}(f(x_*))_n$ above and write out the expression for the quadratic form. This results in the final expression:

$$\mathrm{Var}(f(x^*))_n = k_{**} - \left(\mathbf{k}_*^\top \left(\mathbf{K} + \sigma_y^2 \mathbf{I}_{n-1}\right)^{-1} \mathbf{k}_* + \frac{1}{\sigma_n^2} \left(a^2 - 2\frac{a}{\sigma_n^2} + \frac{1}{\sigma_n^4}\right)\right) \tag{6}$$

where we used shorthand $a = \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_y^2 \mathbf{I}_{n-1}\right)^{-1} \mathbf{k}$ (this is a scalar).

Identifying that the first two terms are the expression for $\mathrm{Var}(f(x_*))_{n-1}$ and that the last term $-\left(a - \frac{1}{\sigma_n^2}\right)^2$ is negative finishes the proof.

Problem 3. extra credit (10pt) Fit a GP to the Johnson&Johnson quarterly earnings dataset. Chose 2 kernel variants and briefly justify your choices. Using the remaining data as training set, predict the values of the function in the period 1970-1975 and 1980-1990 (extrapolation in the future). Compare the two models by comparing the average prediction error at the test points 1970-1975 and 1980. Comment on results.

Solution: I used a simple squared exponential kernel, a periodic kernel, and a spectral mixture kernel (all zero mean). Any other justified choice of mean and kernel is possible here. The main signatures of the data you may want to capture with these choices: 1) increasing mean 2) smoothing 3) some form of periodicity.

Interpretation of results: The SE capture the increasing trend, but none of the fine scale structure – as expected since designed for smooth interpolation. The periodic kernel seems to capture best the fine scale regularities, and extrapolate to periodic predictions of the future. The SM kernel behaves very similar to SE in the dataset, because the amount of data is insufficient to constrain properly the hyperparameters, such that the ML estimates of θ are quite variable and a bit prone to overfitting (may also have to do with the initialization, but seems robust to variations in the number of components). Occasionally periodic solutions do come up, but not nearly as good as the explicitly periodic version. Interpolation seems to work well for all versions (you'll need to also report goodness of fit on 1970-1975 and 1980 test data).

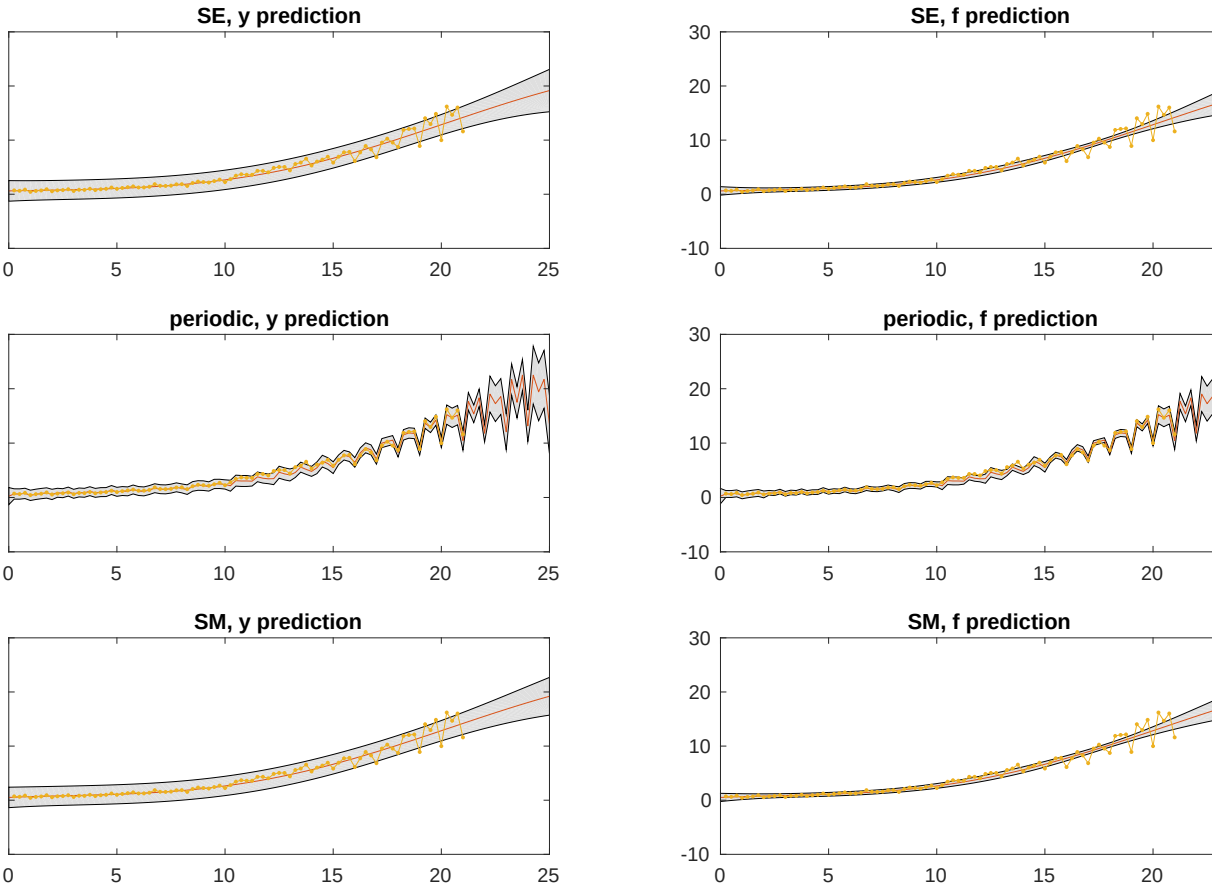

Figure 1: Visualization basic predictions. Plot uses quarters as time units starting from 1960. Data shown in yellow (includes test data), posterior mean in red and 95% confidence intervals in gray.