Understanding Probabilistic Sparse Gaussian Process Approximations

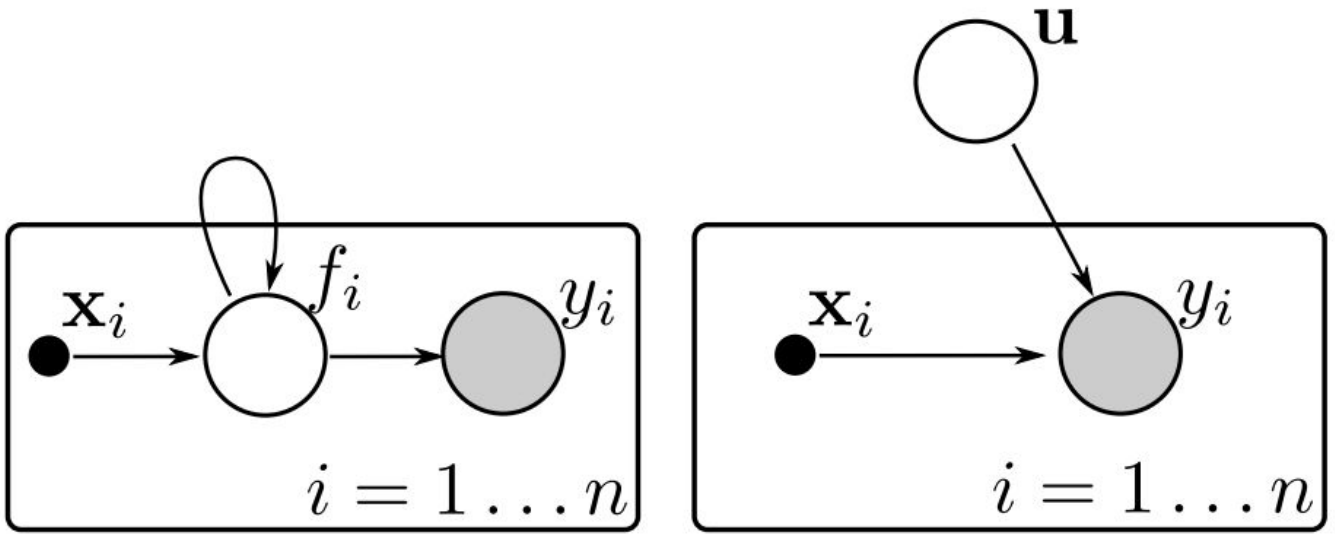
By Matthias Stephan Bauer
Presentation by Salvador Medina Herrera

Introduction: Gaussian Processes

A **Gaussian Process** is a statistical distribution, for which every finite linear combination of samples has a joint **Gaussian Distribution**:

N(**f; 0, Kff**) where [**Kff**] = k(**xi, xj**), where {**xi**} are the inputs

Gaussian Process Regression (I)



(a) Gaussian Process regression (b) Variational GP regression

$$p(\mathbf{f}) = N(\mathbf{f}; \mathbf{0}, \mathbf{K}_{\mathbf{ff}}), \quad p(\mathbf{y}|\mathbf{f}) = \prod_{i=1\ldots n} N(\mathbf{y}_i; \mathbf{f}_i, \boldsymbol{\sigma}_i^2)$$

Gaussian Process Regression (II)

Define the hyperparameter $\boldsymbol{\theta}$ that contains the signal and noise variance. We can determine it by optimizing the marginal likelihood and then marginalize over the posterior of **f**:

$$\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathbf{y}|\boldsymbol{\theta})$$

$$p(\mathbf{y}^*|\mathbf{y}) = p(\mathbf{y}^*,\mathbf{y})/p(\mathbf{y}) = \int p(\mathbf{y}^*|\mathbf{f}^*)\, p(\mathbf{f}^*|\mathbf{f})\, p(\mathbf{f}|\mathbf{y})\, d\mathbf{f}\, d\mathbf{f}^*$$

But the cost of evaluating it scales as $\mathbf{O(N^3)}$ due to the inversion of $\mathbf{K_{ff}} + \sigma_n^2\,\mathbf{I}$!!

SGP Approximations (I)

In order to reduce computational complexity, multiple approximations have been proposed for SGP estimation:

- Subset of Data
- Subset of Regressors (Deterministic Inducing Conditional)
- Deterministic Training Conditional
- Fully or Partially Independent Training Conditional

...

