More on approximate methods for large datasets

Mauricio A. Álvarez, PhD

Curso de entrenamiento ArcelorMittal

A comment on notation

- Before, we used $\mathbf{K}_{\mathbf{f},\mathbf{f}}$ to refer to the matrix $\mathbf{K}(\mathbf{X}, \mathbf{X})$. In this slides, we will also use $\mathbf{K}_{nn}$ to refer to the same matrix.

Contents

Mean GP predictor (I)

- We can get the GP regression predictive equations if we start by defining

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i),$$

  where $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_n]^\top$ and

$$\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^{-1}).$$

- We have used $\mathbf{K} = \mathbf{K}(\mathbf{X}, \mathbf{X})$. We will also use $\mathbf{K}_{nn}$ to refer to the kernel matrix $\mathbf{K}(\mathbf{X}, \mathbf{X})$.

- For the Gaussian regression case and with training data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$, the likelihood function is

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\alpha}, \sigma_n^2) = \mathcal{N}(\mathbf{y}|\mathbf{K}\boldsymbol{\alpha}, \sigma_n^2\mathbf{I}).$$

Mean GP predictor (II)

- By using the prior on $\boldsymbol{\alpha}$, $p(\boldsymbol{\alpha})$, and the likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\alpha}, \sigma_n^2)$, we get the following posterior for $\boldsymbol{\alpha}$

$$p(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{X}, \sigma_n^2) = \mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\Sigma}\mathbf{K}^\top \sigma_n^{-2}\mathbf{y}, \boldsymbol{\Sigma})$$

  where $\boldsymbol{\Sigma} = (\mathbf{K} + \sigma_n^{-2}\mathbf{K}^\top\mathbf{K})^{-1}$.

- $\mathbf{K}$ is a symmetric matrix and

$$\boldsymbol{\Sigma} = (\mathbf{K} + \sigma_n^{-2}\mathbf{K}^\top\mathbf{K})^{-1} = [\sigma_n^{-2}\mathbf{K}(\sigma_n^2\mathbf{I} + \mathbf{K})]^{-1} = (\sigma_n^2\mathbf{I} + \mathbf{K})^{-1}\mathbf{K}^{-1}\sigma_n^2.$$

- The posterior mean in $p(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{X}, \sigma_n^2)$ is simply

$$\overline{\boldsymbol{\alpha}} = \boldsymbol{\Sigma}\mathbf{K}^\top\sigma_n^{-2}\mathbf{y} = (\sigma_n^2\mathbf{I} + \mathbf{K})^{-1}\mathbf{K}^{-1}\sigma_n^2\mathbf{K}\sigma_n^{-2}\mathbf{y} = (\sigma_n^2\mathbf{I} + \mathbf{K})^{-1}\mathbf{y}.$$

Mean GP predictor (III)

- We can now compute the predictive distribution for $p(f(\mathbf{x}_*)|\mathbf{X}, \mathbf{y})$ by marginalising $\boldsymbol{\alpha}$ using the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{X}, \sigma_n^2)$.

- According to what we saw before, $f(\mathbf{x}_*) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_*, \mathbf{x}_i) = \mathbf{k}^\top(\mathbf{x}_*)\boldsymbol{\alpha}$, where $\mathbf{k}^\top(\mathbf{x}_*) = [k(\mathbf{x}_*, \mathbf{x}_1), \cdots, k(\mathbf{x}_*, \mathbf{x}_n)]$.

- The mean predictive is then given as

$$\bar{f}(\mathbf{x}_*) = \mathbb{E}(f(\mathbf{x}_*)) = \mathbf{k}^\top(\mathbf{x}_*)\mathbb{E}(\boldsymbol{\alpha}) = \mathbf{k}^\top(\mathbf{x}_*)(\sigma_n^2\mathbf{I} + \mathbf{K})^{-1}\mathbf{y},$$

  which corresponds to the expression for the mean prediction in GP regression.

- What if instead of having $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$, with $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}^{-1})$, we use $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}_i)$ with $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{mm}^{-1})$?

What is $\mathbf{K}_{mm}$?

- Several methods, including Subset of Regressors, consider selecting a subset I of the n datapoints.
- The set I has size $m < n$.
- The remaining $n - m$ datapoints form the set R.
- I is the subset of included datapoints whereas R is the set of remaining datapoints.
- The matrix $\mathbf{K}$ can be partitioned as

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{mm} & \mathbf{K}_{m(n-m)} \\ \mathbf{K}_{(n-m)m} & \mathbf{K}_{(n-m)(n-m)} \end{bmatrix}$$

- A key difference with the inducing variable methods is that the set I is part of the n datapoints, whereas $\mathbf{u}$ can be any points.

Subset of regressors (I)

- We can consider a subset of regressors $m < n$ such that

$$f_{\text{SR}}(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}_i), \qquad \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{mm}^{-1}).$$

- Following the same procedure than before, the likelihood function is given as

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\alpha}, \sigma_n^2) = \mathcal{N}(\mathbf{y}|\mathbf{K}_{nm}\boldsymbol{\alpha}_m, \sigma_n^2\mathbf{I}).$$

- With the prior over $\boldsymbol{\alpha}_m$ and the likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\alpha}, \sigma_n^2)$, the posterior distribution over $\boldsymbol{\alpha}_m$ follows as

$$p(\boldsymbol{\alpha}_m|\mathbf{y}, \mathbf{X}, \sigma_n^2) = \mathcal{N}(\boldsymbol{\alpha}_m|\boldsymbol{\Sigma}_{mm}\mathbf{K}_{mn}\sigma_n^{-2}\mathbf{y}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}_{mm} = (\mathbf{K}_{mm} + \sigma_n^{-2}\mathbf{K}_{mn}\mathbf{K}_{nm})^{-1} = (\sigma_n^2\mathbf{K}_{mm} + \mathbf{K}_{mn}\mathbf{K}_{nm})^{-1}\sigma_n^2$.

Subset of regressors (II)

- Predictions are made using $f_{\text{SR}}(\mathbf{x}_*) = \mathbf{k}_m^\top(\mathbf{x}_*)\boldsymbol{\alpha}_m$, where

$$\mathbf{k}_m^\top(\mathbf{x}_*) = [k(\mathbf{x}_*, \mathbf{x}_1), \cdots, k(\mathbf{x}_*, \mathbf{x}_m)].$$

- Using the posterior distribution $p(\boldsymbol{\alpha}_m|\mathbf{y}, \mathbf{X}, \sigma_n^2)$ to marginalise $\boldsymbol{\alpha}_m$ from $f_{\text{SR}}(\mathbf{x}_*)$, the predictive distribution for $f_{\text{SR}}(\mathbf{x}_*)$ has moments

$$\bar{f}_{\text{SR}} = \mathbf{k}_m^\top(\mathbf{x}_*)(\sigma_n^2\mathbf{K}_{mm} + \mathbf{K}_{mn}\mathbf{K}_{nm})^{-1}\mathbf{K}_{mn}\mathbf{y}$$
$$\mathbb{V}[f_{\text{SR}}] = \sigma_n^2\mathbf{k}_m^\top(\mathbf{x}_*)(\sigma_n^2\mathbf{K}_{mm} + \mathbf{K}_{mn}\mathbf{K}_{nm})^{-1}\mathbf{k}_m(\mathbf{x}_*).$$

- Computational complexity is $\mathcal{O}(nm^2)$.

