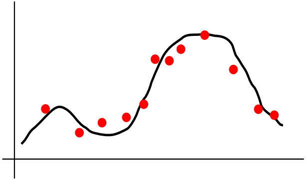
Efficient Sparse Approximations for Convolution Processes

Mauricio A. Álvarez

Joint work with Neil Lawrence, David Luengo and Michalis Titsias

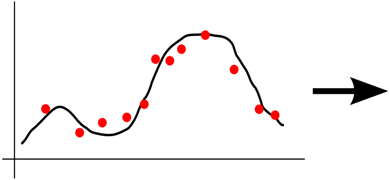
University of Manchester

Introduction: covariances for multiple outputs



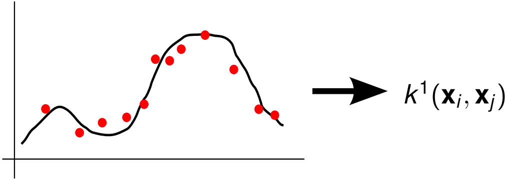
$$\mathcal{D}^1 = \{(\mathbf{x}_i^1, y_i^1) | i = 1, \ldots, N_1\}$$

Introduction: covariances for multiple outputs



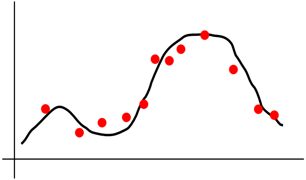
$$\mathcal{D}^1 = \{(\mathbf{x}_i^1, y_i^1) | i = 1, \ldots, N_1\}$$

Introduction: covariances for multiple outputs



$$\mathcal{D}^1 = \{(\mathbf{x}_i^1, y_i^1) | i = 1, \ldots, N_1\}$$

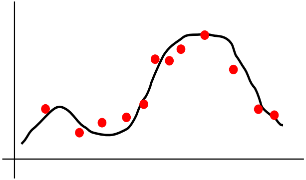
Introduction: covariances for multiple outputs



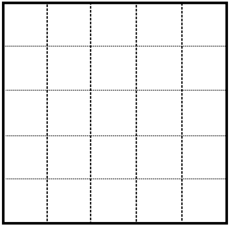
$\rightarrow k^1(\mathbf{x}_i, \mathbf{x}_j) \rightarrow$

$$\mathcal{D}^1 = \{(\mathbf{x}_i^1, y_i^1) | i = 1, \ldots, N_1\}$$

Introduction: covariances for multiple outputs

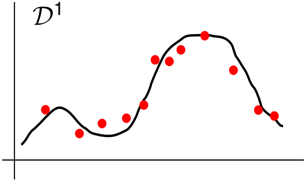


$\rightarrow k^1(\mathbf{x}_i, \mathbf{x}_j) \rightarrow \mathbf{K}^1 =$

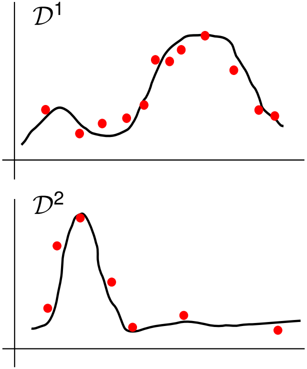


$\mathcal{D}^1 = \{(\mathbf{x}_i^1, y_i^1) | i = 1, \ldots, N_1\}$

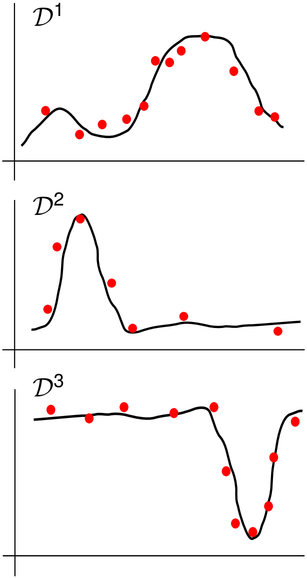
Introduction: covariances for multiple outputs



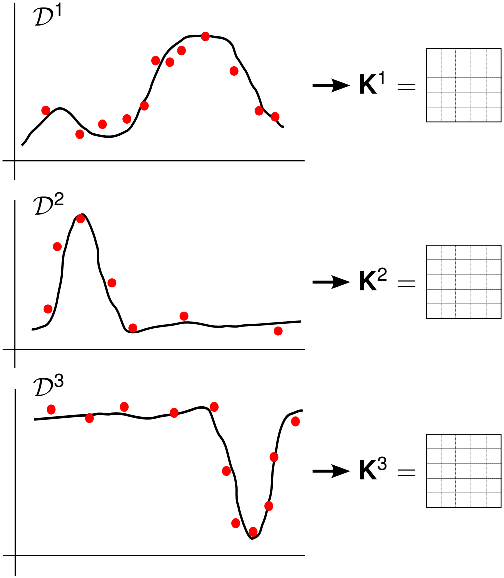
Introduction: covariances for multiple outputs



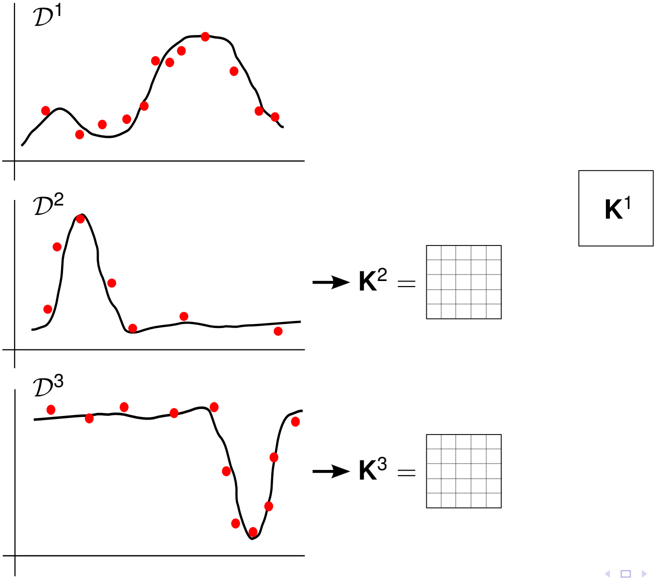
Introduction: covariances for multiple outputs



