Prior Knowledge and Sparse Methods for Convolved Multiple Outputs Gaussian Processes

Mauricio A. Álvarez

Joint work with Neil D. Lawrence, David Luengo and Michalis K. Titsias

School of Computer Science
University of Manchester

Contents

- Latent force models.
- Sparse approximations for latent force models.

Data driven paradigm

- Traditionally, the main focus in machine learning has been model generation through a *data driven paradigm*.

- Combine a data set with a flexible class of models and, through regularization, make predictions on unseen data.

- Problems
  - Data is scarce relative to the complexity of the system.
  - Model is forced to extrapolate.

Mechanistic models

- Models inspired by the underlying knowledge of a physical system are common in many areas.

- Description of a well characterized physical process that underpins the system, typically represented with a set of differential equations.

- Identifying and specifying all the interactions might not be feasible.

- A mechanistic model can enable accurate prediction in regions where there may be no available training data

Hybrid systems

- We suggest a *hybrid approach* involving a mechanistic model of the system augmented through machine learning techniques.

- Dynamical systems (e.g. incorporating first order and second order differential equations).

- Partial differential equations for systems with multiple inputs.

Latent variable model: definition

- Our approach can be seen as a type of latent variable model.

$$\mathbf{Y} = \mathbf{U}\mathbf{W} + \mathbf{E},$$

where $\mathbf{Y} \in \mathbb{R}^{N \times D}$, $\mathbf{U} \in \mathbb{R}^{N \times Q}$, $\mathbf{W} \in \mathbb{R}^{Q \times D}$ $(Q < D)$ and $\mathbf{E}$ is a matrix variate white Gaussian noise with columns $\mathbf{e}_{:,d} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$.

- In PCA and FA the common approach to deal with the unknowns is to integrate out **U** under a Gaussian prior and optimize with respect to **W**.

Latent variable model: alternative view

- Data with temporal nature and Gaussian (Markov) prior for rows of **U** leads to the Kalman filter/smoother.

- Consider a joint distribution for $p\left(\mathbf{U}|\mathbf{t}\right)$, $\mathbf{t} = \left[t_1 \ldots t_N\right]^\top$, with the form of a Gaussian process (GP),

$$p\left(\mathbf{U}|\mathbf{t}\right) = \prod_{q=1}^{Q} \mathcal{N}\left(\mathbf{u}_{:,q}|\mathbf{0}, \mathbf{K}_{u_{:,q},u_{:,q}}\right).$$

  The latent variables are random functions, $\{u_q(t)\}_{q=1}^{Q}$ with associated covariance $\mathbf{K}_{u_{:,q},u_{:,q}}$.

- The GP for **Y** can be readily implemented. In [TSJ05] this is known as a semi-parametric latent factor model (SLFM).

Latent force model: mechanistic interpretation (1)

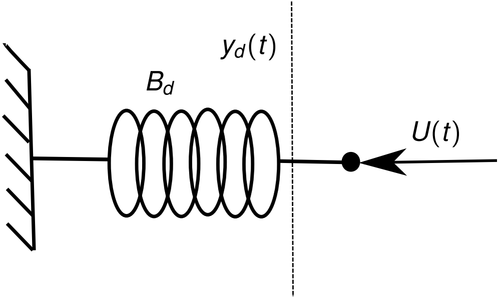
- We include a further dynamical system with a *mechanistic* inspiration.

- Reinterpret equation $\mathbf{Y} = \mathbf{U}\mathbf{W} + \mathbf{E}$, as a force balance equation

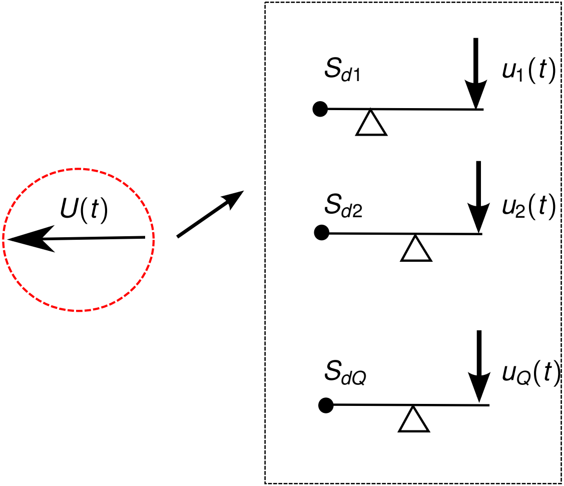
$$\mathbf{Y}\mathbf{B} = \mathbf{U}\mathbf{S} + \widetilde{\mathbf{E}},$$

where $\mathbf{S} \in \mathbb{R}^{Q \times D}$ is a matrix of sensitivities, $\mathbf{B} \in \mathbb{R}^{D \times D}$ is diagonal matrix of spring constants, $\mathbf{W} = \mathbf{S}\mathbf{B}^{-1}$ and $\widetilde{\mathbf{e}}_{:,d} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{B}^{\top}\boldsymbol{\Sigma}\mathbf{B}\right)$.

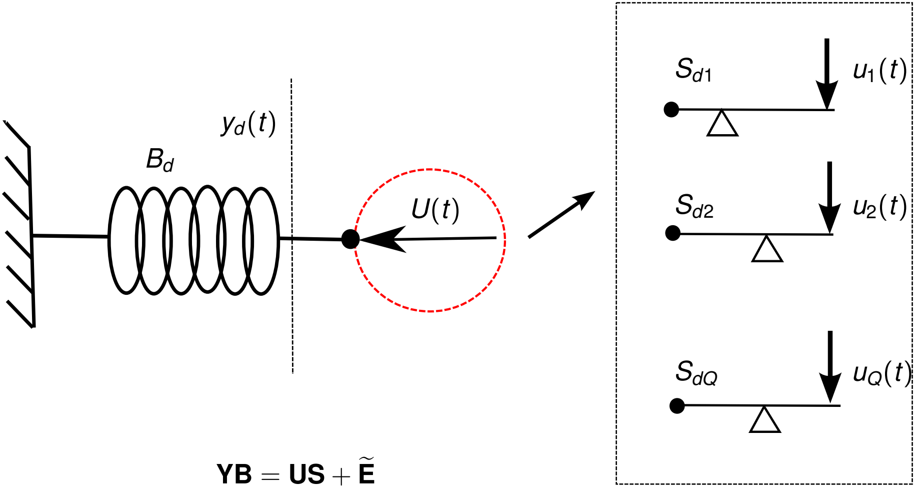
Latent force model: mechanistic interpretation (2)



Latent force model: mechanistic interpretation (2)



Latent force model: mechanistic interpretation (2)



$$\mathbf{YB} = \mathbf{US} + \widetilde{\mathbf{E}}$$

Latent force model: extension (1)

- The model can be extended including dampers and masses.