SR marginal likelihood

- Using

$$f_{\mathrm{SR}}(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}_i), \qquad \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{mm}^{-1}),$$

the marginal distribution for $p(\mathbf{f}_{\mathrm{SR}}) = \mathcal{N}(\mathbf{f}_{\mathrm{SR}}|\mathbf{0}, \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn})$.

- Let $\widetilde{\mathbf{K}} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$.

- The log-marginal likelihood under this model follows as

$$\log p_{\mathrm{SR}}(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\log|\widetilde{\mathbf{K}} + \sigma_n^2\mathbf{I}| - \frac{1}{2}\mathbf{y}^\top(\widetilde{\mathbf{K}} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{n}{2}\log(2\pi).$$

How to choose I?

- Randomly from $\mathbf{X}$.

- Run clustering over $\{\mathbf{x}_i\}_{i=1}^n$ and obtain m points that are closest to the m centres.

Contents

Approximation of the eigenfunctions of $k(\mathbf{x}, \mathbf{x}')$

- The Nyström method approximates the i eigenfunction of a kernel function $k(\mathbf{x}, \mathbf{x}')$ using

$$\phi_i(\mathbf{x}) \simeq \frac{\sqrt{n}}{\lambda_i^{\text{mat}}} \mathbf{k}^\top(\mathbf{x}) \mathbf{u}_i,$$

  where $\mathbf{k}^\top(\mathbf{x}) = [k(\mathbf{x}_1, \mathbf{x}), \ldots, k(\mathbf{x}_n, \mathbf{x})]$ and λ_i^{mat} and $\mathbf{u}_i$ are obtained from solving the matrix eigenproblem

$$\mathbf{K}\mathbf{u}_i = \lambda_i^{\text{mat}} \mathbf{u}_i.$$

- The eigenvectors are normalised $\mathbf{u}_i^\top \mathbf{u}_i = 1$.

Approximation of the eigenvectors of $\mathbf{K}$

- We compute the eigenvalues/vectors for $\mathbf{K}_{mm}$, denoted as $\{\lambda_i^{(m)}\}_{i=1}^m$ and $\left\{\mathbf{u}_i^{(m)}\right\}_{i=1}^m$.
- We use these to compute the eigenvalues/vectors for $\mathbf{K}$,

$$\begin{aligned}
\tilde{\lambda}_i^{(n)} &\triangleq \frac{n}{m}\lambda_i^{(m)}, & i = 1, \ldots, m \\
\tilde{\mathbf{u}}_i^{(n)} &\triangleq \sqrt{\frac{m}{n}}\frac{1}{\lambda_i^{(m)}}K_{nm}\mathbf{u}_i^{(m)}, & i = 1, \ldots, m
\end{aligned}$$

- We approximate $\mathbf{K}$ using

$$\mathbf{K} \approx \widetilde{\mathbf{K}} = \sum_{i=1}^{p} \tilde{\lambda}_i^{(n)}\tilde{\mathbf{u}}_i^{(n)} \left(\tilde{\mathbf{u}}_i^{(n)}\right)^\top.$$

- Setting $p = m$ then leads to

$$\widetilde{\mathbf{K}} = \sum_{i=1}^{m} \tilde{\lambda}_i^{(n)}\tilde{\mathbf{u}}_i^{(n)} \left(\tilde{\mathbf{u}}_i^{(n)}\right)^\top = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}.$$

$\mathbf{K}$ by $\widetilde{\mathbf{K}}$

- The Nyström method replaces $\mathbf{K}$ by $\widetilde{\mathbf{K}} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$ in the mean and variance prediction equations of GP regression.

- The original GP predictive distribution $p(f(\mathbf{x}_*)|\mathbf{X}, \mathbf{y})$ has moments

$$\bar{f}(\mathbf{x}_*) = \mathbf{k}^\top(\mathbf{x}_*)\left[\mathbf{K} + \sigma_n^2\mathbf{I}\right]^{-1}\mathbf{y}$$

$$\mathbb{V}(f(\mathbf{x}_*)) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^\top(\mathbf{x}_*)\left[\mathbf{K} + \sigma_n^2\mathbf{I}\right]^{-1}\mathbf{k}(\mathbf{x}_*).$$

- For the Nyström method, the predictive distribution $p(f(\mathbf{x}_*)|\mathbf{X}, \mathbf{y})$ has moments

$$\bar{f}_N(\mathbf{x}_*) = \mathbf{k}^\top(\mathbf{x}_*)\left[\widetilde{\mathbf{K}} + \sigma_n^2\mathbf{I}\right]^{-1}\mathbf{y}$$

$$\mathbb{V}(f_N(\mathbf{x}_*)) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^\top(\mathbf{x}_*)\left[\widetilde{\mathbf{K}} + \sigma_n^2\mathbf{I}\right]^{-1}\mathbf{k}(\mathbf{x}_*).$$

Contents

Subset of datapoints

- A simple approximation to the full-sample GP predictor is to keep the GP predictor on a smaller subset of size m of the data.
- We can use a greedy algorithm to select which points are taken into the active set I.

Greedy approximation

- The algorithm starts with the active set I being empty, and the set R containing the indices of all training examples.

- On each iteration one index is selected from R and added to I.

- This is achieved by evaluating some criterion Δ and selecting the data point that optimizes this criterion.

- It can be too expensive to evaluate Δ on all points in R, so some working set $J \subset R$ can be chosen instead, usually at random from R.

Algorithm: greedy approximation

```
input: m, desired size of active set
Initialization I = ∅, R = {1, ..., n}
for j := 1 ... m do
    Create working set J ⊆ R
    Compute Δ_j for all j ∈ J
    i = argmax_{j∈J} Δ_j
    Update model to include data from example i
    I ← I ∪ {i}, R ← R\{i}
end for
return: I
```

Algorithm 8.1: General framework for greedy subset selection. Δ_j is the criterion function evaluated on data point j.