SGP Approximations (II)

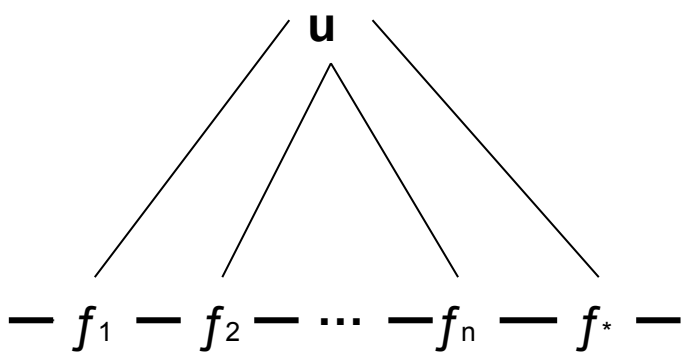
This paper compares **Fully Independent Training Conditional** (FITC) and **Variational Free Energy** (VFE).

They allow both the hyperparameters and inducing inputs to be learned from the date through gradient-based optimisation and rely on a lower-rank matrix $\mathbf{Q}_{ff} = \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}$.

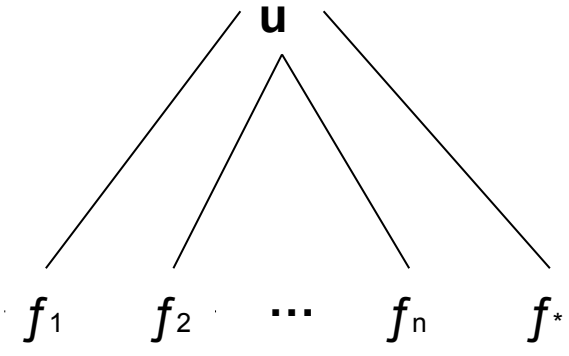
Hence reducing the size of any matrix inversion from **N** to **M**.

Fully Independent Training Conditional

The approximation is made by imposing a conditional independence assumption on the joint prior over the training and test cases (See annex for formulas*)



No approximation
(Fully Connected)



FITC approximation
(Latent function values not connected)

Figure: Relation between inducing variables, training latent functions and test function value.

Variational Free Energy

- We augment the model with the gaussian process u at the inducing inputs and then define a lower bound to the log-likelihood using Jensen's inequality.
- Define 'q' so that q(f,u) = q(f|u)*q(u) and q(f|u) = p(f|y). We suppose that p(f|y) is approximately equal to p(f|u). Substituting in the log-likelihood bound we have:

$$\log p(\mathbf{y}) \geq \int q(\mathbf{u}, \mathbf{f}) \log \frac{p(\mathbf{y}|\mathbf{f}) \cancel{p(\mathbf{f}|\mathbf{u})} p(\mathbf{u})}{\cancel{p(\mathbf{f}|\mathbf{u})} q(\mathbf{u})} \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f}$$

Using variational calculus we get the following lower bound:

$$\log p(\mathbf{y}) \geq \log \mathcal{N}\left(\mathbf{y}; 0, Q_{\mathbf{ff}} + \sigma_n^2 I\right) - \frac{1}{2\sigma_n^2} \operatorname{tr}\left(K_{\mathbf{ff}} - Q_{\mathbf{ff}}\right)$$

Objective function for VFE and FITC (I)

$$\mathcal{F} = \frac{N}{2}\log(2\pi) + \underbrace{\frac{1}{2}\log|Q_{\mathbf{ff}} + G|}_{\text{complexity penalty}} + \underbrace{\frac{1}{2}\mathbf{y}^{\intercal}(Q_{\mathbf{ff}} + G)^{-1}\mathbf{y}}_{\text{data fit}} + \underbrace{\frac{1}{2\sigma_n^2}\operatorname{tr}(T)}_{\text{trace term}}$$

where:

$$G_{\text{FITC}} = \operatorname{diag}[K_{\mathbf{ff}} - Q_{\mathbf{ff}}] + \sigma_n^2 I \qquad G_{\text{VFE}} = \sigma_n^2 I$$
$$T_{\text{FITC}} = 0 \qquad T_{\text{VFE}} = K_{\mathbf{ff}} - Q_{\mathbf{ff}}.$$

Objective function for VFE and FITC (II)

- **Data Fit:** Penalises the data lying outside the covariance ellipse $\mathbf{Q}_{ff}$ + $\mathbf{G}$.
- **Complexity Penalty**: Penalises the method for being able to predict too many datasets. Characterises the volume of possible datasets that are compatible with the data fit term.
- **Trace Term (In VFE):** Penalises the sum of conditional variances at the training inputs, conditioned on the inducing inputs. Ensures that VFE approximates the covariance structure of the full GP $\mathbf{K}_{ff}$.