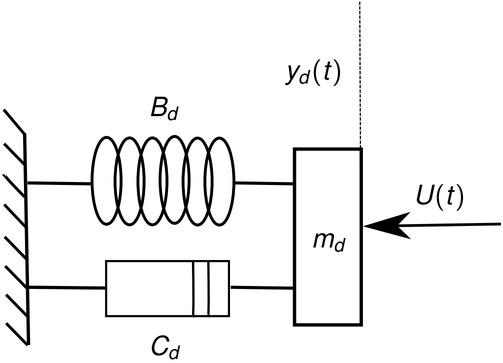
- We can write

$$\mathbf{YB} + \dot{\mathbf{Y}}\mathbf{C} + \ddot{\mathbf{Y}}\mathbf{M} = \mathbf{US} + \widehat{\boldsymbol{E}},$$

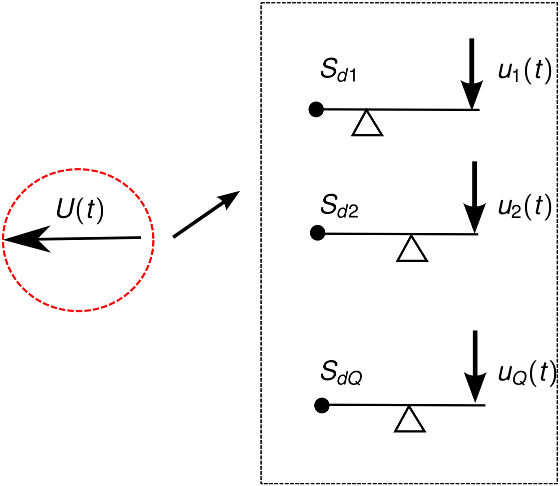
where

$\dot{\mathbf{Y}}$ is the first derivative of $\mathbf{Y}$ w.r.t. time
$\ddot{\mathbf{Y}}$ is the second derivative of $\mathbf{Y}$ w.r.t. time
$\mathbf{C}$ is a diagonal matrix of damping coefficients
$\mathbf{M}$ is a diagonal matrix of masses
$\widehat{\boldsymbol{E}}$ is a matrix variate white Gaussian noise.

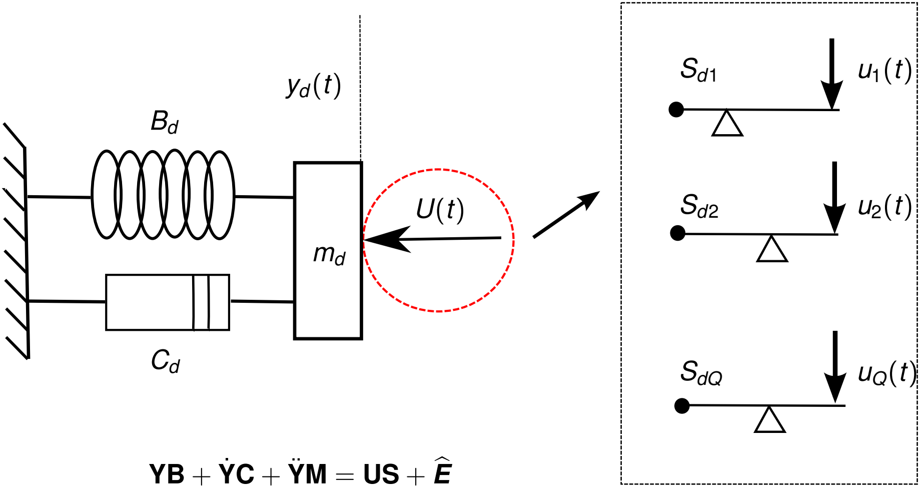
Latent force model: extension (2)



Latent force model: extension (2)



Latent force model: extension (2)



$$\mathbf{Y}\mathbf{B} + \dot{\mathbf{Y}}\mathbf{C} + \ddot{\mathbf{Y}}\mathbf{M} = \mathbf{U}\mathbf{S} + \widehat{\boldsymbol{E}}$$

Latent force model: properties

- This model allows to include behaviors like inertia and resonance.
- We refer to these systems as *latent force models* (LFMs).
- One way of thinking of our model is to consider puppetry.

Second Order Dynamical System

Using the system of second order differential equations

$$m_d \frac{\mathrm{d}^2 y_d(t)}{\mathrm{d}t^2} + C_d \frac{\mathrm{d}y_d(t)}{\mathrm{d}t} + B_d y_d(t) = \sum_{q=1}^{Q} S_{dq} u_q(t),$$

where

$u_q(t)$	latent forces
$y_d(t)$	displacements over time
C_d	damper constant for the *d*-th output
B_d	spring constant for the *d*-th output
m_d	mass constant for the *d*-th output
S_{dq}	sensitivity of the *d*-th output to the *q*-th input.

Second Order Dynamical System: solution

Solving for $y_d(t)$, we obtain

$$y_d(t) = \sum_{q=1}^{Q} L_{dq}[u_q](t),$$

where the linear operator is given by a convolution:

$$L_{dq}[u_q](t) = \frac{S_{dq}}{\omega_d} \int_0^t \exp(-\alpha_d(t-\tau)) \sin(\omega_d(t-\tau)) u_q(\tau) \mathrm{d}\tau,$$

with $\omega_d = \sqrt{4B_d - C_d^2}/2$ and $\alpha_d = C_d/2$.