Selection criteria

- The *informative vector machine* (IVM) (Lawrence et al., 2003) efficiently computes the *differential entropy score*

$$\Delta_j \triangleq H\left[p\left(f_j\right)\right] - H\left[p^{\text{new}}\left(f_j\right)\right],$$

where $H\left[p\left(f_j\right)\right]$ is the entropy of the Gaussian process at $j \in R$ without including observation j and $H\left[p^{\text{new}}\left(f_j\right)\right]$ is the entropy at $j \in R$ when including the observation j.

- The *information gain* criterion $\mathrm{KL}\left(p^{\text{new}}\left(f_j\right) \| p\left(f_j\right)\right)$ can also be used as a selection criterion (Seeger, 2003).

Contents

Split the data into p parts

- Let $\mathbf{f}_*$ be the vector of function values at the test locations.
- The Bayesian committee machine (BCM) splits the dataset into p parts, $\mathcal{D}_1, \ldots, \mathcal{D}_p$, where $\mathcal{D}_i = \{\mathbf{X}_i, \mathbf{y}_i\}$.
- It assumes that

$$p\left(\mathbf{y}_1, \ldots, \mathbf{y}_p \mid \mathbf{f}_*, \mathbf{X}\right) \simeq \prod_{i=1}^{p} p\left(\mathbf{y}_i \mid \mathbf{f}_*, \mathbf{X}_i\right).$$

- The above approximation leads to

$$q\left(\mathbf{f}_* \mid \mathcal{D}_1, \ldots, \mathcal{D}_p\right) \propto p\left(\mathbf{f}_*\right) \prod_{i=1}^{p} p\left(\mathbf{y}_i \mid \mathbf{f}_*, X_i\right) = c \frac{\prod_{i=1}^{p} p\left(\mathbf{f}_* \mid \mathcal{D}_i\right)}{p^{p-1}\left(\mathbf{f}_*\right)},$$

where we have used $p\left(\mathbf{y}_i \mid \mathbf{f}_*, \mathbf{X}_i\right) \propto p(\mathbf{f}_* \mid \mathcal{D}_i)/p\left(\mathbf{f}_*\right)$ and c is a constant.

Predictive distribution for the BCM (I)

- The numerator and denominator in the expression before only involve Gaussian distributions.

- We can use the technique of *completing the square* to compute the mean and covariance for $q\left(\mathbf{f}_* \mid \mathcal{D}_1, \ldots, \mathcal{D}_p\right) = q\left(\mathbf{f}_* \mid \mathcal{D}\right)$.

Predictive distribution for the BCM (II)

- It can be shown that the predictive mean and predictive covariance for $q(\mathbf{f}_* \mid \mathcal{D})$ are given as

$$\mathbb{E}_q[\mathbf{f}_* \mid \mathcal{D}] = [\mathrm{cov}_q(\mathbf{f}_* \mid \mathcal{D})] \sum_{i=1}^{p} [\mathrm{cov}(\mathbf{f}_* \mid \mathcal{D}_i)]^{-1} \mathbb{E}[\mathbf{f}_* \mid \mathcal{D}_i]$$

$$[\mathrm{cov}_q(\mathbf{f}_* \mid \mathcal{D})]^{-1} = -(p-1)\mathbf{K}_{**}^{-1} + \sum_{i=1}^{p} [\mathrm{cov}(\mathbf{f}_* \mid \mathcal{D}_i)]^{-1},$$

  where $\mathbf{K}_{**}$ corresponds to the covariance matrix at the test points.

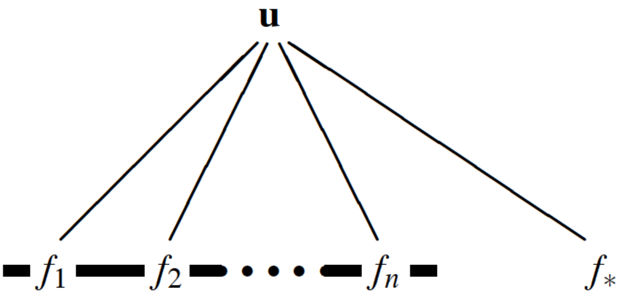
- $\mathbb{E}[\mathbf{f}_* \mid \mathcal{D}_i]$ and $\mathrm{cov}(\mathbf{f}_* \mid \mathcal{D}_i)$ are computed using the expressions for the preditictive distribution in GP regression.

Contents

Inducing variable methods

- There is a set of methods that explicitly introduce additional variables into the GP prior.
- Such additional variables $\mathbf{u} = [u_1, \ldots, u_m]^\top$ are known as *inducing variables*.
- These latent variables are values of the GP evaluated at a set of inputs $\mathbf{Z}$.
- The main idea of these methods is to exploit *conditional independencies* between $\mathbf{f}$ and $\mathbf{f}_*$ based on $\mathbf{u}$.
- The role of $\mathbf{u}$ is to *induce* dependencies between training and test cases.

Pictorial representation



A comment on notation

- Depending on context, $\mathbf{K}_{\mathbf{u},\mathbf{u}}$ might also be called $\mathbf{K}_{mm}$

Approximate the likelihood or the prior

- We can introduce these methods by using the inducing variables to exploit conditional dependencies in the likelihood or in the prior.

- In what follows, we will use the unifying view by Quiñonero-Candela and Rasmussen (2005) where the conditional dependencies are in the prior.

- This view assumes the exact likelihood function for GP regression $p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma^2_{\text{noise}}\mathbf{I}\right)$.

Exact prior

- The exact GP prior $p(\mathbf{f}_*, \mathbf{f})$ can be recovered by marginalising $\mathbf{u}$ from $p(\mathbf{f}_*, \mathbf{f}, \mathbf{u})$,

$$p(\mathbf{f}_*, \mathbf{f}) = \int p(\mathbf{f}_*, \mathbf{f}, \mathbf{u})\, \mathrm{d}\mathbf{u} = \int p(\mathbf{f}_*, \mathbf{f} \mid \mathbf{u})\, p(\mathbf{u})\mathrm{d}\mathbf{u}$$

  where $p(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}})$.

- The exact prior conditionals $p(\mathbf{f} \mid \mathbf{u})$ and $p(\mathbf{f}_* \mid \mathbf{u})$ are given as

$$p(\mathbf{f} \mid \mathbf{u}) = \mathcal{N}\left(\mathbf{f} \mid \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}}\right)$$
$$p(\mathbf{f}_* \mid \mathbf{u}) = \mathcal{N}\left(\mathbf{f}_* \mid \mathbf{K}_{*,\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \mathbf{K}_{*,*} - \mathbf{Q}_{*,*}\right),$$

  with $\mathbf{Q}_{\mathbf{a},\mathbf{b}} \triangleq \mathbf{K}_{\mathbf{a},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{b}}$.