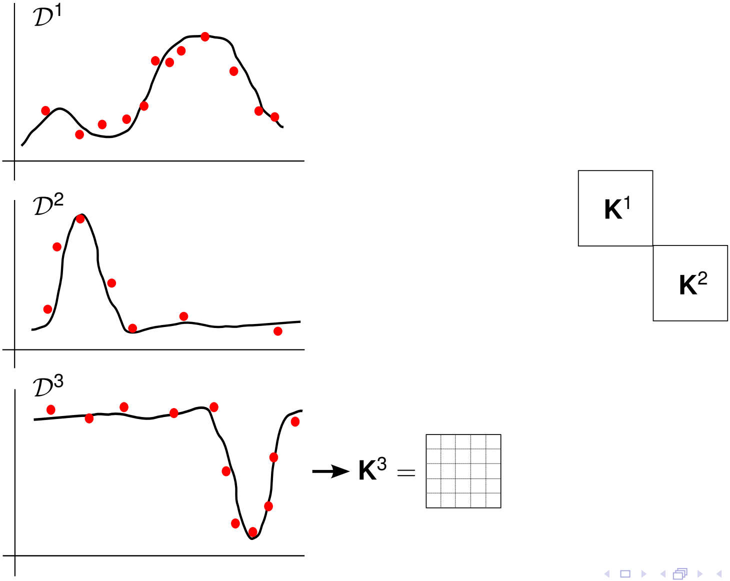
Introduction: covariances for multiple outputs



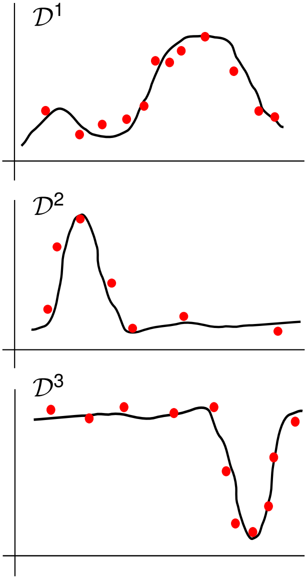
Introduction: covariances for multiple outputs

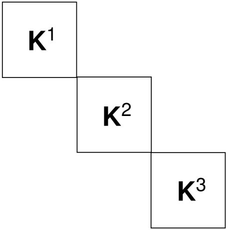


Introduction: covariances for multiple outputs

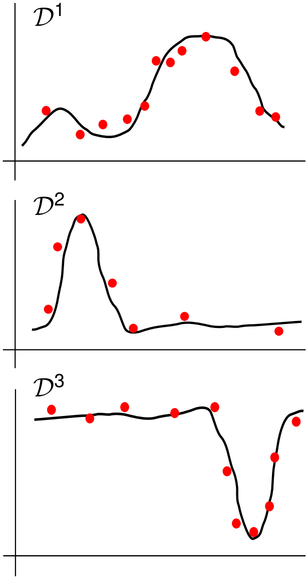


Introduction: covariances for multiple outputs



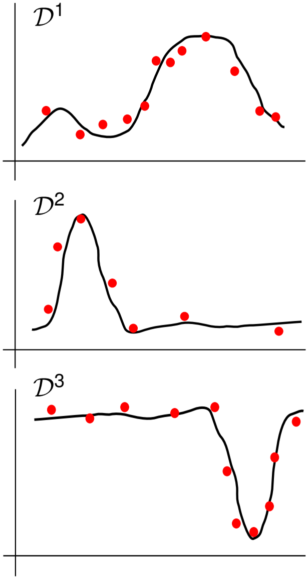


Introduction: covariances for multiple outputs



$$\mathbf{K} = \begin{bmatrix} \mathbf{K}^1 & & \\ & \mathbf{K}^2 & \\ & & \mathbf{K}^3 \end{bmatrix}$$

Introduction: covariances for multiple outputs



Joint covariance

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}^1 & ? & ? \\ ? & \mathbf{K}^2 & ? \\ ? & ? & \mathbf{K}^3 \end{bmatrix}$$

K be a valid covariance matrix

Some approaches

- Linear model of coregionalization.
- Intrinsic coregionalization model.
- Multitask kernels.
- Convolution of covariances.
- Convolution of processes or convolution process.

Convolution Process

- A convolution process is a moving-average construction that guarantees a valid covariance function.
- Consider a set of functions $\{f_d(\mathbf{x})\}_{d=1}^{D}$.
- Each function can be expressed as

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

- Influence of more than one latent function, $\{u_q(\mathbf{z})\}_{q=1}^{Q}$ and inclusion of an independent process $w_d(\mathbf{x})$

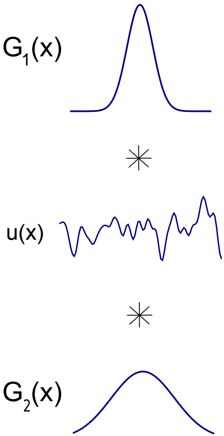
$$y_d(\mathbf{x}) = f_d(\mathbf{x}) + w_d(\mathbf{x}) = \sum_{q=1}^{Q} \int_{\mathcal{X}} G_{d,q}(\mathbf{x}-\mathbf{z})u_q(\mathbf{z})\mathrm{d}\mathbf{z} + w_d(\mathbf{x}).$$

A pictorial representation

u(x) 

u(x): latent function.