Second Order Dynamical System: covariance matrix

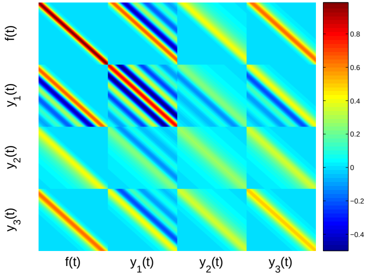
Behaviour of the system summarized by the damping ratio:

$$\zeta_d = \frac{1}{2} C_d / \sqrt{B_d}$$

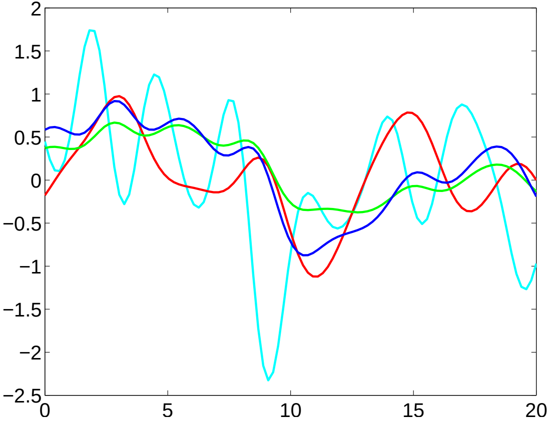
$\zeta_d > 1$	overdamped system
$\zeta_d = 1$	critically damped system
$\zeta_d < 1$	underdamped system
$\zeta_d = 0$	undamped system (no friction)

Example covariance matrix:

$\zeta_1 = 0.125$	underdamped
$\zeta_2 = 2$	overdamped
$\zeta_3 = 1$	critically damped

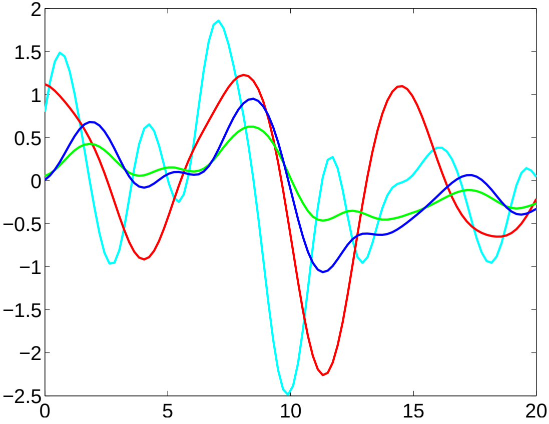


Second Order Dynamical System: samples from GP



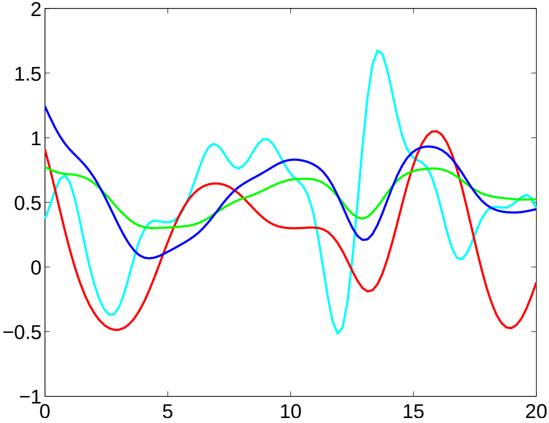
Joint samples from the ODE covariance, *cyan*: $u(t)$, *red*: $y_1(t)$(underdamped) and *green*: $y_2(t)$ (overdamped) and *blue*: $y_3(t)$ (critically damped).

Second Order Dynamical System: samples from GP



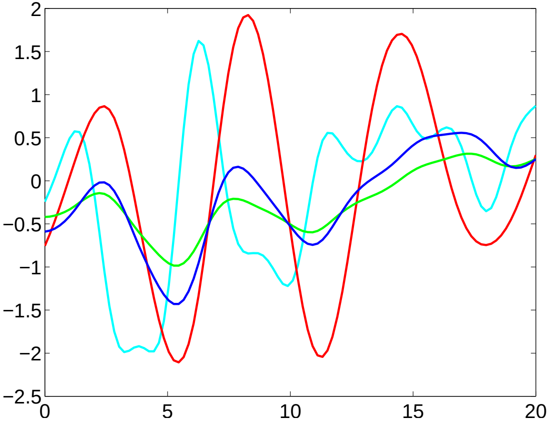
Joint samples from the ODE covariance, *cyan*: $u(t)$, *red*: $y_1(t)$(underdamped) and *green*: $y_2(t)$ (overdamped) and *blue*: $y_3(t)$ (critically damped).

Second Order Dynamical System: samples from GP



Joint samples from the ODE covariance, *cyan*: $u(t)$, *red*: $y_1(t)$(underdamped) and *green*: $y_2(t)$ (overdamped) and *blue*: $y_3(t)$ (critically damped).

Second Order Dynamical System: samples from GP



Joint samples from the ODE covariance, *cyan*: $u\,(t)$, *red*: $y_1\,(t)$(underdamped) and *green*: $y_2\,(t)$ (overdamped) and *blue*: $y_3\,(t)$ (critically damped).

Motion Capture Data (1)

- CMU motion capture data, motions 18, 19 and 20 from subject 49.

- Motions 18 and 19 for training and 20 for testing.

Motion Capture Data (2)

- The data down-sampled by 32 (from 120 frames per second to 3.75).

- We focused on the subject's left arm.

- For testing, we condition only on the observations of the shoulder's orientation (motion 20) to make predictions for the rest of the arm's angles.

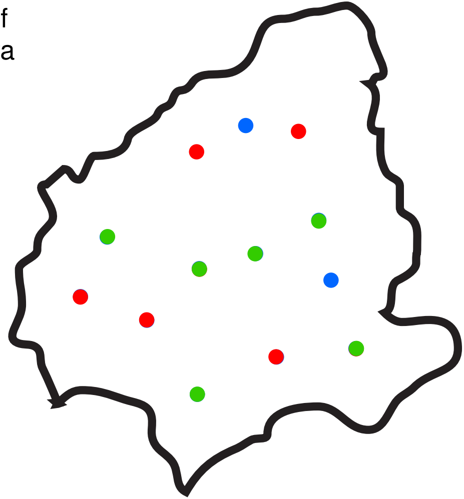
Motion Capture Results

Root mean squared (RMS) angle error for prediction of the left arm's configuration in the motion capture data. Prediction with the latent force model outperforms the prediction with regression for all apart from the radius's angle.