Approximate prior

- The joint prior $p(\mathbf{f}_*, \mathbf{f})$ can be approximated by $q(\mathbf{f}_*, \mathbf{f})$ using

$$p\left(\mathbf{f}_*, \mathbf{f}\right) \simeq q\left(\mathbf{f}_*, \mathbf{f}\right) = \int q\left(\mathbf{f}_* \mid \mathbf{u}\right) q(\mathbf{f} \mid \mathbf{u}) p(\mathbf{u}) \mathrm{d}\mathbf{u},$$

  where $p(\mathbf{u}) = \mathcal{N}\left(\mathbf{u} \mid \mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}}\right)$.

- Depending on how we approximate the conditional priors $q\left(\mathbf{f}_* \mid \mathbf{u}\right)$ and $q(\mathbf{f} \mid \mathbf{u})$, we get different type of approximations.

Contents

DTC: priors

- DTC has been introduced as a likelihood approximation method based on inducing inputs, under the names of Projected Latent Variables (PLV) or Projected Process Approximation (PPA).

- In the DTC approximation, the training conditional distribution is deterministic and the test conditional distribution is exact, this is

$$
\begin{aligned}
q_{\mathrm{DTC}}(\mathbf{f} \mid \mathbf{u}) &= \mathcal{N}\left(\mathbf{f} \mid \mathbf{K}_{\mathbf{f},\mathbf{u}} \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u}, \mathbf{0}\right) \\
q_{\mathrm{DTC}}\left(\mathbf{f}_* \mid \mathbf{u}\right) &= p\left(\mathbf{f}_* \mid \mathbf{u}\right) \\
&= \mathcal{N}\left(\mathbf{f}_* \mid \mathbf{K}_{*,\mathbf{u}} \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u}, \mathbf{K}_{*,*} - \mathbf{Q}_{*,*}\right).
\end{aligned}
$$

- The joint prior implied by DTC follows as

$$
q_{\mathrm{DTC}}\left(\mathbf{f}, \mathbf{f}_*\right) = \mathcal{N}\left(\left[\begin{array}{c} \mathbf{f} \\ \mathbf{f}_* \end{array}\right] \middle| \mathbf{0}, \left[\begin{array}{cc} \mathbf{Q}_{\mathbf{f},\mathbf{f}} & \mathbf{Q}_{\mathbf{f},*} \\ \mathbf{Q}_{*,\mathbf{f}} & \mathbf{K}_{*,*} \end{array}\right]\right)
$$

DTC: predictive distribution

- Using the Gaussian likelihood model, $p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma^2_{\text{noise}}\mathbf{I}\right)$, the predictive distribution follows as

$$q_{\text{DTC}}\left(\mathbf{f}_* \mid \mathbf{y}\right) = \mathcal{N}\left(\mathbf{f}_* \mid \boldsymbol{\mu}_{*,\text{DTC}}, \boldsymbol{\Sigma}_{*,\text{DTC}}\right),$$

  where

$$\boldsymbol{\mu}_{*,\text{DTC}} = \mathbf{Q}_{*,\mathbf{f}}\left(\mathbf{Q}_{\mathbf{f},\mathbf{f}} + \sigma^2_{\text{noise}}\mathbf{I}\right)^{-1}\mathbf{y} = \sigma^{-2}\mathbf{K}_{*,\mathbf{u}}\Sigma\mathbf{K}_{\mathbf{u},\mathbf{f}}\mathbf{y},$$

$$\boldsymbol{\Sigma}_{*,\text{DTC}} = \mathbf{K}_{*,*} - \mathbf{Q}_{*,\mathbf{f}}\left(\mathbf{Q}_{\mathbf{f},\mathbf{f}} + \sigma^2_{\text{noise}}\mathbf{I}\right)^{-1}\mathbf{Q}_{\mathbf{f},*} = \mathbf{K}_{*,*} - \mathbf{Q}_{*,*} + \mathbf{K}_{*,\mathbf{u}}\Sigma\mathbf{K}^\top_{*,\mathbf{u}},$$

  with $\boldsymbol{\Sigma} = \left(\sigma^{-2}_{\text{noise}}\mathbf{K}_{\mathbf{u},\mathbf{f}}\mathbf{K}_{\mathbf{f},\mathbf{u}} + \mathbf{K}_{\mathbf{u},\mathbf{u}}\right)^{-1}$.

- Comparing against the subset of regressors, the predictive means are the same in both methods. The covariance for the SoR was equal to $\mathbf{K}_{*,\mathbf{u}}\Sigma\mathbf{K}^\top_{*,\mathbf{u}}$.

- Because $\mathbf{K}_{*,*} - \mathbf{Q}_{*,*}$ is always positive, the predictive variance in DTC is larger than the SoR's predictive variance.;

Contents

FITC: priors

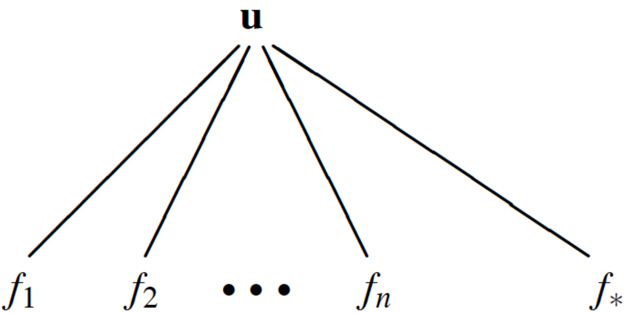
- FITC has been introduced as a likelihood approximation method based on inducing inputs, under the name Sparse Gaussian Processes using Pseudo-inputs (SGPP) by Snelson and Ghahramani (2005).

- In the FITC approximation, the training conditional distribution includes a variance term and the test conditional distribution is exact, this is

$$q_{\mathrm{FITC}}(\mathbf{f} \mid \mathbf{u}) = \prod_{i=1}^{n} p\left(f_i \mid \mathbf{u}\right) = \mathcal{N}\left(\mathbf{f} \mid \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \operatorname{diag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}}\right]\right)$$

$$q_{\mathrm{FITC}}\left(\mathbf{f}_* \mid \mathbf{u}\right) = p\left(\mathbf{f}_* \mid \mathbf{u}\right) = \mathcal{N}\left(\mathbf{f}_* \mid \mathbf{K}_{*,\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \mathbf{K}_{*,*} - \mathbf{Q}_{*,*}\right).$$

Pictorial representation



FITC vs DTC

- The joint prior implied by FITC follows as

$$q_{\text{FITC}}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \middle| \mathbf{0}, \begin{bmatrix} \mathbf{Q}_{\mathbf{f},\mathbf{f}} - \mathrm{diag}[\mathbf{Q}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{f}}] & \mathbf{Q}_{\mathbf{f},*} \\ \mathbf{Q}_{*,\mathbf{f}} & \mathbf{K}_{*,*} \end{bmatrix}\right)$$

- From before, the joint prior implied by DTC was

$$q_{\text{DTC}}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \middle| \mathbf{0}, \begin{bmatrix} \mathbf{Q}_{\mathbf{f},\mathbf{f}} & \mathbf{Q}_{\mathbf{f},*} \\ \mathbf{Q}_{*,\mathbf{f}} & \mathbf{K}_{*,*} \end{bmatrix}\right)$$

- Compared to DTC, FITC uses the exact covariance function in the main diagonal.

