Multi-Kernel Gaussian Processes

Arman Melkumyan
Australian Centre for Field Robotics
School of Aerospace, Mechanical and
Mechatronic Engineering
The University of Sydney
NSW 2006, Australia
a.melkumyan@acfr.usyd.edu.au

Fabio Ramos
Australian Centre for Field Robotics
School of Information Technologies
The University of Sydney
NSW 2006, Australia
f.ramos@acfr.usyd.edu.au

Abstract

Multi-task learning remains a difficult yet important problem in machine learning. In Gaussian processes the main challenge is the definition of valid kernels (covariance functions) able to capture the relationships between different tasks. This paper presents a novel methodology to construct valid multi-task covariance functions (Mercer kernels) for Gaussian processes allowing for a combination of kernels with different forms. The method is based on Fourier analysis and is general for arbitrary stationary covariance functions. Analytical solutions for cross covariance terms between popular forms are provided including Matérn, squared exponential and sparse covariance functions. Experiments are conducted with both artificial and real datasets demonstrating the benefits of the approach.

1 Introduction

Over the past years Gaussian processes (GPs) have become an important tool for machine learning. Initially proposed under the name *kriging* in the geostatistical literature [Cressie, 1993], its formulation as a non-parametric Bayesian regression technique boosted the application of these models to problems beyond spatial stochastic process modeling [MacKay, 1997; Rasmussen and Williams, 2006].

Gaussian process inference is usually formulated for a single output, i.e. given a training set consisting of inputs $\mathbf{x}$ and outputs $\mathbf{y}$ compute the mean and variance of the predictive distribution for a new point $\mathbf{x}^*$. However, in many machine learning problems the objective is to infer multiple tasks jointly, possibly using the dependencies between them to improve results. Real world examples of this problem include ore mining where the objective is to infer the concentration of several chemical components to assess the ore quality. Similarly, in robotics and control problems there are several actuators and the understanding and accurate modeling of the dependencies between the control outputs can significantly improve the controller.

Given its importance for mining applications, the geostatistical community has studied the multiple output case for several years under the name *co-kriging* [Wackernagel, 2003]. The main difficulty in cokriging or multiple output GPs is the definition of a valid covariance function, able to model dependencies between different outputs in the cross terms. Recently, new approaches were proposed where the cross terms are obtained from a convolution process [Boyle and Frean, 2005; Alvarez and Lawrence, 2009]. These approaches use the same covariance function to model different tasks. In many problems however, tasks can have very different behaviors and still be dependent. For example, in mining a chemical component might have a very variable concentration but still be correlated to another component with a smooth concentration in the same area.

In this paper we generalize the multi-task Gaussian process through convolutional processes to allow the use of multiple covariance functions, possibly having a different covariance function per task. We develop a general mathematical framework to build valid cross covariance terms and demonstrate the applicability to real world problems. As examples, we provide closed form solutions to cross covariance terms between Matérn and squared exponential, Matérn and sparse, and sparse and squared exponential covariance functions. The sparse covariance function was recently proposed for exact GP inference in large datasets [Melkumyan and Ramos, 2009]. This property can be naturally incorporated in the multiple output case with the definition of valid cross sparse terms as described in this paper.

The paper is organized as follows: in Section 2 we discuss the previous work. Section 3 reviews multiple output Gaussian process inference and learning. In Section 4 we propose a new methodology for constructing multi-task covariance functions. Section 5 presents analytical solutions for six cross covariance terms and Section 6 demonstrates experimental results. Finally, Section 7 concludes the paper.

2 Related Work

Multi-task learning has received a lot of attention recently. In regression problems, this can be achieved in the linear case with multidimensional linear regression [Hastie *et al.*, 2001]. For non-linear cases, neural networks were employed in [Caruana, 1997] where hidden units represent the sharing knowledge between multiple tasks. In [Evgeniou *et al.*, 2005] correlation between multiple tasks is obtained by specifying a correlated prior over linear regression parameters for support vector machines. The kernels obtained can model linear dependencies aptly but are not suitable for more complex (non-linear) dependencies between tasks.

An interesting multi-task formulation combining Gaussian processes and parametric models was proposed in [Teh *et al.*, 2005]. Each task is modeled with a different covariance function and correlated with a positive definite matrix learned by optimizing a variational lower bound on the marginal likelihood.

In geostatistics the multi-task inference problem has been investigated under the name of *co-kriging* [Wackernagel, 2003]. Different forms of co-kriging have been proposed. Ordinary co-kriging assumes zero mean while simple co-kriging computes an arbitrary mean in the same inference process. In both cases the covariance function is usually assumed to have the same form and parametrization for the different tasks. The linear model of coregionalization proposes the introduction of positive definite matrix multiplying the covariance function similar to the prior model proposed in [Bonilla *et al.*, 2008].