Angle	Latent Force Error	Regression Error
Radius	4.11	**4.02**
Wrist	**6.55**	6.65
Hand X rotation	**1.82**	3.21
Hand Z rotation	**2.76**	6.14
Thumb X rotation	**1.77**	3.10
Thumb Z rotation	**2.73**	6.09

Diffussion in the Swiss Jura

Region of
Swiss Jura



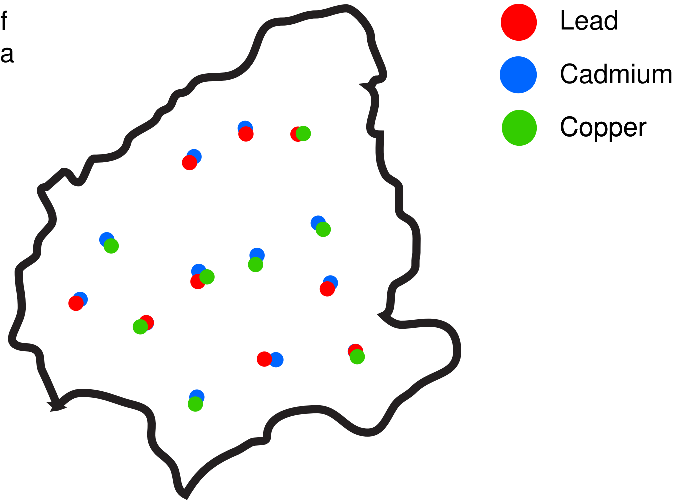
🔴 Lead

🔵 Cadmium

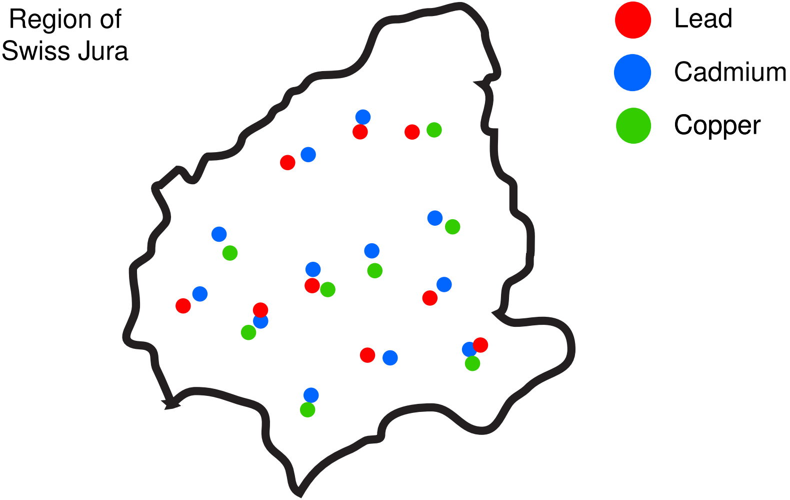
🟢 Copper

Diffussion in the Swiss Jura

Region of
Swiss Jura



Diffussion in the Swiss Jura



Region of Swiss Jura

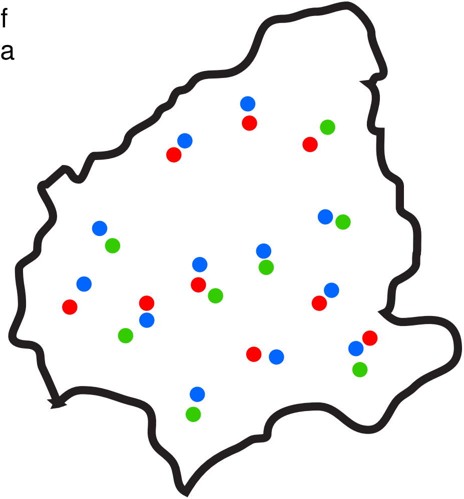
● Lead

● Cadmium

● Copper

Diffussion in the Swiss Jura

Region of Swiss Jura



● Lead

● Cadmium

● Copper

Diffusion equation

- A simplified version of the diffusion equation is

$$\frac{\partial y_d(\mathbf{x},t)}{\partial t} = \sum_{j=1}^{p} \kappa_d \frac{\partial^2 y_d(\mathbf{x},t)}{\partial x_j^2},$$

where $y_d(\mathbf{x},t)$ are the concentrations of each pollutant.

- The solution to the system is then given by

$$y_d(\mathbf{x},t) = \sum_{q=1}^{Q} S_{dq} \int_{\mathbb{R}^p} G_d(\mathbf{x},\mathbf{x}',t) u_q(\mathbf{x}')\mathrm{d}\mathbf{x}',$$

where $u_q(\mathbf{x})$ represents the concentration of pollutants at time zero and $G_d(\mathbf{x},\mathbf{x}',t)$ is the Green's function given as

$$G_d(\mathbf{x},\mathbf{x}',t) = \frac{1}{2^p \pi^{p/2} T_d^{p/2}} \exp\left[-\sum_{j=1}^{p} \frac{(x_j - x_j')^2}{4T_d}\right],$$

with $T_d = \kappa_d t$.

Prediction of Metal Concentrations

- Prediction of a *primary variable* by conditioning on the values of some *secondary variables*.

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

- Comparison bewteen diffusion kernel, independent GPs and “ordinary co-kriging”.