FITC: predictive distribution

- Using the Gaussian likelihood model, $p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma^2_{\text{noise}}\mathbf{I}\right)$, the predictive distribution follows as

$$q_{\text{FITC}}\left(\mathbf{f}_* \mid \mathbf{y}\right) = \mathcal{N}\left(\mathbf{f}_* \mid \boldsymbol{\mu}_{*,\text{FITC}}, \boldsymbol{\Sigma}_{*,\text{FITC}}\right),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{*,\text{FITC}} &= \mathbf{Q}_{*,\mathbf{f}}\left(\mathbf{Q}_{\mathbf{f},\mathbf{f}} + \Lambda\right)^{-1}\mathbf{y} = \mathbf{K}_{*,\mathbf{u}}\Sigma\mathbf{K}_{\mathbf{u},\mathbf{f}}\Lambda^{-1}\mathbf{y}, \\
\boldsymbol{\Sigma}_{*,\text{FITC}} &= \mathbf{K}_{*,*} - \mathbf{Q}_{*,\mathbf{f}}\left(\mathbf{Q}_{\mathbf{f},\mathbf{f}} + \Lambda\right)^{-1}\mathbf{Q}_{\mathbf{f},*} = \mathbf{K}_{*,*} - \mathbf{Q}_{*,*} + \mathbf{K}_{*,\mathbf{u}}\Sigma\mathbf{K}_{*,\mathbf{u}}^{\top},
\end{aligned}$$

with $\boldsymbol{\Sigma} = \left(\mathbf{K}_{\mathbf{u},\mathbf{f}}\boldsymbol{\Lambda}^{-1}\mathbf{K}_{\mathbf{f},\mathbf{u}} + \mathbf{K}_{\mathbf{u},\mathbf{u}}\right)^{-1}$ and $\boldsymbol{\Lambda} = \text{diag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}} + \sigma^2_{\text{noise}}\,\mathbf{I}\right]$.

Contents

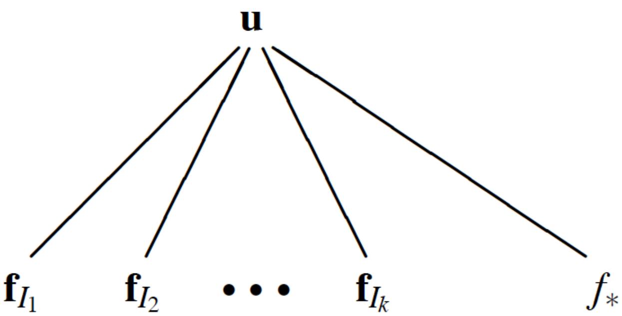
PITC: priors

- In the PITC approximation, the training conditional distribution has a block-diagonal covariance and the test conditional distribution is exact, this is

$$
\begin{aligned}
q_{\mathrm{PITC}}(\mathbf{f} \mid \mathbf{u}) &= \mathcal{N}\left(\mathbf{f} \mid \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \mathrm{blockdiag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}}\right]\right) \\
q_{\mathrm{PITC}}\left(\mathbf{f}_* \mid \mathbf{u}\right) &= p\left(\mathbf{f}_* \mid \mathbf{u}\right) = \mathcal{N}\left(\mathbf{f}_* \mid \mathbf{K}_{*,\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \mathbf{K}_{*,*} - \mathbf{Q}_{*,*}\right),
\end{aligned}
$$

where $\mathrm{blockdiag}[A]$ is a block diagonal matrix. The blocks have not been specified.

Pictorial representation



PITC: prior

- The joint prior implied by PITC follows as

$$q_{\mathrm{PITC}}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \middle| \mathbf{0}, \begin{bmatrix} \mathbf{Q}_{\mathbf{f},\mathbf{f}} - \mathrm{blockdiag}[\mathbf{Q}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{f}}] & \mathbf{Q}_{\mathbf{f},*} \\ \mathbf{Q}_{*,\mathbf{f}} & \mathbf{K}_{*,*} \end{bmatrix} \right)$$

PITC: predictive distribution

- Using the Gaussian likelihood model, $p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma^2_{\text{noise}}\mathbf{I}\right)$, the predictive distribution follows as

$$q_{\text{PITC}}\left(\mathbf{f}_* \mid \mathbf{y}\right) = \mathcal{N}\left(\mathbf{f}_* \mid \boldsymbol{\mu}_{*,\text{PITC}}, \boldsymbol{\Sigma}_{*,\text{PITC}}\right),$$

where

$$\boldsymbol{\mu}_{*,\text{PITC}} = \mathbf{Q}_{*,\mathbf{f}}\left(\mathbf{Q}_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Lambda}\right)^{-1}\mathbf{y} = \mathbf{K}_{*,\mathbf{u}}\boldsymbol{\Sigma}\mathbf{K}_{\mathbf{u},\mathbf{f}}\boldsymbol{\Lambda}^{-1}\mathbf{y},$$
$$\boldsymbol{\Sigma}_{*,\text{PITC}} = \mathbf{K}_{*,*} - \mathbf{Q}_{*,\mathbf{f}}\left(\mathbf{Q}_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Lambda}\right)^{-1}\mathbf{Q}_{\mathbf{f},*} = \mathbf{K}_{*,*} - \mathbf{Q}_{*,*} + \mathbf{K}_{*,\mathbf{u}}\boldsymbol{\Sigma}\mathbf{K}^{\top}_{*,\mathbf{u}},$$

with $\boldsymbol{\Sigma} = \left(\mathbf{K}_{\mathbf{u},\mathbf{f}}\boldsymbol{\Lambda}^{-1}\mathbf{K}_{\mathbf{f},\mathbf{u}} + \mathbf{K}_{\mathbf{u},\mathbf{u}}\right)^{-1}$ and
$\boldsymbol{\Lambda} = \text{blockdiag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}} + \sigma^2_{\text{noise}}\mathbf{I}\right]$.

Contents

Marginal likelihood of the full Gaussian process.

- The marginal likelihood of the model is given by

$$p(\mathbf{y}|\mathbf{X},\phi) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Sigma})$$

  where $\mathbf{y} = \left[\mathbf{y}_1^\top, \ldots, \mathbf{y}_D^\top\right]^\top$ is the set of output functions, $\mathbf{K}_{\mathbf{f},\mathbf{f}}$ covariance matrix with blocks $\mathrm{cov}\left[f_d, f_{d'}\right]$, $\boldsymbol{\Sigma}$ matrix of noise variances, ϕ is the set of parameters of the covariance matrix and $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is the set of input vectors.