A convolution process between a smoothing kernel and a latent function was used by Boyle and Frean to specify cross covariance terms in [Boyle and Frean, 2005]. Alvarez and Lawrence provide a sparse approximation for that approach in [Alvarez and Lawrence, 2009]. Our model can also be seen as a convolution process of two smoothing kernels (basis functions) assuming the influence of one latent function. Extensions to multiple latent functions are also possible by following the procedure in [Alvarez and Lawrence, 2009].

Our method differs from previous approaches by directly modeling the dependencies between multiple tasks through new cross covariance terms. A mathematical procedure is presented to obtain basis functions for general stationary covariance functions which can be used to construct new cross covariance terms. This is expected to provide much more flexibility in representing complex dependencies between the different outputs. It also generalizes the use of kernels with different forms for the joint multiple output prediction.

3 Multiple Output Gaussian Processes

Consider the supervised learning problem of estimating M tasks $\mathbf{y}^*$ for a query point $\mathbf{x}^*$ given a set X of inputs $\mathbf{x}_{11}, \ldots, \mathbf{x}_{N_1 1}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{N_2 2}, \ldots, \mathbf{x}_{1M}, \ldots, \mathbf{x}_{N_M M}$ and corresponding noisy outputs $\mathbf{y} = (y_{11}, \ldots, y_{N_1 1}, y_{12}, \ldots, y_{N_2 2}, \ldots, y_{1M}, \ldots, y_{N_M M})^T$, where $\mathbf{x}_{il}$ and y_{il} correspond to the ith input and output for task l respectively, and N_l is the number of training examples for task l.

The Gaussian processes approach to this problem is to place a Gaussian prior over the latent functions f_l mapping inputs to outputs. Assuming zero mean for the outputs we define a covariance matrix over all latent functions in order to explore the dependencies between different tasks

$$\operatorname{cov}\left[f_l(\mathbf{x}), f_k(\mathbf{x}')\right] = K_{lk}(\mathbf{x}, \mathbf{x}'), \qquad (1)$$

where K_{lk} with $l, k = 1 : M$ define the positive semi-definite (PSD) block matrix K. In this work we allow K_{lk} to be computed with multiple covariance functions (or kernels) resulting in a final PSD matrix. To fully define the model we need to specify the auto covariance terms k_{lk} with $l = k$ and the cross covariance terms k_{lk} with $l \neq k$. The main difficulty in this problem is to define cross covariance terms that provide PSD matrices. A general framework for this is proposed in the next section.

With these definitions, inference can be computed using the conventional GP equations for mean and variance:

$$\bar{f}_l(\mathbf{x}^*) = \mathbf{k}_l^T K_y^{-1}\mathbf{y}, \quad \mathbb{V}[f_l(\mathbf{x}^*)] = \mathbf{k}_{l*} - \mathbf{k}_l^T K_y^{-1}\mathbf{k}_l, \quad (2)$$

where $K_y = K + \sigma^2 I$ is the covariance matrix for the targets $\mathbf{y}$ and

$$\mathbf{k}_l = [k_{1l}(\mathbf{x}^*, \mathbf{x}_{11}) \ldots k_{1l}(\mathbf{x}^*, \mathbf{x}_{N_1 1}) \ldots$$
$$\ldots k_{Ml}(\mathbf{x}^*, \mathbf{x}_{1M}) \ldots k_{Ml}(\mathbf{x}^*, \mathbf{x}_{N_M M})]^T.$$

Similarly, learning can be performed by maximizing the log marginal likelihood

$$\mathcal{L}(\Theta) = -\frac{1}{2}\mathbf{y}^T K_y^{-1}\mathbf{y} - \frac{1}{2}\log|K_y| - \frac{\log 2\pi}{2}\sum_{i=1}^{M} N_i \quad (3)$$

where Θ is a set of hyper-parameters.

4 Constructing Multi-Kernel Covariance Functions

4.1 General Form

To construct valid cross covariance terms between M covariance functions $k_{11}, k_{22}, \ldots, k_{MM}$ we need to go back to their basis functions and construct cross covariance terms between all the kernel pairs. The proposition below states that if the M covariance functions k_{ii}, $i = 1 : M$ can be written as convolution of their basis functions (or smoothing kernels) g_i, defining the cross covariance terms as the convolutions of g_i with g_j where $i, j = 1 : M$ results in a PSD multi-task covariance function.

Proposition 1. Suppose $k_{ii}\left(x, x'\right)$, $i = 1 : M$ are single-task stationary covariance functions and can be written in the following form:

$$k_{ii}\left(x, x'\right) = \int_{-\infty}^{\infty} g_i\left(x - u\right) g_i\left(x' - u\right) \mathrm{d}u, \; i = 1 : M \quad (4)$$

Then the M task covariance function defined as

$$K\left(x_i, x_j'\right) = \int_{-\infty}^{\infty} g_i\left(x_i - u\right) g_j\left(x_j' - u\right) \mathrm{d}u \qquad (5)$$

where x_i and x_j' belong to the tasks i and j, respectively, is a PSD multi-task covariance function.