Metals	IGPs	GPDK	OCK
Cd	0.5823±0.0133	**0.4505±0.0126**	0.5
Cu	15.9357±0.0907	**7.1677±0.2266**	7.8
Pb	22.9141±0.6076	**10.1097±0.2842**	10.7
Co	2.0735±0.1070	1.7546±0.0895	**1.5**

LFM in the context of convolution processes

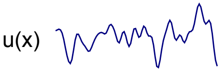
- Consider a set of functions $\{f_d(\mathbf{x})\}_{d=1}^{D}$.
- Each function can be expressed as

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

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

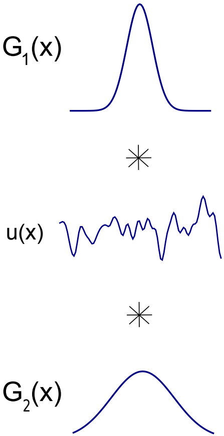
$$y_d(\mathbf{x}) = f_d(\mathbf{x}) + w_d(\mathbf{x}) = \sum_{q=1}^{Q} \int_{\mathcal{X}} G_{dq}(\mathbf{x} - \mathbf{z}) u_q(\mathbf{z}) \mathrm{d}\mathbf{z} + w_d(\mathbf{x}).$$

A pictorial representation



u(x): latent function.

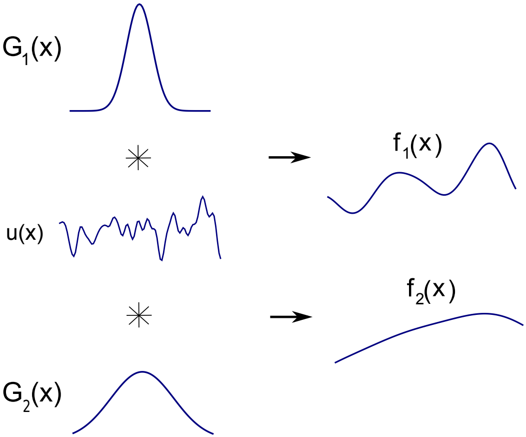
A pictorial representation



u(x): latent function.

G(x): smoothing kernel.

A pictorial representation

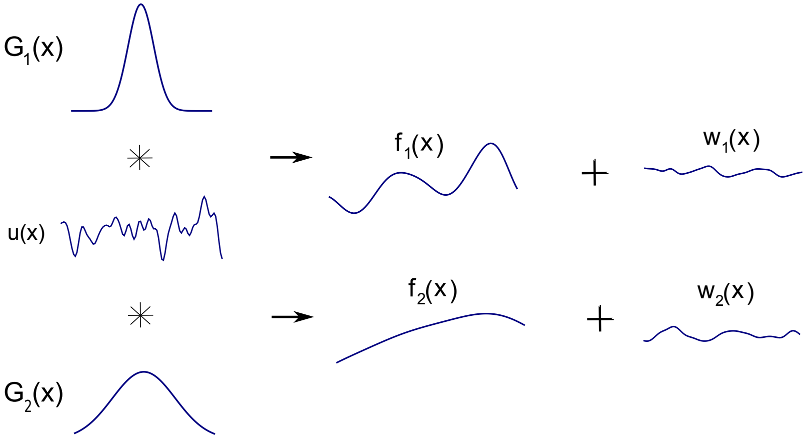


$u(x)$: latent function.

$G(x)$: smoothing kernel.

$f(x)$: output function.

A pictorial representation



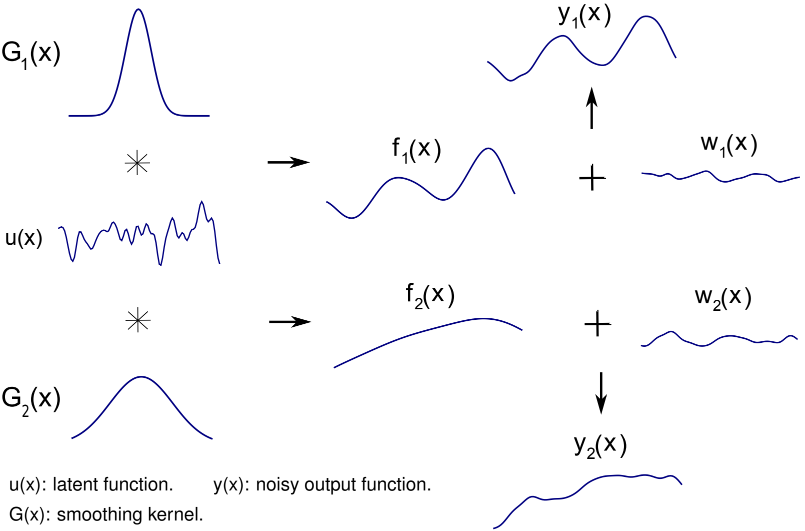
u(x): latent function.

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

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

where $\mathrm{cov}\left[f_d(\mathbf{x}), f_{d'}(\mathbf{x}')\right]$

$$\sum_{q=1}^{Q} \sum_{q'=1}^{Q} \int_{\mathcal{X}} G_{dq}(\mathbf{x}-\mathbf{z}) \int_{\mathcal{X}} G_{d'q'}(\mathbf{x}'-\mathbf{z}')\, \mathrm{cov}\left[u_q(\mathbf{z}), u_{q'}(\mathbf{z}')\right] \mathrm{d}\mathbf{z}' \mathrm{d}\mathbf{z}$$

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 $\mathrm{cov}\left[f_d, f_{d'}\right]$, 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 $\mathbf{X}_*$

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

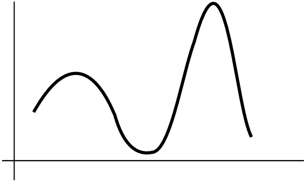
with

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

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

Conditional prior distribution.

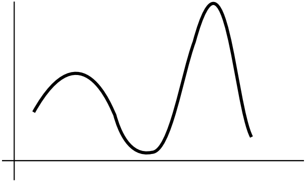
Sample from $p(u)$



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

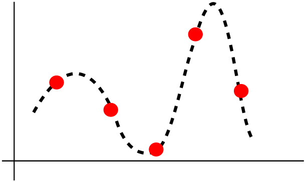
Conditional prior distribution.

Sample from $p(u)$



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

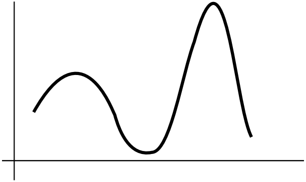
Discretize u



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

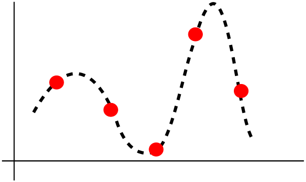
Conditional prior distribution.