- Learning from the log-likelihood involves the inverse of $\mathbf{K}_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Sigma}$, which grows with complexity $\mathcal{O}(N^3 D^3)$

Predictive distribution of the full Gaussian process.

- Predictive distribution at $\mathbf{X}_*$

$$p(\mathbf{y}_*|\mathbf{y}, \mathbf{X}, \mathbf{X}_*, \phi) = \mathcal{N}\left(\boldsymbol{\mu}_*, \boldsymbol{\Lambda}_*\right)$$

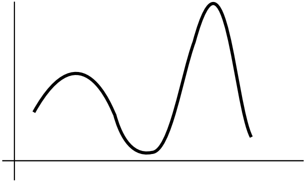
with

$$\begin{aligned}\boldsymbol{\mu}_* &= \mathbf{K}_{\mathbf{f}_*,\mathbf{f}}(\mathbf{K}_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Sigma})^{-1}\mathbf{y} \\ \boldsymbol{\Lambda}_* &= \mathbf{K}_{\mathbf{f}_*,\mathbf{f}_*} - \mathbf{K}_{\mathbf{f}_*,\mathbf{f}}(\mathbf{K}_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Sigma})^{-1}\mathbf{K}_{\mathbf{f},\mathbf{f}_*} + \boldsymbol{\Sigma}\end{aligned}$$

- Prediction is $\mathcal{O}(ND)$ for the mean and $\mathcal{O}(N^2D^2)$ for the variance.

Conditional prior distribution.

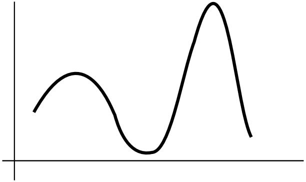
Sample from $p(u)$



$$f_d(\mathbf{x}) = \int_{\mathcal{X}} G_d(\mathbf{x} - \mathbf{z}) u(\mathbf{z}) \mathrm{d}\mathbf{z}$$

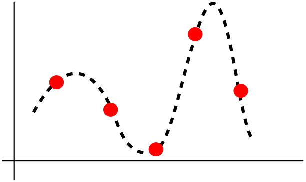
Conditional prior distribution.

Sample from $p(u)$



$$f_d(\mathbf{x}) = \int_{\mathcal{X}} G_d(\mathbf{x} - \mathbf{z}) u(\mathbf{z}) \mathrm{d}\mathbf{z}$$

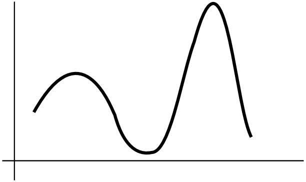
Discretize u



$$f_d(\mathbf{x}) \approx \sum_{\forall k} G_d(\mathbf{x} - \mathbf{z}_k) u(\mathbf{z}_k)$$

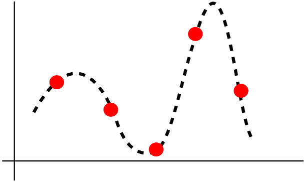
Conditional prior distribution.

Sample from $p(u)$



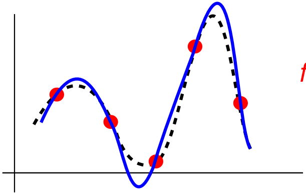
$$f_d(\mathbf{x}) = \int_{\mathcal{X}} G_d(\mathbf{x}-\mathbf{z})u(\mathbf{z})\mathrm{d}\mathbf{z}$$

Discretize u



$$f_d(\mathbf{x}) \approx \sum_{\forall k} G_d(\mathbf{x}-\mathbf{z}_k)u(\mathbf{z}_k)$$

Sample from $p(u|\mathbf{u})$



$$f_d(\mathbf{x}) \approx \int_{\mathcal{X}} G_d(\mathbf{x}-\mathbf{z})\,\mathrm{E}\left[u(\mathbf{z})|\mathbf{u}\right]\mathrm{d}\mathbf{z}$$

The conditional independence assumption (I)

- This form for $f_d(\mathbf{x})$ leads to the following likelihood

$$p(\mathbf{f}|\mathbf{u},\mathbf{Z}) = \mathcal{N}\left(\mathbf{f}|\mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right),$$

  where

$\mathbf{u}$	discrete sample from the latent function
$\mathbf{Z}$	set of input vectors corresponding to $\mathbf{u}$
$\mathbf{K}_{\mathbf{u},\mathbf{u}}$	cross-covariance matrix between latent functions
$\mathbf{K}_{\mathbf{f},\mathbf{u}} = \mathbf{K}_{\mathbf{u},\mathbf{f}}^{\top}$	cross-covariance matrix between latent and output functions

- Even though we conditioned on $\mathbf{u}$, we still have dependencies between outputs due to the uncertainty in $p(u|\mathbf{u})$.

The conditional independence assumption (II)

Our key assumption is that the outputs will be independent even if we have only observed **u** rather than the whole function u.

$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} - \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1}$	$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_2} - \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2}$	$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_3} - \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3}$
$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_1} - \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1}$	$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} - \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2}$	$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_3} - \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3}$
$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_1} - \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1}$	$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_2} - \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2}$	$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} - \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3}$

The conditional independence assumption (II)

Our key assumption is that the outputs will be independent even if we have only observed **u** rather than the whole function u.

$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} - \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1}$	$\mathbf{0}$	$\mathbf{0}$
$\mathbf{0}$	$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} - \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2}$	$\mathbf{0}$
$\mathbf{0}$	$\mathbf{0}$	$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} - \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3}$

Better approximations can be obtained when $\mathrm{E}[u|\mathbf{u}]$ approximates u.