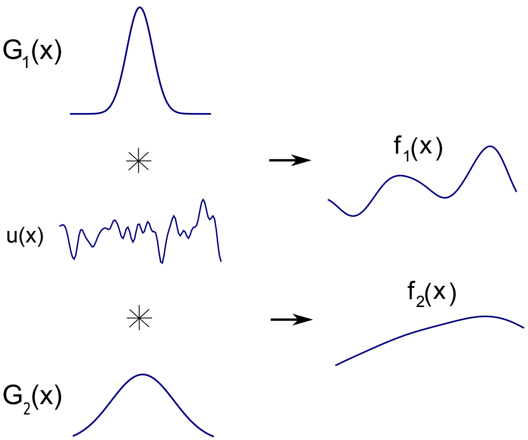
A pictorial representation



u(x): latent function.

G(x): smoothing kernel.

A pictorial representation

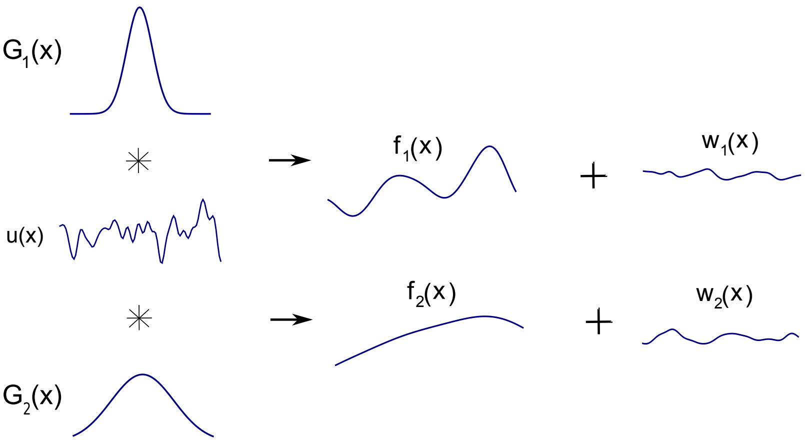


u(x): latent function.

G(x): smoothing kernel.

f(x): output function.

A pictorial representation



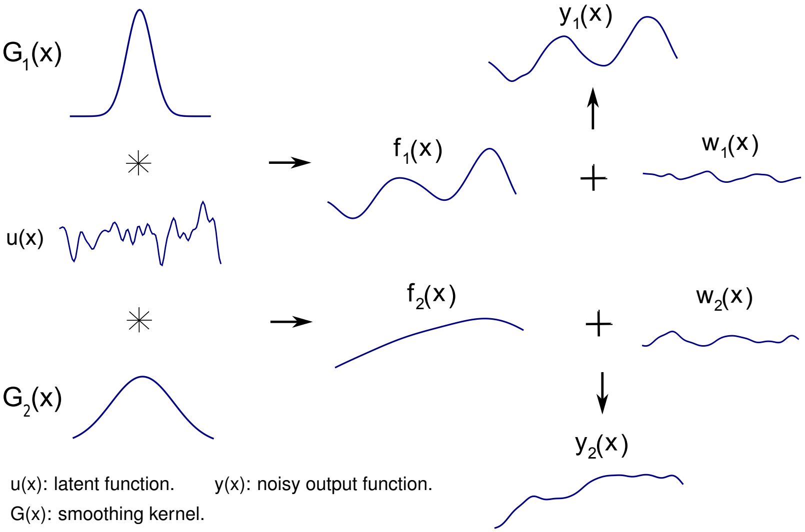
u(x): latent function.

G(x): smoothing kernel.

f(x): output function.

w(x): independent process.

A pictorial representation



u(x): latent function. y(x): noisy output function.

G(x): smoothing kernel.

f(x): output function.

w(x): independent process.

Covariance of the output functions.

The covariance between $y_d(\mathbf{x})$ and $y_{d'}(\mathbf{x}')$ is given as

$$\text{cov}\left[y_d(\mathbf{x}), y_{d'}(\mathbf{x}')\right] = \text{cov}\left[f_d(\mathbf{x}), f_{d'}(\mathbf{x}')\right] + \text{cov}\left[w_d(\mathbf{x}), w_{d'}(\mathbf{x}')\right]\delta_{d,d'}$$

where

$$\text{cov}\left[f_d(\mathbf{x}), f_{d'}(\mathbf{x}')\right] = \int_{\mathcal{X}} G_d(\mathbf{x}-\mathbf{z}) \int_{\mathcal{X}} G_{d'}(\mathbf{x}'-\mathbf{z}')\, \text{cov}\left[u(\mathbf{z}), u(\mathbf{z}')\right] \mathrm{d}\mathbf{z}' \mathrm{d}\mathbf{z}$$

Different forms of covariance for the output functions.

- Input *Gaussian process*

$$\text{cov}\left[f_d, f_{d'}\right] = \int_{\mathcal{X}} G_d(\mathbf{x}-\mathbf{z}) \int_{\mathcal{X}} G_{d'}(\mathbf{x}'-\mathbf{z}') k_{u,u}(\mathbf{z},\mathbf{z}')\mathrm{d}\mathbf{z}'\mathrm{d}\mathbf{z}$$

- Input *white noise process*

$$\text{cov}\left[f_d, f_{d'}\right] = \int_{\mathcal{X}} G_d(\mathbf{x}-\mathbf{z}) G_{d'}(\mathbf{x}'-\mathbf{z})\mathrm{d}\mathbf{z}$$

- Covariance between output functions and latent functions

$$\text{cov}\left[f_d, u\right] = \int_{\mathcal{X}} G_d(\mathbf{x}-\mathbf{z}') k_{u,u}(\mathbf{z}',\mathbf{z})\mathrm{d}\mathbf{z}'$$

Likelihood of the full Gaussian process.