Sample from $p(u)$



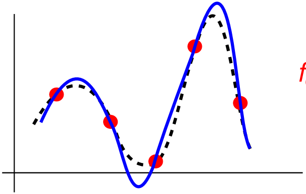
$$f_d(\mathbf{x}) = \int_{\mathcal{X}} G_d(\mathbf{x}-\mathbf{z})u(\mathbf{z})\mathrm{d}\mathbf{z}$$

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*.

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

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*.

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

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

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

Comparison of marginal likelihoods

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

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

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

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

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

Predictive distribution for the sparse approximation

Predictive distribution

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

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

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.

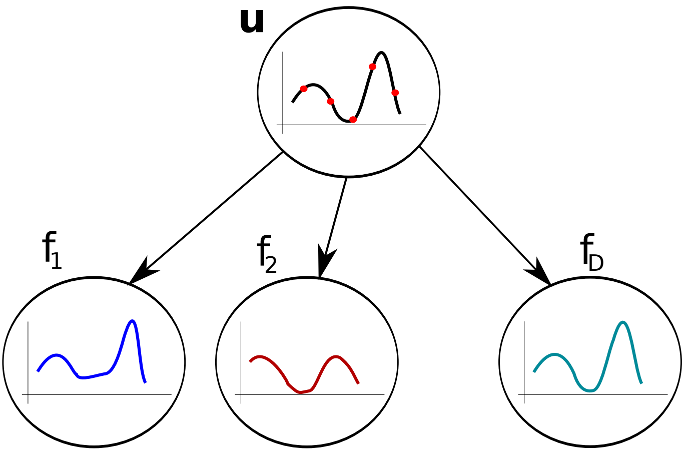
- The functional form of the approximation is almost identical to that of the Partially Independent Training Conditional (PITC) approximation [QR05].

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.

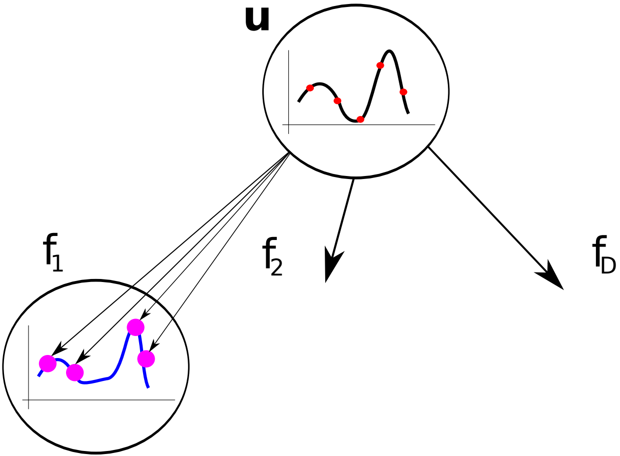
Additional conditional independencies

- 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

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

Comparison of marginal likelihoods

The marginal likelihood is given as

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

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

Comparison of marginal likelihoods

The marginal likelihood is given as

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

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

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

Comparison of marginal likelihoods

The marginal likelihood is given as

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

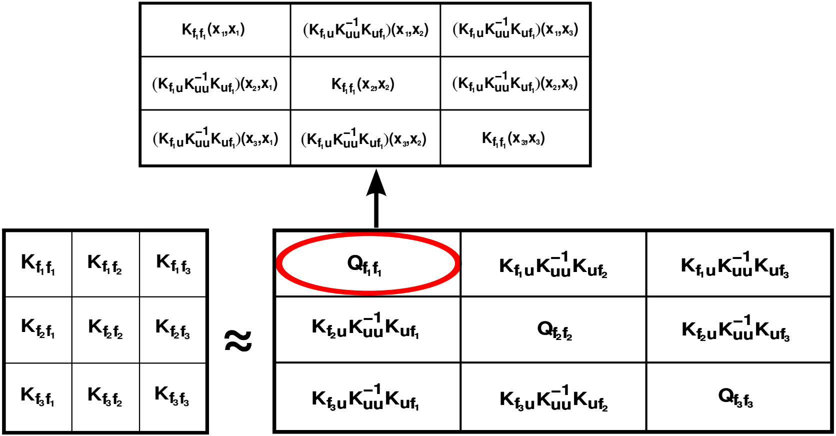
$$\begin{bmatrix} \mathbf{K}_{f_1 f_1} & \mathbf{K}_{f_1 f_2} & \mathbf{K}_{f_1 f_3} \\ \mathbf{K}_{f_2 f_1} & \mathbf{K}_{f_2 f_2} & \mathbf{K}_{f_2 f_3} \\ \mathbf{K}_{f_3 f_1} & \mathbf{K}_{f_3 f_2} & \mathbf{K}_{f_3 f_3} \end{bmatrix} \approx \begin{bmatrix} \mathbf{K}_{f_1 f_1} & \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2} & \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3} \\ \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1} & \mathbf{K}_{f_2 f_2} & \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3} \\ \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1} & \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2} & \mathbf{K}_{f_3 f_3} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{K}_{f_1 f_1} & \mathbf{K}_{f_1 f_2} & \mathbf{K}_{f_1 f_3} \\ \mathbf{K}_{f_2 f_1} & \mathbf{K}_{f_2 f_2} & \mathbf{K}_{f_2 f_3} \\ \mathbf{K}_{f_3 f_1} & \mathbf{K}_{f_3 f_2} & \mathbf{K}_{f_3 f_3} \end{bmatrix} \approx \begin{bmatrix} \boxed{\mathbf{Q}_{f_1 f_1}} & \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2} & \mathbf{K}_{f_1 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3} \\ \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1} & \mathbf{Q}_{f_2 f_2} & \mathbf{K}_{f_2 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_3} \\ \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_1} & \mathbf{K}_{f_3 u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf_2} & \mathbf{Q}_{f_3 f_3} \end{bmatrix}$$

Comparison of marginal likelihoods

The marginal likelihood is given as

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



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

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

$\approx$

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

Comparison of marginal likelihoods

The marginal likelihood is given as

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

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

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

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 [QR05, SG06].

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).