Comparison of marginal likelihoods

Integrating out **u**, the marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z}, \mathbf{X}, \theta) = \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}} + \text{blockdiag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right] + \Sigma\right).$$

Comparison of marginal likelihoods

Integrating out $\mathbf{u}$, the marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z},\mathbf{X},\theta) = \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}} + \text{blockdiag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right] + \Sigma\right).$$

$$\begin{bmatrix} \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_2\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_3\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} \end{bmatrix} \approx \begin{bmatrix} \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} \end{bmatrix}$$

Comparison of marginal likelihoods

Integrating out $\mathbf{u}$, the marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z},\mathbf{X},\theta) = \mathcal{N}\left(\mathbf{y}|\mathbf{0},\mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}} + \operatorname{blockdiag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right] + \Sigma\right).$$

$$\begin{bmatrix} \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_2\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_3\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} \end{bmatrix} \approx \begin{bmatrix} \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_3} \\ \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_1} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_3} \\ \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_2} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_2\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_3\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} \end{bmatrix} \approx \boxed{\mathbf{G}} \;\mathbf{X}\; \boxed{\mathbf{G}^{\mathsf{T}}}$$

Discrete case $[\mathbf{G}]_{i,k} = G_d(\mathbf{x}_i - \mathbf{z}_k)$

Predictive distribution for the sparse approximation

Predictive distribution

$$
\begin{aligned}
p(\mathbf{y}_*|\mathbf{y}, \mathbf{X}, \mathbf{X}_*, \mathbf{Z}, \boldsymbol{\theta}) &= \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}_*, \widetilde{\boldsymbol{\Lambda}}_*\right), \text{ with} \\
\widetilde{\boldsymbol{\mu}}_* &= \mathbf{K}_{\mathbf{f}_*,\mathbf{u}}\mathbf{A}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}(\mathbf{D}+\boldsymbol{\Sigma})^{-1}\mathbf{y} \\
\widetilde{\boldsymbol{\Lambda}}_* &= \mathbf{D}_* + \mathbf{K}_{\mathbf{f}_*,\mathbf{u}}\mathbf{A}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}_*} + \boldsymbol{\Sigma} \\
\mathbf{A} &= \mathbf{K}_{\mathbf{u},\mathbf{u}} + \mathbf{K}_{\mathbf{u},\mathbf{f}}(\mathbf{D}+\boldsymbol{\Sigma})^{-1}\mathbf{K}_{\mathbf{f},\mathbf{u}} \\
\mathbf{D}_* &= \operatorname{blockdiag}\left[\mathbf{K}_{\mathbf{f}_*,\mathbf{f}_*} - \mathbf{K}_{\mathbf{f}_*,\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}_*}\right]
\end{aligned}
$$

Remarks

- For learning the computational demand is in the calculation of the block-diagonal term which grows as $\mathcal{O}(N^3 D) + \mathcal{O}(NDM^2)$ (with $Q = 1$). Storage is $\mathcal{O}(N^2 D) + \mathcal{O}(NDM)$.

- For inference, the computation of the mean grows as $\mathcal{O}(DM)$ and the computation of the variance as $\mathcal{O}(DM^2)$, after some pre-computations and for one test point.

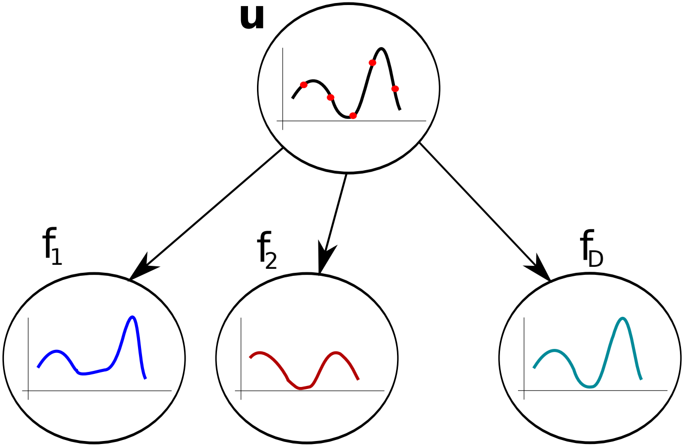
- The functional form of the approximation is almost identical to that of the Partially Independent Training Conditional (PITC) approximation Quiñonero-Candela and Rasmussen (2005).

Additional conditional independencies

- The N^3 term in the computational complexity and the N^2 term in storage in PITC are still expensive for larger data sets.
- An additional assumption is independence over the data points.

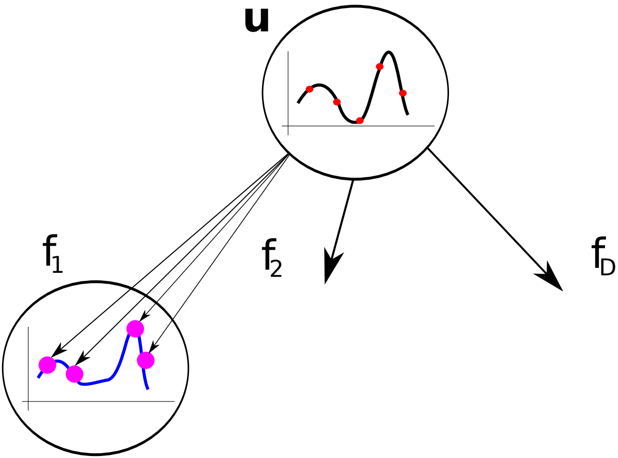
Additional conditional independencies

- The N^3 term in the computational complexity and the N^2 term in storage in PITC are still expensive for larger data sets.
- An additional assumption is independence over the data points.



Additional conditional independencies

- The N^3 term in the computational complexity and the N^2 term in storage in PITC are still expensive for larger data sets.
- An additional assumption is independence over the data points.



Comparison of marginal likelihoods

The marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z}, \mathbf{X}, \theta) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}} + \mathrm{diag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right] + \boldsymbol{\Sigma}\right).$$

Comparison of marginal likelihoods

The marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z},\mathbf{X},\theta) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}} + \operatorname{diag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right] + \Sigma\right).$$

$$\begin{bmatrix} \mathbf{K}_{f_1 f_1} & \mathbf{K}_{f_1 f_2} & \mathbf{K}_{f_1 f_3} \\ \mathbf{K}_{f_2 f_1} & \mathbf{K}_{f_2 f_2} & \mathbf{K}_{f_2 f_3} \\ \mathbf{K}_{f_3 f_1} & \mathbf{K}_{f_3 f_2} & \mathbf{K}_{f_3 f_3} \end{bmatrix} \approx \begin{bmatrix} \mathbf{K}_{f_1 f_1} & \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2} & \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3} \\ \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1} & \mathbf{K}_{f_2 f_2} & \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3} \\ \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1} & \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2} & \mathbf{K}_{f_3 f_3} \end{bmatrix}$$