- The likelihood of the model is given by

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

  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 cov $[f_d, f_{d'}]$, Σ 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 $\mathbf{K}_{\mathbf{f},\mathbf{f}} + \Sigma$, which grows with complexity $\mathcal{O}(N^3 D^3)$

Predictive distribution of the full Gaussian process.

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

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

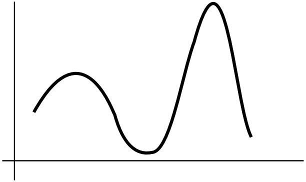
with

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

- Prediction is $\mathcal{O}(DN)$ for the mean and $\mathcal{O}(D^2N^2)$ for the variance, for one test point. Storage is $\mathcal{O}(D^2N^2)$.

Conditional prior distribution.

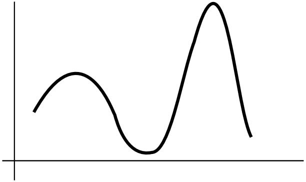
Sample from $p(u)$



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

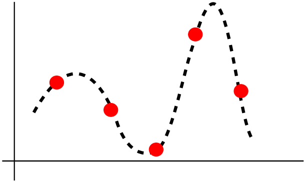
Conditional prior distribution.

Sample from $p(u)$



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

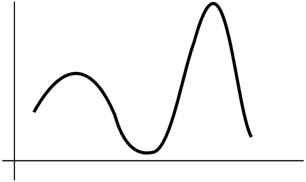
Discretize u



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

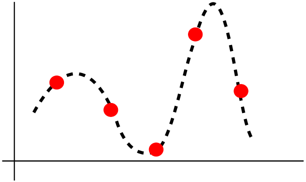
Conditional prior distribution.

Sample from $p(u)$



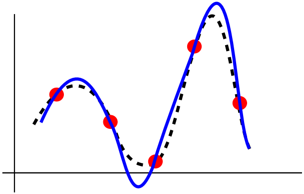
$$f_d(\mathbf{x}) = \int_{\mathcal{X}} G_d(\mathbf{x}-\mathbf{z})u(\mathbf{z})\mathrm{d}\mathbf{z}$$

Discretize u



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

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



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

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.

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

Better approximations can be obtained when $\mathrm{E}\left[u|\mathbf{u}\right]$ 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 **u**, the marginal likelihood is given as

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

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2}$	$\mathbf{K}_{f_1 f_3}$
$\mathbf{K}_{f_2 f_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3}$
$\mathbf{K}_{f_3 f_1}$	$\mathbf{K}_{f_3 f_2}$	$\mathbf{K}_{f_3 f_3}$

$\approx$

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2}$	$\mathbf{K}_{f_1 f_3} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3}$
$\mathbf{K}_{f_2 f_1} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3}$
$\mathbf{K}_{f_3 f_1} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1}$	$\mathbf{K}_{f_3 f_2} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2}$	$\mathbf{K}_{f_3 f_3}$

Comparison of marginal likelihoods

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

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

$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_1}$	$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_2}$	$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_3}$
$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_1}$	$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_2}$	$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_3}$
$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_1}$	$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_2}$	$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_3}$

$\approx$

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

$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_1}$	$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_2}$	$\mathbf{K}_{\mathbf{f}_1\mathbf{f}_3}$
$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_1}$	$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_2}$	$\mathbf{K}_{\mathbf{f}_2\mathbf{f}_3}$
$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_1}$	$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_2}$	$\mathbf{K}_{\mathbf{f}_3\mathbf{f}_3}$

$\approx$ $\mathbf{G}$ $\times$ $\mathbf{G}^{T}$

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

Predictive distribution for the sparse approximation

Predictive distribution

$$p(\mathbf{y}_*|\mathbf{y}, \mathbf{X}, \mathbf{X}_*, \mathbf{Z}, \theta) = \mathcal{N}\left(\widetilde{\mu}_*, \widetilde{\Lambda}_*\right), \text{ with}$$

$$\widetilde{\mu}_* = \mathbf{K}_{\mathbf{f}_*,\mathbf{u}}\mathbf{A}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}}(\mathbf{D} + \Sigma)^{-1}\mathbf{y}$$

$$\widetilde{\Lambda}_* = \mathbf{D}_* + \mathbf{K}_{\mathbf{f}_*,\mathbf{u}}\mathbf{A}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}_*} + \Sigma$$

$$\mathbf{A} = \mathbf{K}_{\mathbf{u},\mathbf{u}} + \mathbf{K}_{\mathbf{u},\mathbf{f}}(\mathbf{D} + \Sigma)^{-1}\mathbf{K}_{\mathbf{f},\mathbf{u}}$$

$$\mathbf{D}_* = \text{blockdiag}\left[\mathbf{K}_{\mathbf{f}_*,\mathbf{f}_*} - \mathbf{K}_{\mathbf{f}_*,\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{u},\mathbf{f}_*}\right]$$

Remarks

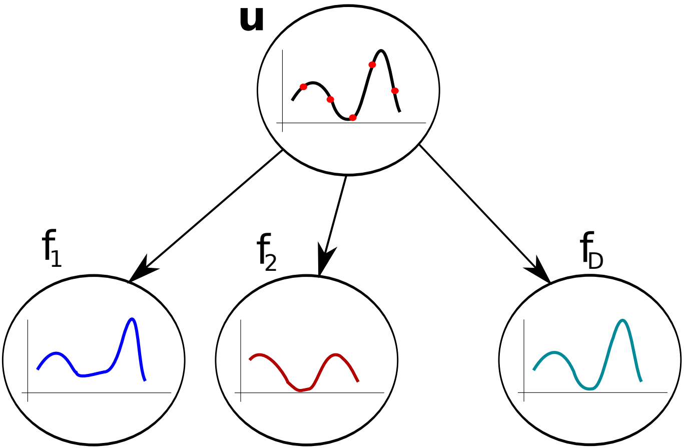
- For learning the computational demand is in the calculation of $\mathbf{D}^{-1}$, which grows as $\mathcal{O}(N^3D) + \mathcal{O}(NDM^2)$ (with $R = 1$). Storage is $\mathcal{O}(N^2D) + \mathcal{O}(NDM)$.
- For inference, the computation of the mean grows as $\mathcal{O}(DM)$ and the computation of 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 [QR05].

