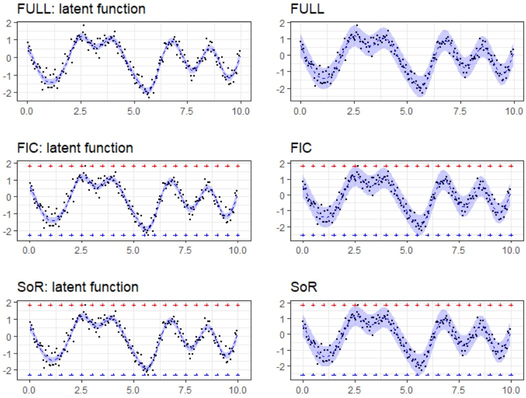
Inducing Point GP Approximations

October 31, 2019

Motivating Example



GP Notation

$$\phi \sim \mathcal{GP}(m(\cdot), k_\theta(\cdot, \cdot))$$

- y: vector of observed response values
- x_i: vector of explanatory variables for response value y_i (x is the matrix with x_i as rows)
- $\tilde{x}$: matrix of inducing point locations with rows $\tilde{x}_i$
- $f_i = f(x_i)$: realized values of the latent function at N input locations $x_i \in \mathcal{X} \subset \mathbb{R}^d$
- f_x: vector of (realized) N function values at observed data locations
- $\phi_x \sim \mathcal{N}\left(m_x, \Sigma_{xx}\right)$

Subset of Regressors (SoR) Approximation

- $p_{SoR}(\phi_x, \phi_{\tilde{x}}) = p(\phi_x, \phi_{\tilde{x}}) = p(\phi_x | \phi_{\tilde{x}}) p(\phi_{\tilde{x}})$
  - $\phi_x | \phi_{\tilde{x}} \sim N\left(m_x + \Sigma_{x\tilde{x}} \Sigma_{\tilde{x}\tilde{x}}^{-1} (\phi_{\tilde{x}} - m_{\tilde{x}}), 0\right)$
  - $\phi_{\tilde{x}} \sim \mathcal{N}(m_{\tilde{x}}, \Sigma_{\tilde{x}\tilde{x}})$

Fully Independent Conditional (FIC) Approximation

- $p_{FIC}(\phi_x, \phi_{\tilde{x}}) = p(\phi_x, \phi_{\tilde{x}}) = p(\phi_x|\phi_{\tilde{x}})p(\phi_{\tilde{x}})$
  - $\phi_x|\phi_{\tilde{x}} \sim N\left(m_x + \Sigma_{x\tilde{x}}\Sigma_{\tilde{x}\tilde{x}}^{-1}(\phi_{\tilde{x}} - m_{\tilde{x}}), \mathsf{diag}\left[\Sigma_{xx} - \Sigma_{x\tilde{x}}\Sigma_{\tilde{x}\tilde{x}}^{-1}\Sigma_{\tilde{x}x}\right]\right)$
  - $\phi_{\tilde{x}} \sim \mathcal{N}(m_{\tilde{x}}, \Sigma_{\tilde{x}\tilde{x}})$

Changing the approximation for predictions

- x_i^*: vector of explanatory variables for test set observation i
- Projected Process Approximation (PPA) / Deterministic Training Conditional (DTC)
- Fully Independent Training Conditional (FITC)
- $\phi_x^* | \phi_{\tilde{x}} \sim \mathcal{N}\left(m_{x^*} + \Sigma_{x^*\tilde{x}} \Sigma_{\tilde{x}\tilde{x}}^{-1} (\phi_{\tilde{x}} - m_{\tilde{x}}), \Sigma_{x^*x^*} - \Sigma_{x^*\tilde{x}} \Sigma_{\tilde{x}\tilde{x}}^{-1} \Sigma_{\tilde{x}x^*}\right)$

Summary of Inducing Point Methods

Do the following quantities match the full GP?

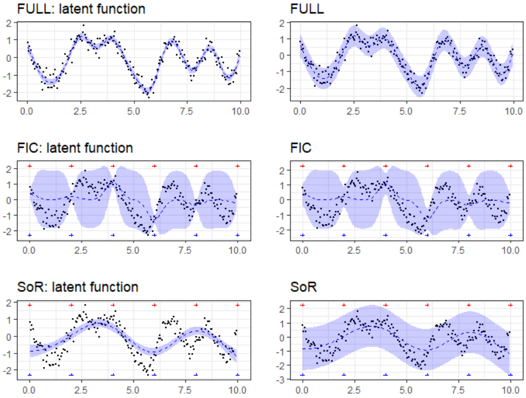
	Training variances	Test variances	Test covariances
SoR	NO	YES	NO
PPA/DTC	NO	YES	YES
FIC	YES	YES	NO
FITC	YES	YES	YES

In a single layer GP, the *posterior approximation* used in Damianou and Lawrence is the *same* as that resulting from the PPA/DTC model.

Summary of Inducing Point Methods (cont.)

- All methods result in conditional independence of $\phi_{x_i}|\phi_{\tilde{x}}$, so no sparse method defines the same prior covariances on the *training set* as the full GP.
- Posterior predictive variances/covariances are different for all methods.
- Changing the prior for *only* the variances / covariances on the test set does *not* change the posterior predictive mean.

Model Comparison when K is too small



Fitting models with likelihood optimization

PPA/DTC and SoR both use the *same* likelihood, so fitted paramters will be the same. The same goes for FIC and FITC. So what?

- SoR and PPA/DTC overestimate noise variance to compensate for lack of model flexibility/poorly approximated marginal function variances. *If* the PPA/DTC "correction" fixes the posterior function variance, the PPA/DTC models *must* overestimate posterior variance of the response variables.
- FIC and FITC do not share this relationship, but there may be undesirable consequences in terms of the posterior covariances.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen
A Unifying View of Sparse Approximate Gaussian Process Regression
Journal of Machine Learning Research, 2005, 6, 1939-1959.

Michalis K. Titsias
Variational Learning of Inducing Variables in Sparse Gaussian Processes
In *Artificial Intelligence and Statistics, 2009, 567-574* .

Michalis K. Titsias
Variational Model Selection for Sparse Gaussian Process Regression
Technical report, School of Computer Science, University of Manchester, 2009.