Comparison

FITC can underestimate the noise variance whereas VFE overestimates it

- The diagonal correction term $\mathbf{diag(K_{ff} - Q_{ff})}$ of **FITC** is zero exactly at an inducing input and grows to the prior variance away from it. Consequently underestimating the variance near the inducing inputs.
- The data fit and trace terms $\mathbf{Q_{ff} + \sigma_n^2 I}$ have $\mathbf{N - M}$ eigenvectors with an eigenvalue of σ_n^2. Any component of y in these directions will result in a larger penalty than for $\mathbf{K_{ff}}$, which can only be reduced by increasing σ_n^2. It can overestimate σ_n^2 as a result.

FITC can underestimate the noise variance whereas VFE overestimates it

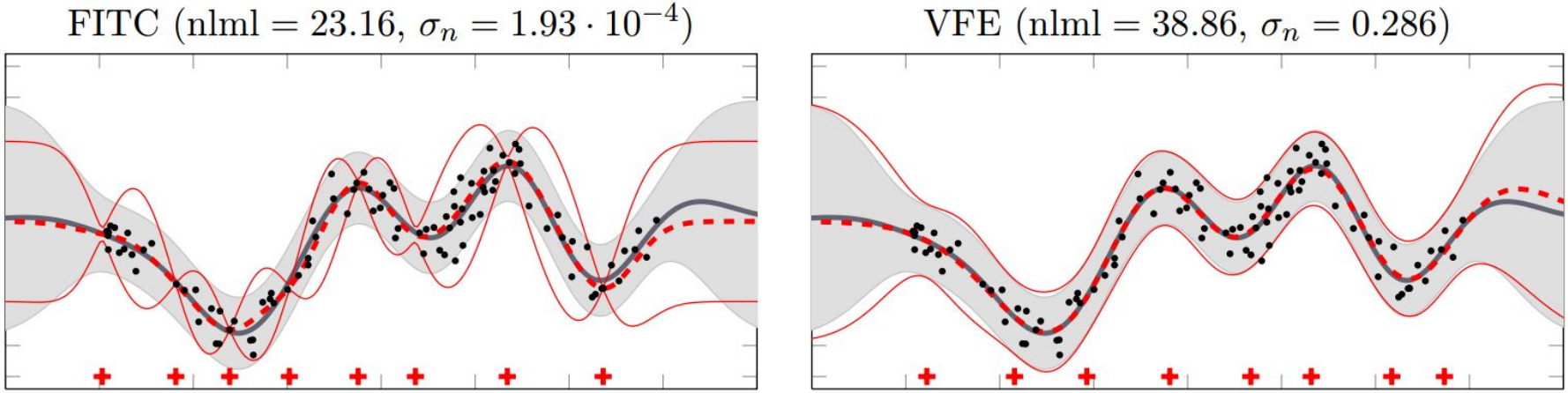


Figure 1: Behaviour of FITC and VFE on subset of 100 data points of the Snelson dataset for 8 inducing inputs (red crosses indicate inducing inputs; red lines indicate mean and 2σ) compared to the prediction of the full GP in grey. Optimised values for the full GP: nlml = 34.15, $\sigma_n = 0.274$

VFE improves with additional inducing inputs, FITC may ignore them

In **FITC**, adding additional inducing inputs reduces the complexity penalty since the diagonal component of $\mathbf{Q}_{ff} + \mathbf{G}$ is reduced and replaced by a more strongly correlated $\mathbf{Q}_{ff}$. **BUT** it worsens the data fit term as the heteroscedastic term is required to fit the data when the homoscedastic noise is underestimated.