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.

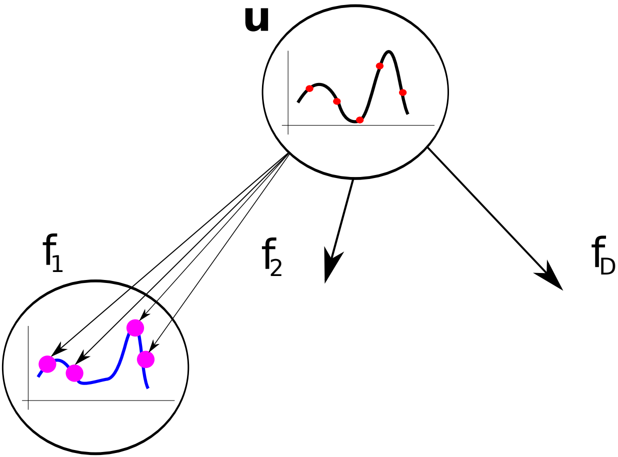
Additional conditional independencies

- 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

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

Comparison of marginal likelihoods

The marginal likelihood is given as

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

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2}$	$\mathbf{K}_{f_1 f_3}$
$\mathbf{K}_{f_2 f_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3}$
$\mathbf{K}_{f_3 f_1}$	$\mathbf{K}_{f_3 f_2}$	$\mathbf{K}_{f_3 f_3}$

$\approx$

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2}$	$\mathbf{K}_{f_1 f_3} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3}$
$\mathbf{K}_{f_2 f_1} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3}$
$\mathbf{K}_{f_3 f_1} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1}$	$\mathbf{K}_{f_3 f_2} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2}$	$\mathbf{K}_{f_3 f_3}$

Comparison of marginal likelihoods

The marginal likelihood is given as

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

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2}$	$\mathbf{K}_{f_1 f_3}$
$\mathbf{K}_{f_2 f_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3}$
$\mathbf{K}_{f_3 f_1}$	$\mathbf{K}_{f_3 f_2}$	$\mathbf{K}_{f_3 f_3}$

$\approx$

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2}$	$\mathbf{K}_{f_1 f_3} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3}$
$\mathbf{K}_{f_2 f_1} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3}$
$\mathbf{K}_{f_3 f_1} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1}$	$\mathbf{K}_{f_3 f_2} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2}$	$\mathbf{K}_{f_3 f_3}$

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2}$	$\mathbf{K}_{f_1 f_3}$
$\mathbf{K}_{f_2 f_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3}$
$\mathbf{K}_{f_3 f_1}$	$\mathbf{K}_{f_3 f_2}$	$\mathbf{K}_{f_3 f_3}$

$\approx$

$\mathbf{Q}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2}$	$\mathbf{K}_{f_1 f_3} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3}$
$\mathbf{K}_{f_2 f_1} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1}$	$\mathbf{Q}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3}$
$\mathbf{K}_{f_3 f_1} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1}$	$\mathbf{K}_{f_3 f_2} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2}$	$\mathbf{Q}_{f_3 f_3}$

Comparison of marginal likelihoods

The marginal likelihood is given as

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

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2}$	$\mathbf{K}_{f_1 f_3}$
$\mathbf{K}_{f_2 f_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3}$
$\mathbf{K}_{f_3 f_1}$	$\mathbf{K}_{f_3 f_2}$	$\mathbf{K}_{f_3 f_3}$

$\approx$

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2}$	$\mathbf{K}_{f_1 f_3} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3}$
$\mathbf{K}_{f_2 f_1} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3}$
$\mathbf{K}_{f_3 f_1} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1}$	$\mathbf{K}_{f_3 f_2} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2}$	$\mathbf{K}_{f_3 f_3}$

$\mathbf{K}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2}$	$\mathbf{K}_{f_1 f_3}$
$\mathbf{K}_{f_2 f_1}$	$\mathbf{K}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3}$
$\mathbf{K}_{f_3 f_1}$	$\mathbf{K}_{f_3 f_2}$	$\mathbf{K}_{f_3 f_3}$

$\approx$

$\mathbf{Q}_{f_1 f_1}$	$\mathbf{K}_{f_1 f_2} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2}$	$\mathbf{K}_{f_1 f_3} - \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3}$
$\mathbf{K}_{f_2 f_1} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1}$	$\mathbf{Q}_{f_2 f_2}$	$\mathbf{K}_{f_2 f_3} - \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_3}$
$\mathbf{K}_{f_3 f_1} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_1}$	$\mathbf{K}_{f_3 f_2} - \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{u f_2}$	$\mathbf{Q}_{f_3 f_3}$

Comparison of marginal likelihoods

The marginal likelihood is given as

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

$\mathbf{K}_{f_1f_1}(x_1,x_1)$	$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(x_1,x_2)$	$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(x_1,x_3)$
$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(x_2,x_1)$	$\mathbf{K}_{f_1f_1}(x_2,x_2)$	$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(x_2,x_3)$
$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(x_3,x_1)$	$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(x_3,x_2)$	$\mathbf{K}_{f_1f_1}(x_3,x_3)$

$\mathbf{K}_{f_1f_1}$	$\mathbf{K}_{f_1f_2}$	$\mathbf{K}_{f_1f_3}$
$\mathbf{K}_{f_2f_1}$	$\mathbf{K}_{f_2f_2}$	$\mathbf{K}_{f_2f_3}$
$\mathbf{K}_{f_3f_1}$	$\mathbf{K}_{f_3f_2}$	$\mathbf{K}_{f_3f_3}$