Examples

- 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 **L** is a diagonal matrix which allows for different length-scales along each dimension.

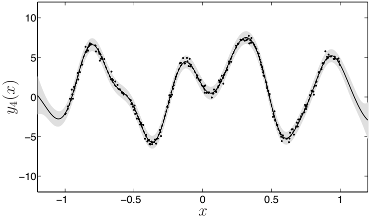
- The smoothing kernel had the same form,

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

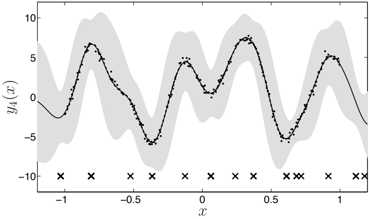
  where $S_d \in \mathbb{R}$ and $\mathbf{L}_d$ is a symmetric positive definite matrix.

Examples: Artificial data 1D

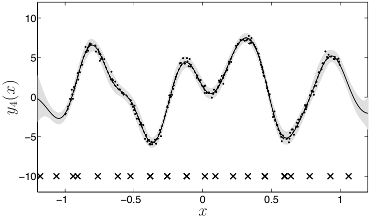
Four outputs generated from the full GP ($D = 4$).



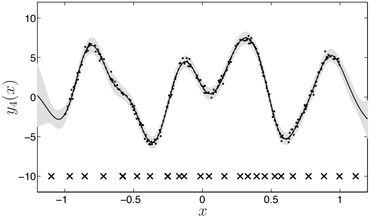
$y_4(x)$ using the full GP



$y_4(x)$ using the DTC approximation



$y_4(x)$ using the FITC approximation



$y_4(x)$ using the PITC approximation

Artificial example (cont.)

Method	SMSE $y_1(x)$	SMSE $y_2(x)$	SMSE $y_3(x)$	SMSE $y_4(x)$
Full GP	1.06 ± 0.08	0.99 ± 0.06	1.10 ± 0.09	1.05 ± 0.09
DTC	1.06 ± 0.08	0.99 ± 0.06	1.12 ± 0.09	1.05 ± 0.09
FITC	1.06 ± 0.08	0.99 ± 0.06	1.10 ± 0.08	1.05 ± 0.08
PITC	1.06 ± 0.08	0.99 ± 0.06	1.10 ± 0.09	1.05 ± 0.09

Standarized mean square error (SMSE). All numbers are to be multiplied by 10^{-2}.

Method	MSLL $y_1(x)$	MSLL $y_2(x)$	MSLL $y_3(x)$	MSLL $y_4(x)$
Full GP	-2.27 ± 0.04	-2.30 ± 0.03	-2.25 ± 0.04	-2.27 ± 0.05
DTC	-0.98 ± 0.18	-0.98 ± 0.18	-1.25 ± 0.16	-1.25 ± 0.16
FITC	-2.26 ± 0.04	-2.29 ± 0.03	-2.16 ± 0.04	-2.23 ± 0.05
PITC	-2.27 ± 0.04	-2.30 ± 0.03	-2.23 ± 0.04	-2.26 ± 0.05

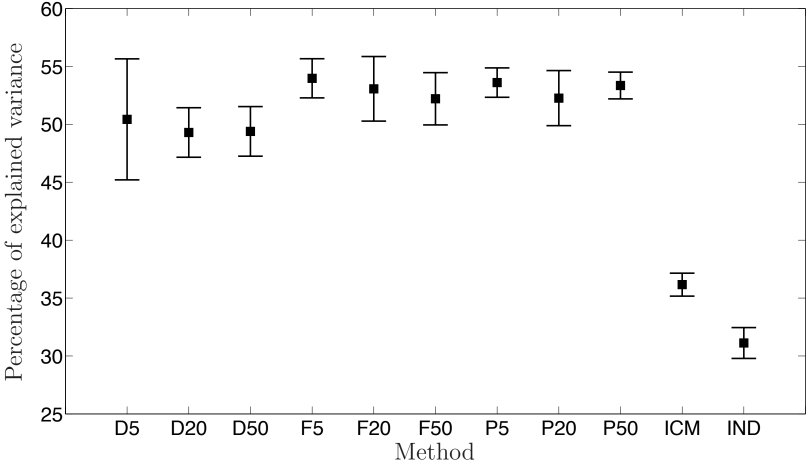
Mean standardized log loss (MSLL). More negative values indicate better models.

Training times for iteration of each model are 1.97 ± 0.02 secs for the full GP, 0.20 ± 0.01 secs for DTC, 0.41 ± 0.03 for FITC and 0.59 ± 0.05 for the PITC.

Predicting school examination scores

- Multitask learning problem.
- The goal is to predict the exam score obtained by a particular student described by a set of 20 features belonging to a specific school (task).
- It consists of examination records from 139 secondary schools in years 1985, 1986 and 1987.
- Features include year of the exam, gender, VR band and ethnic group for each student, which are transformed to dummy variables.
- Dataset consists of 4004 samples. Ten repetitions with 75% training and 25% testing.
- Gaussian smoothing function.

Predicting school examination scores (cont.)



D: DTC. F: FITC. P: PITC. ICM: Intrinsic coregionalization model [BCW08]. IND: Independent GPs [BCW08].

A dynamic model for transcription regulation

- Microarray studies have made the simultaneous measurement of mRNA from thousands of genes practical.
- Transcription is governed by the presence of absence of transcription factor proteins that act as switches to turn on and off the expression of the genes.
- The active concentration of these transcription factors is typically much more difficult to measure.

A dynamic model for transcription regulation (cont.)

- There are Q transcription factors $\{u_q(t)\}_{q=1}^{Q}$, each of them represented through a Gaussian process, $u_q(t) \sim \mathcal{GP}\left(0, k_{u_q u_q}(t, t')\right)$.

- Our model is based on the following differential equation [ALL09],

$$\frac{\mathrm{d}f_d}{\mathrm{d}t} = \gamma_d + \sum_{q=1}^{Q} S_{dq} u_q(t) - B_d f_d(t),$$

where γ_d is the basal transcription rate of gene d, S_{dq} is the sensitivity of gene d to the transcription factor $u_q(t)$ and B_d is the decay rate of mRNA.