VFE improves with additional inducing inputs, FITC may ignore them

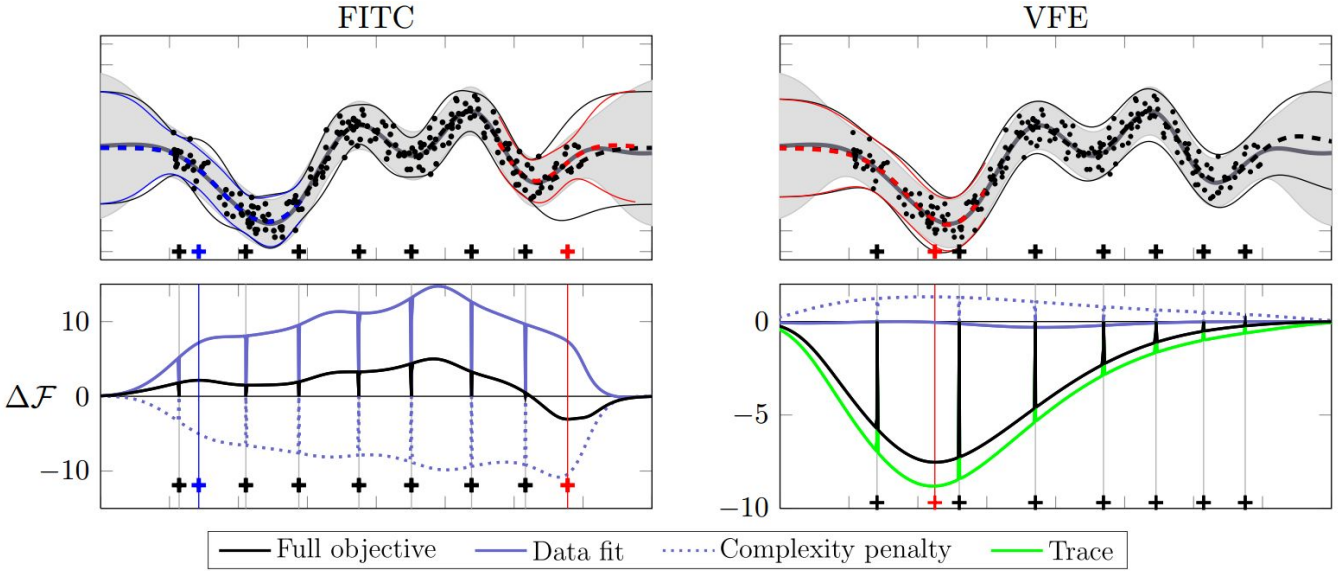


Figure 2: *Top:* Fits for FITC and VFE on 200 data points of the Snelson dataset for 7 optimised inducing inputs (black). *Bottom:* Change in objective function from adding an inducing input anywhere along the x-axis. The overall change is decomposed into the change in the individual terms (see legend). Two particular additional inducing inputs and their effect on the predictive distribution shown in red and blue.

Other properties

- FITC does not recover the true posterior, VFE does: Both VFE and FITC can recover the true posterior by placing an inducing input on every training input. For VFE, this must be a global minimum. For FITC, however, this solution is merely a saddle point.
- FITC relies on local optima: When running FITC with many inducing inputs its resemblance to the full GP solution relies on local optima, rather than the objective function changing.
- VFE is hindered by local optima: VFE has a tendency to find under-fitting solutions. However, this is an optimisation issue. The bound correctly identifies good solutions.

Conclusions (I)

The paper highlights some of the usual shortcomings of FITC:

- Overestimation of the marginal likelihood
- Underestimation of the noise variance
- Wasting of modelling resources
- Not able to recover the true posterior.

And VFE:

- Susceptible to local optima
- Harder to optimise

Conclusions (II)

VFE is usually superior to FITC:

- It is a true bound to the marginal likelihood of the full GP: It does not ignore connections between latent inputs
- Always improves with more resources
- Recovers the true posterior when possible

Hence the usage of VFE over FITC is recommended, specially if ‘optimization tricks’ are used to prevent it from falling local optima.

Showtime

Questions

Annex:
Gimme my Maths!