Comparison of marginal likelihoods

The marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z}, \mathbf{X}, \theta) = \mathcal{N}\left(\mathbf{0}, \mathbf{K_{f,u}}\mathbf{K_{u,u}^{-1}}\mathbf{K_{u,f}} + \mathrm{diag}\left[\mathbf{K_{f,f}} - \mathbf{K_{f,u}}\mathbf{K_{u,u}^{-1}}\mathbf{K_{u,f}}\right] + \Sigma\right).$$

$$\begin{array}{|c|c|c|} \hline \mathbf{K_{f_1 f_1}} & \mathbf{K_{f_1 f_2}} & \mathbf{K_{f_1 f_3}} \\ \hline \mathbf{K_{f_2 f_1}} & \mathbf{K_{f_2 f_2}} & \mathbf{K_{f_2 f_3}} \\ \hline \mathbf{K_{f_3 f_1}} & \mathbf{K_{f_3 f_2}} & \mathbf{K_{f_3 f_3}} \\ \hline \end{array} \approx \begin{array}{|c|c|c|} \hline \mathbf{K_{f_1 f_1}} & \mathbf{K_{f_1 u}K_{uu}^{-1}K_{uf_2}} & \mathbf{K_{f_1 u}K_{uu}^{-1}K_{uf_3}} \\ \hline \mathbf{K_{f_2 u}K_{uu}^{-1}K_{uf_1}} & \mathbf{K_{f_2 f_2}} & \mathbf{K_{f_2 u}K_{uu}^{-1}K_{uf_3}} \\ \hline \mathbf{K_{f_3 u}K_{uu}^{-1}K_{uf_1}} & \mathbf{K_{f_3 u}K_{uu}^{-1}K_{uf_2}} & \mathbf{K_{f_3 f_3}} \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|} \hline \mathbf{K_{f_1 f_1}} & \mathbf{K_{f_1 f_2}} & \mathbf{K_{f_1 f_3}} \\ \hline \mathbf{K_{f_2 f_1}} & \mathbf{K_{f_2 f_2}} & \mathbf{K_{f_2 f_3}} \\ \hline \mathbf{K_{f_3 f_1}} & \mathbf{K_{f_3 f_2}} & \mathbf{K_{f_3 f_3}} \\ \hline \end{array} \approx \begin{array}{|c|c|c|} \hline \mathbf{Q_{f_1 f_1}} & \mathbf{K_{f_1 u}K_{uu}^{-1}K_{uf_2}} & \mathbf{K_{f_1 u}K_{uu}^{-1}K_{uf_3}} \\ \hline \mathbf{K_{f_2 u}K_{uu}^{-1}K_{uf_1}} & \mathbf{Q_{f_2 f_2}} & \mathbf{K_{f_2 u}K_{uu}^{-1}K_{uf_3}} \\ \hline \mathbf{K_{f_3 u}K_{uu}^{-1}K_{uf_1}} & \mathbf{K_{f_3 u}K_{uu}^{-1}K_{uf_2}} & \mathbf{Q_{f_3 f_3}} \\ \hline \end{array}$$

Comparison of marginal likelihoods

The marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z},\mathbf{X},\theta) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}} + \operatorname{diag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right] + \Sigma\right).$$

$$\begin{bmatrix} \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_2\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_3\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} \end{bmatrix} \approx \begin{bmatrix} \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_2\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_3\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} \end{bmatrix} \approx \begin{bmatrix} \mathbf{Q}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1} & \mathbf{Q}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3} \\ \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2} & \mathbf{Q}_{\mathbf{f}_3\mathbf{f}_3} \end{bmatrix}$$

Comparison of marginal likelihoods

The marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z},\mathbf{X},\theta) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}} + \operatorname{diag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right] + \Sigma\right).$$

$$\begin{array}{|c|c|c|}\hline
\mathbf{K}_{\mathbf{f}_1\mathbf{f}_1}(\mathbf{x}_1,\mathbf{x}_1) & (\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1})(\mathbf{x}_1,\mathbf{x}_2) & (\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1})(\mathbf{x}_1,\mathbf{x}_3) \\ \hline
(\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1})(\mathbf{x}_2,\mathbf{x}_1) & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1}(\mathbf{x}_2,\mathbf{x}_2) & (\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1})(\mathbf{x}_2,\mathbf{x}_3) \\ \hline
(\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1})(\mathbf{x}_3,\mathbf{x}_1) & (\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1})(\mathbf{x}_3,\mathbf{x}_2) & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_1}(\mathbf{x}_3,\mathbf{x}_3) \\ \hline
\end{array}$$

$$\begin{array}{|c|c|c|}\hline
\mathbf{K}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{f}_3} \\ \hline
\mathbf{K}_{\mathbf{f}_2\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{f}_3} \\ \hline
\mathbf{K}_{\mathbf{f}_3\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_3\mathbf{f}_3} \\ \hline
\end{array}
\approx
\begin{array}{|c|c|c|}\hline
\mathbf{Q}_{\mathbf{f}_1\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3} \\ \hline
\mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1} & \mathbf{Q}_{\mathbf{f}_2\mathbf{f}_2} & \mathbf{K}_{\mathbf{f}_2\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_3} \\ \hline
\mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_1} & \mathbf{K}_{\mathbf{f}_3\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}\mathbf{f}_2} & \mathbf{Q}_{\mathbf{f}_3\mathbf{f}_3} \\ \hline
\end{array}$$

Comparison of marginal likelihoods

The marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z},\mathbf{X},\theta) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}} + \mathrm{diag}\left[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}\right] + \Sigma\right).$$

$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_1}(\mathbf{x}_1,\mathbf{x}_1)$	$(\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_1})(\mathbf{x}_1,\mathbf{x}_2)$	$(\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_1})(\mathbf{x}_1,\mathbf{x}_3)$
$(\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_1})(\mathbf{x}_2,\mathbf{x}_1)$	$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_1}(\mathbf{x}_2,\mathbf{x}_2)$	$(\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_1})(\mathbf{x}_2,\mathbf{x}_3)$
$(\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_1})(\mathbf{x}_3,\mathbf{x}_1)$	$(\mathbf{K}_{\mathbf{f}_1\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}_1})(\mathbf{x}_3,\mathbf{x}_2)$	$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_1}(\mathbf{x}_3,\mathbf{x}_3)$

$$\mathbf{Q}_{\mathbf{f}_1,\mathbf{f}_1}$$

Computational requirements

- The computational demand is now equal to $\mathcal{O}(NDM^2)$. Storage is $\mathcal{O}(NDM)$.
- For inference, the computation of the mean grows as $\mathcal{O}(DM)$ and the computation of the variance as $\mathcal{O}(DM^2)$, after some pre-computations and for one test point.
- Similar to the Fully Independent Training Conditional (FITC) approximation Quiñonero-Candela and Rasmussen (2005); Snelson and Ghahramani (2005).

Deterministic approximation

- We could also assume that given the latent functions the outputs are deterministic.
- The marginal likelihood is given as

$$p(\mathbf{y}|\mathbf{Z}, \mathbf{X}, \theta) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}} + \Sigma\right).$$

- Computation complexity is the same as FITC.
- Deterministic training conditional approximation (DTC).

References I

Neil Lawrence, Matthias Seeger, and Ralf Herbrich. Fast sparse gaussian process methods: The informative vector machine. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 625–632. MIT Press, 2003.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

Matthias Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, School of Informatics, University of Edinburgh, 2003.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *NIPS 2005*, 2005.