$\approx$

$\mathbf{Q}_{f_1f_1}$	$\mathbf{K}_{f_1f_2} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2}$	$\mathbf{K}_{f_1f_3} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3}$
$\mathbf{K}_{f_2f_1} - \mathbf{K}_{f_2u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1}$	$\mathbf{Q}_{f_2f_2}$	$\mathbf{K}_{f_2f_3} - \mathbf{K}_{f_2u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3}$
$\mathbf{K}_{f_3f_1} - \mathbf{K}_{f_3u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1}$	$\mathbf{K}_{f_3f_2} - \mathbf{K}_{f_3u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2}$	$\mathbf{Q}_{f_3f_3}$

Comparison of marginal likelihoods

The marginal likelihood is given as

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

$\mathbf{K}_{f_1f_1}(\mathbf{x}_1,\mathbf{x}_1)$	$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(\mathbf{x}_1,\mathbf{x}_2)$	$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(\mathbf{x}_1,\mathbf{x}_3)$
$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(\mathbf{x}_2,\mathbf{x}_1)$	$\mathbf{K}_{f_1f_1}(\mathbf{x}_2,\mathbf{x}_2)$	$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(\mathbf{x}_2,\mathbf{x}_3)$
$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(\mathbf{x}_3,\mathbf{x}_1)$	$(\mathbf{K}_{f_1f_1} - \mathbf{K}_{f_1u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1})(\mathbf{x}_3,\mathbf{x}_2)$	$\mathbf{K}_{f_1f_1}(\mathbf{x}_3,\mathbf{x}_3)$

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

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 [QR05, SG06].

Examples using PITC and FITC

- For all our experiments we considered squared exponential covariance functions for the latent process of the form

$$k_{u,u}(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{1}{2}\left(\mathbf{x} - \mathbf{x}'\right)^{\top} \mathbf{L}\left(\mathbf{x} - \mathbf{x}'\right)\right],$$

where $\mathbf{L}$ is a diagonal matrix which allows for different length-scales along each dimension.

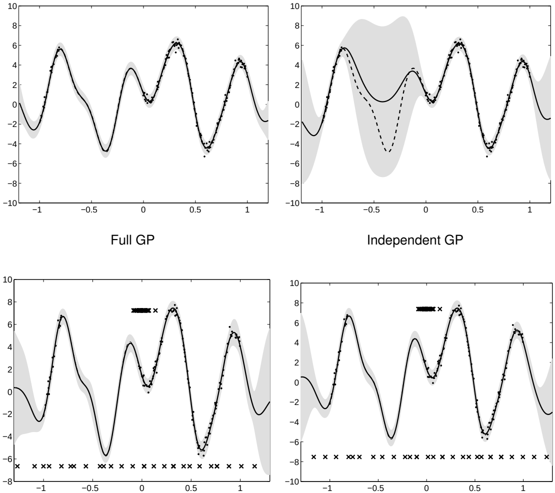
- The smoothing kernel had the same form,

$$G_d(\boldsymbol{\tau}) = \frac{S_d|\mathbf{L}_d|^{1/2}}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}\boldsymbol{\tau}^{\top}\mathbf{L}_d\boldsymbol{\tau}\right],$$

where $S_d \in \mathbb{R}$ and $\mathbf{L}_d$ is a symmetric positive definite matrix.

Examples using PITC and FITC: Artificial data 1D

Four outputs generated from the full GP ($D = 4$).



Full GP Independent GP

FITC PITC

Jura Data set I

- Measurements of concentrations of seven heavy metals collected in the topsoil of a 14.5 km^2 region of the Swiss Jura.
- Prediction set (259 locations) and a validation set (100 locations).

Primary variable	**Secondary Variables**
Cd	Ni, Zn
Cu	Pb, Ni, Zn

- Optimisation of the locations of the inducing inputs.

Jura Data set II

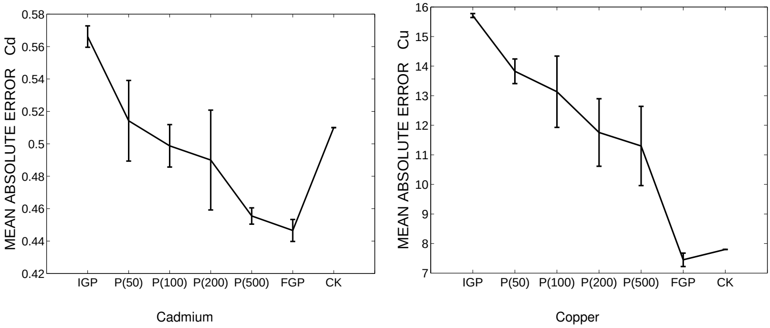


Figure: Mean absolute error for IGP: Independent GP, P(M): PITC with M inducing points, FGP: Full GP, CK: Ordinary Co-kriging

Comparison of marginal likelihoods

- To obtain the above approximations, we have replaced the exact likelihood
$$p(\mathbf{f}|\theta) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_{\mathbf{f},\mathbf{f}} + \Sigma)$$
for the approximated one
$$p(\mathbf{f}|\theta, \mathbf{Z}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{Q}_{\mathbf{f},\mathbf{f}}(\mathbf{Z}) + \Sigma),$$
where θ corresponds to the hyperparameters of the model.

- In other words, we have changed the model and additionally, we have introduced new hyperparameters $\mathbf{Z}$.

- Without additional restrictions, maximization of the approximated marginal likelihood over $\mathbf{Z}$ might lead to overfitting.

An alternative

- A different way to face the problem is to use approximate inference to the exact model.
- Since obtaining the posterior over u is intractable (computational complexity grows as $\mathcal{O}(N^3D^3)$), we propose to approximate the posterior using variational inference.