Fully Independent Training Conditional (*)

The approximation is made by imposing a conditional independence assumption on the joint prior over the training and test cases:

$$p(\mathbf{f}, \mathbf{f}_\star|\mathbf{X}, \mathbf{X}_\star, \bar{\mathbf{X}}) = \int p(\mathbf{f}, \mathbf{f}_\star|\mathbf{u}, \mathbf{X}, \mathbf{X}_\star)p(\mathbf{u}|\bar{\mathbf{X}})\mathrm{d}\mathbf{u} \approx \int q(\mathbf{f}|\mathbf{u}, \mathbf{X})q(\mathbf{f}_\star|\mathbf{u}, \bar{\mathbf{X}})p(\mathbf{u}|\bar{\mathbf{X}})\mathrm{d}\mathbf{u}$$

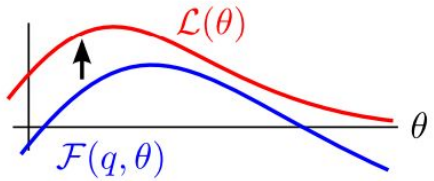
$$p(\mathbf{u}|\bar{\mathbf{X}}) = \mathcal{N}(\mathbf{u}\,;\,\mathbf{0}\,,\,\mathbf{K_{uu}})$$
$$q(\mathbf{f}|\mathbf{u}, \mathbf{X}) = \mathcal{N}\big(\mathbf{f}\,;\,\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}\,,\,\mathrm{diag}\,(\mathbf{K_{ff}} - \mathbf{Q_{ff}})\big)$$
$$q(\mathbf{f}_\star|\mathbf{u}, \mathbf{X}_\star) = \mathcal{N}\big(\mathbf{f}_\star\,;\,\mathbf{K_{\star u}}\mathbf{K_{uu}^{-1}}\mathbf{u}\,,\,\mathrm{diag}\,(\mathbf{K_{\star\star}} - \mathbf{Q_{\star\star}})\big)$$
$$\mathbf{Q_{ab}} = \mathbf{K_{au}}\mathbf{K_{uu}^{-1}}\mathbf{K_{ub}}$$

$$\begin{aligned} q(\mathbf{f}|\mathbf{X}) &= \int \mathcal{N}\big(\mathbf{f}\,;\,\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}\,,\,\mathrm{diag}\,(\mathbf{K_{ff}} - \mathbf{Q_{ff}})\big)\,\mathcal{N}(\mathbf{u}\,;\,\mathbf{0}\,,\,\mathbf{K_{uu}})\,\mathrm{d}\mathbf{u} \\ &= \mathcal{N}(\mathbf{f}\,;\,\mathbf{0}\,,\,\mathbf{Q_{ff}} + \mathrm{diag}\,(\mathbf{K_{ff}} - \mathbf{Q_{ff}})) \end{aligned}$$

Variational Free Energy (*)

augment the model with pseudo-data: $p(\mathbf{y}, \mathbf{f}, \mathbf{u}|\theta) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}, \mathbf{u})$

lower bound the likelihood



$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathbf{df}\, \mathbf{du}\; p(\mathbf{y}, \mathbf{f}, \mathbf{u})$$

$$= \log \int \mathbf{df}\, \mathbf{du}\; p(\mathbf{y}, \mathbf{f}, \mathbf{u}) \frac{q(\mathbf{f}, \mathbf{u})}{q(\mathbf{f}, \mathbf{u})} \geq \int \mathbf{df}\, \mathbf{du}\; q(\mathbf{f}, \mathbf{u}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{u})}{q(\mathbf{f}, \mathbf{u})} = \mathcal{F}(q, \theta)$$

assume approximate posterior factorisation with special form

$$q(\mathbf{f}, \mathbf{u}) = q(\mathbf{f}|\mathbf{u})q(\mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \qquad \text{(exact } q(\mathbf{f}|\mathbf{u}) = p(\mathbf{f}|\mathbf{y})\text{)}$$

$$\mathcal{F}(q, \theta) = \int \mathbf{df}\, \mathbf{du}\; q(\mathbf{f}, \mathbf{u}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{u})}{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} = \int \mathbf{df}\, \mathbf{du}\; q(\mathbf{f}, \mathbf{u}) \log \frac{p(\mathbf{y}|\mathbf{f})\cancel{p(\mathbf{f}|\mathbf{u})}p(\mathbf{u})}{\cancel{p(\mathbf{f}|\mathbf{u})}q(\mathbf{u})}$$

make bound as tight as possible: $q^*(\mathbf{u}) = \arg\max_{q(\mathbf{u})} \mathcal{F}(q, \theta)$

Thanks