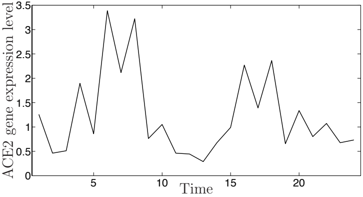
A dynamic model for transcription regulation (cont.)

- Benchmark yeast cell cycle dataset of [SSZ$^+$98].

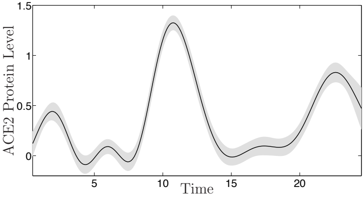
- Data is preprocessed as described in [SLR06] with a final dataset of 1975 genes and 104 transcription factors. There are 24 time points for each gene.

- We optimize the marginal likelihood through scaled conjugate gradient.

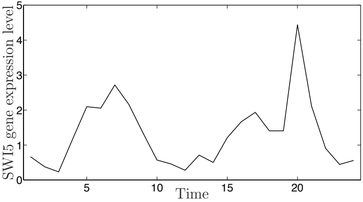
A dynamic model for transcription regulation (cont.)



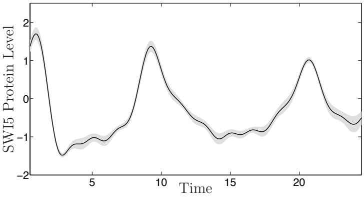
Gene expression profile for ACE2.



Protein concentration for ACE2.

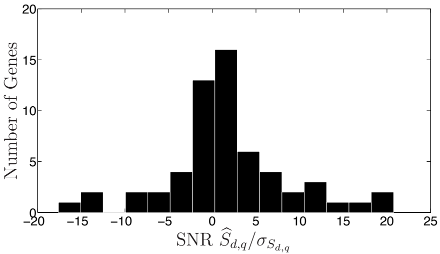


Gene expression profile for SWI5.

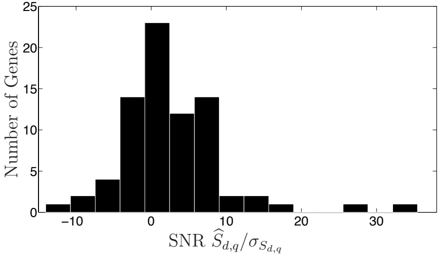


Protein concentration for SWI5.

A dynamic model for transcription regulation (cont.)



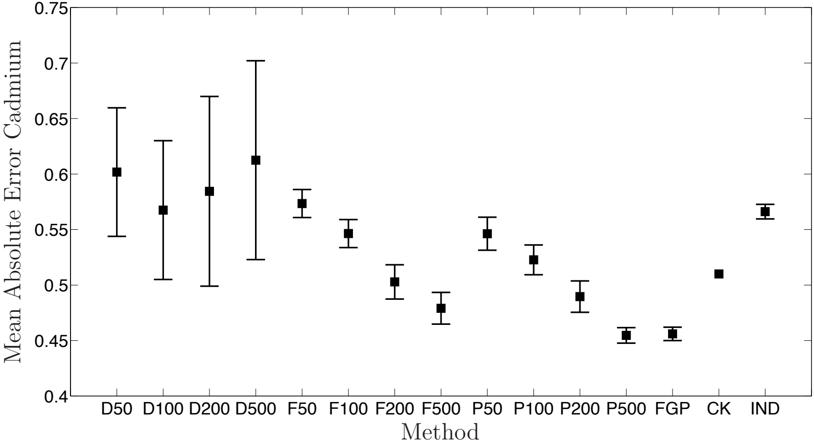
SNR associated to ACE2.



SNR associated to SWI5.

- For ACE2, highest SNR values are obtained for CTS1, SCW11, DSE1 and DSE2, while, for example, NCE4 appears to be repressed with a low SNR value ([SSZ$^+$98, SLR06]).
- SWI5 appears to activate genes AMN1 and PLC2 ([CLCB01]).

Swiss Jura example revisited



D: DTC. F: FITC. P: PITC. FGP: Full Gaussian Process. CK: Cokriging [Goo97]. IND: Independent GPs [BCW08].

Conclusions

❑ Hybrid approach for the use of simple mechanistic models with Gaussian processes.

❑ Convolution processes as a way to augment data-driven models with characteristics of physical systems.

❑ Gaussian process as meaningful prior distributions.

❑ Sparse approximations for multiple outputs convolved GP exploiting conditional independencies.

Acknowledgments

❑ To Google, for a Google Research Award.

❑ To EPSRC, for Grant No EP/F005687/1 “Gaussian Processes for Systems Identification with Applications in Systems Biology”.

References I

Mauricio Álvarez, David Luengo, and Neil D. Lawrence.
Latent Force Models.
In David van Dyk and Max Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 9–16, Clearwater Beach, Florida, 16-18 April 2009. JMLR W&CP 5.

Edwin V. Bonilla, Kian Ming Chai, and Christopher K. I. Williams.
Multi-task Gaussian process prediction.
In John C. Platt, Daphne Koller, Yoram Singer, and Sam Roweis, editors, *NIPS*, volume 20, Cambridge, MA, 2008. MIT Press.

Alejandro Colman-Lerner, Tina E. Chin, and Roger Brent.
Yeast cbk1 and mob2 activate daughter-specific genetic programs to induce asymmetric cell fates.
Cell, 107:739–750, 2001.

Pierre Goovaerts.
Geostatistics For Natural Resources Evaluation.
Oxford University Press, USA, 1997.

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.

Guido Sanguinetti, Neil D. Lawrence, and Magnus Rattray.
Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities.
Bioinformatics, 22:2275–2281, 2006.

References II

Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher.
Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.
Molecular Biology of the Cell, 9(12):3273–3297, 1998.

Yee Whye Teh, Matthias Seeger, and Michael I. Jordan.
Semiparametric latent factor models.
In Robert G. Cowell and Zoubin Ghahramani, editors, *AISTATS 10*, pages 333–340, Barbados, 6-8 January 2005. Society for Artificial Intelligence and Statistics.