Variational inference in one slide

- Variational inference idea: to fit a variational distribution to the true posterior minimizing the Kullback-Leibler divergence

$$\mathrm{KL}(q \parallel p) = -\int q(u) \log \left\{ \frac{p(u|\mathbf{y})}{q(u)} \right\} \mathrm{d}u.$$

- Minimizing the KL divergence is equivalent to maximize the lower bound

$$\log \int p(\mathbf{y}, u) \mathrm{d}u \geq \mathcal{L}(q) = \int q(u) \log \left\{ \frac{p(u, \mathbf{y})}{q(u)} \right\} \mathrm{d}u$$

Variational inference for convolution processes

- We augment the joint distribution $p(\mathbf{y}, u)$ with a set of variables $\mathbf{u}$

$$p\left(\mathbf{y}, u, \mathbf{u}\right) = p(\mathbf{y}|u)p(u|\mathbf{u})p(\mathbf{u}).$$

- We want to approximate the true posterior $p(u, \mathbf{u}|\mathbf{y})$ with a distribution

$$q(u, \mathbf{u}) = p(u|\mathbf{u})\phi(\mathbf{u}),$$

where $\phi(\mathbf{u})$ represents the approximated posterior over the latent variables $\mathbf{u}$.

Lower bound for the marginal likelihood

- The distribution $q(u, \mathbf{u})$ is approximated minimizing the KL distance.
- Equivalently, we maximize the following lower bound

$$\mathcal{L}(\mathbf{Z}, \phi(\mathbf{u})) = \int_{u,\mathbf{u}} q(u, \mathbf{u}) \log \left\{ \frac{p(\mathbf{y}, u, \mathbf{u})}{q(u, \mathbf{u})} \right\} \mathrm{d}\mathbf{u}\, \mathrm{d}u$$
$$= \int_{u,\mathbf{u}} p(u|\mathbf{u})\phi(\mathbf{u}) \log \left\{ \frac{p(\mathbf{y}|u)p(u|\mathbf{u})p(\mathbf{u})}{p(u|\mathbf{u})\phi(\mathbf{u})} \right\} \mathrm{d}\mathbf{u}\, \mathrm{d}u$$

- Maximizing the lower bound with respect to $\phi(\mathbf{u})$

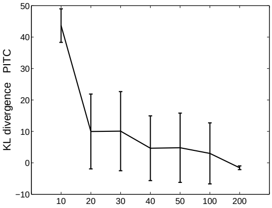
$$\mathcal{L}(\mathbf{Z}, \theta) = \log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K_{f,u}K_{u,u}^{-1}K_{u,f}} + \Sigma\right) - \frac{1}{2} \operatorname{trace}\left[\Sigma^{-1}\left(\mathbf{K_{f,f}} - \mathbf{K_{f,u}K_{u,u}^{-1}K_{u,f}}\right)\right].$$

Remarks

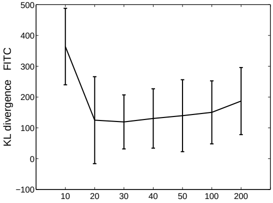
- Expressions for the (approximated) posterior $\phi(\mathbf{u})$ and the predictive distribution follow similar forms that for the PITC and FITC approximations.
- The computational complexity is again $\mathcal{O}(NDM^2)$ plus an aditional trace operation.
- The form of the likelihood obtained if we remove the trace term is similar to the Deterministic Training Conditional (DTC).
- Since we have an additional trace term and a variational treatment we call this approximation DTC-VAR.

An illustration: artificial data 1D revisited

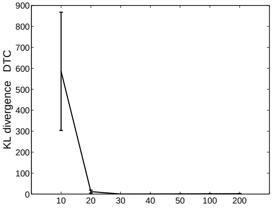
Measuring the KL divergence for the 1D toy example above



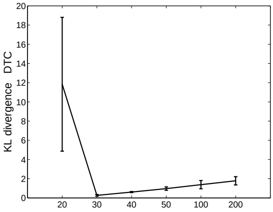
PITC



FITC



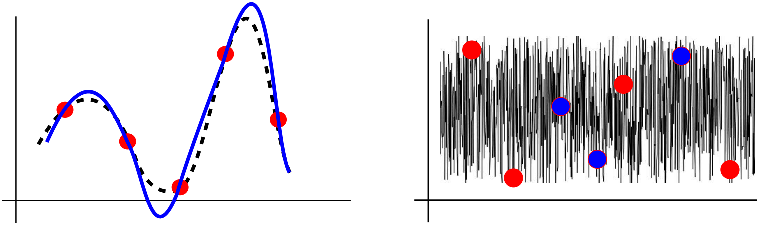
DTC-VAR



DTC-VAR with zoom

Input functions as white noise processes

- The key assumption for the approximations before is that we can express the conditional prior $p(u|\mathbf{u})$.
- In other words, that the latent functions can be summarized using just a few points.
- If the input function corresponds to a white noise process this is certainly not true.



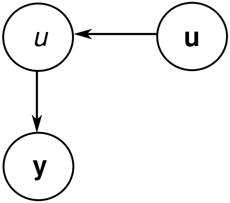
Variational inducing kernel

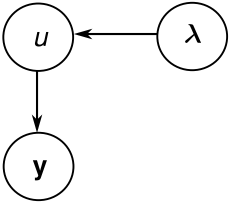
- Instead of applying the variational framework described before to a finite set of inducing points $\mathbf{u}$, we compute the bound with respect to a finite set of points $\boldsymbol{\lambda}$ obtained from the process

$$\lambda(\mathbf{z}) = \int_{\mathcal{X}} T(\mathbf{z}-\mathbf{z}')u(\mathbf{z}')\mathrm{d}\mathbf{z}'.$$

- We refer to the smoothing kernel $T(\mathbf{z}-\mathbf{z}')$ as the inducing kernel.
- Under this setup, the set of points $\boldsymbol{\lambda}$ are informative about the white noise process.

Comparison





$$p\left(\mathbf{y}, u, \mathbf{u}\right) = p(\mathbf{y}|u)p(u|\mathbf{u})p(\mathbf{u}).$$

$$p\left(\mathbf{y}, u, \boldsymbol{\lambda}\right) = p(\mathbf{y}|u)p(u|\boldsymbol{\lambda})p(\boldsymbol{\lambda}).$$

$\mathbf{u}$ is uninformative

$\boldsymbol{\lambda}$ is informative

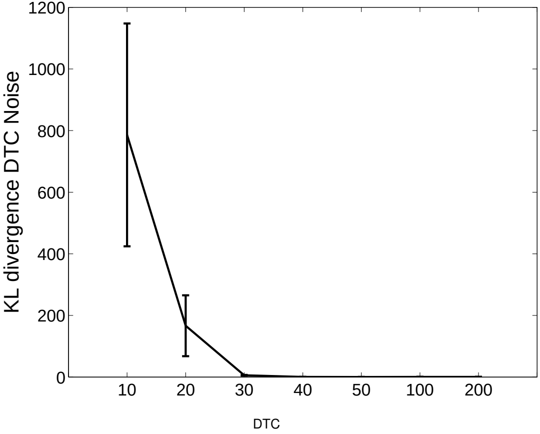
Lower bound

Under the same analysis that before, the variational lower bound is obtained as

$$\mathcal{L}(\mathbf{Z}, T, \theta) = \log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}_{\mathbf{f},\lambda}\mathbf{K}_{\lambda,\lambda}^{-1}\mathbf{K}_{\lambda,\mathbf{f}} + \Sigma\right) - \frac{1}{2}\,\text{trace}\left[\Sigma^{-1}\left(\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{K}_{\mathbf{f},\lambda}\mathbf{K}_{\lambda,\lambda}^{-1}\mathbf{K}_{\lambda,\mathbf{f}}\right)\right]$$

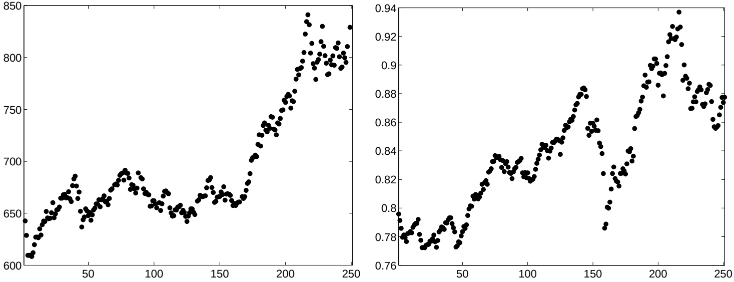
Example

Measuring the KL divergence for a 1D toy example



Financial data set

Multivariate financial data set: the dollar prices of the 3 precious metals and top 10 currencies.



Gold AUD

Dynamic model

- Our model: a set of coupled differential equations, driven by either a Gaussian process, a white noise process or both,

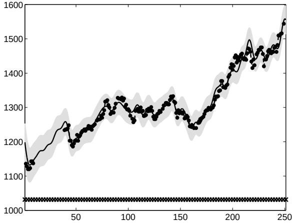
$$\frac{\mathrm{d}f_d(t)}{\mathrm{d}t} = B_d f_d(t) + S_d u(t),$$

  where B_d is a decay coefficient and S_d quantifies the influence of the process $u(t)$.

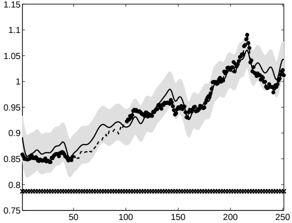
- If $u(t)$ is a white noise process $\rightarrow$ Langevin equation $\rightarrow$ a linear stochastic differential equation.

- Solution for $f_d(t)$ has the form of convolutions. For a single output and white noise process, $f_d(t) \rightarrow$ Ornstein-Uhlenbeck (OU) process.

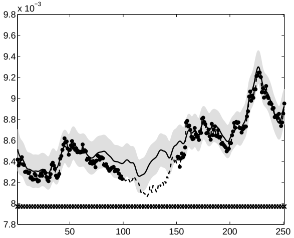
Some results



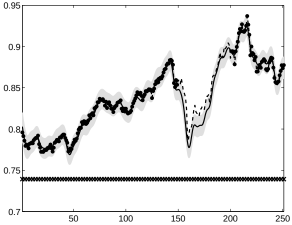
XPT: Real data and prediction



CAD: Real data and prediction



JPY: Real data and prediction



AUD: Real data and prediction

Open questions

- Choice of the kernel function
- Experimental comparison
- Online learning
- Theoretical connections between methods.
- Computational complexity
- How the inference is affected with different variants of spatial configuration (isotopic vs heterotopic).
- Is there any theoretical way to know beforehand when considering the cross-covariance might help?

References I

David M. Higdon.
Space and space-time modelling using process convolutions.
In C. Anderson, V. Barnett, P. Chatwin, and A. El-Shaarawi, editors, *Quantitative methods for current environmental issues*, pages 37–56. Springer-Verlag, 2002.

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.

Edward Snelson and Zoubin Ghahramani.
Sparse Gaussian processes using pseudo-inputs.
In Yair Weiss, Bernhard Schölkopf, and John C. Platt, editors, *NIPS*, volume 18, Cambridge, MA, 2006. MIT Press.

Michalis K. Titsias.
Variational learning of inducing variables in sparse Gaussian processes.
In David van Dyk and Max Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 567–574, Clearwater Beach, Florida, 16-18 April 2009. JMLR W&CP 5.