This preposition can be proved by considering the quadratic form that it generates. After some algebraic manipulations this quadratic form can be rearranged into a sum of squares which proves that K is a M task PSD covariance function. The details of the proof can be found in the technical report ACFR-TR-2011-002 [1]. The covariance functions k_{ii}, $i = 1 : M$ can have the same form with different hyperparameters or can have completely different forms. When the covariance functions can be written as in Eq. (4) the cross covariance terms can be calculated as in Eq. (5). The main difficulty is finding g_i (smoothing kernel) for popular covariance functions and computing the integrals in Eq. (5). The following section demonstrates how to obtain smoothing kernels for stationary covariance functions through the Fourier analysis.

[1] The technical report ACFR-TR-2011-002 is available at: http://www.acfr.usyd.edu.au/techreports/

4.2 Constructing Cross Covariance Terms with Fourier Analysis

Consider the covariance function $k(\mathbf{x}, \mathbf{x}')$ which can be represented in the form

$$k(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^D} g(\mathbf{x} - \mathbf{u})\, g(\mathbf{u} - \mathbf{x}')\, \mathrm{d}\mathbf{u} \qquad (6)$$

where $g(\mathbf{u}) \equiv g(-\mathbf{u})$:

Changing the variable of integration in Eq. (6) we obtain

$$k(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^D} g(\mathbf{u})\, g(\tau - \mathbf{u})\, d\mathbf{u} = (g * g)(\tau) \qquad (7)$$

where $*$ stands for convolution and $\tau = \mathbf{x} - \mathbf{x}'$.

Applying the Fourier transformation

$$h^*(\mathbf{s}) = \mathcal{F}_{\tau \to \mathbf{s}}[h(\tau)] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^D} h(\tau)\, e^{i\mathbf{s}\cdot\tau} \mathrm{d}\tau$$
$$h(\tau) = \mathcal{F}^{-1}_{\mathbf{s} \to \tau}[h^*(\mathbf{s})] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^D} h^*(\mathbf{s})\, e^{-i\mathbf{s}\cdot\tau} \mathrm{d}\mathbf{s} \qquad (8)$$

to Eq. (7) and using the equality

$$(g_1(\tau) * g_2(\tau))^*(\mathbf{s}) = \sqrt{2\pi} g_1^*(\mathbf{s})\, g_2^*(\mathbf{s})$$

one has that

$$k^*(\mathbf{s}) = \sqrt{2\pi}\, (g^*(\mathbf{s}))^2 . \qquad (9)$$

Using Eqs. (9), (8) one can calculate the basis function $g(\tau)$ of arbitrary stationary covariance function $k(\tau)$ via the formula:

$$g(\tau) = \frac{1}{(2\pi)^{1/4}} \mathcal{F}^{-1}_{\mathbf{s} \to \tau}\left[\sqrt{\mathcal{F}_{\tau \to \mathbf{s}}[k(\tau)]}\right]. \qquad (10)$$

5 Examples

In this section we provide analytical solutions for three cross covariance functions using the framework described above. Proofs can be found in the technical report ACFR-TR-2011-002 referenced above. We have included the sparse covariance function proposed in [Melkumyan and Ramos, 2009] as this provides intrinsically sparse matrices which can be inverted efficiently. The definitions for the squared exponential, Matérn ($\nu = 3/2$ see [Rasmussen and Williams, 2006], p.85) and sparse covariance functions are given below

$$k_{SE}(r; l_{SE}) = \exp\left[-\frac{1}{2}\left(\frac{r}{l_{SE}}\right)^2\right] \qquad (11)$$

$$k_M(r; l_M) = \left(1 + \sqrt{3}r/l_M\right) \exp\left(-\sqrt{3}r/l_M\right) \qquad (12)$$

$$k_S(r; l_S) = \left[\frac{2 + \cos(2\pi r/l_S)}{3}(1 - r/l_S) + \right.$$
$$\left. + \frac{1}{2\pi}\sin(2\pi r/l_S)\right] H(l_S - r) \qquad (13)$$

where l_{SE}, l_M and l_S are the corresponding length scales, $H(x)$ is the Heaviside unit step function and $r = |x - x'|$.

5.1 Squared Exponential × Matern 3/2

Given the squared exponential and Matérn 3/2 covariance functions defined in Eq. (11) and Eq. (12) respectively, the cross covariance term can be analytically calculated via the proposed methodology resulting in:

$$k_{SE\times M}(r; l_{SE}, l_M) = \sqrt{\lambda}\left(\frac{\pi}{2}\right)^{1/4} e^{\lambda^2}\left[2\cosh\left(\frac{\sqrt{3}r}{l_M}\right)\right.$$
$$\left. - e^{\frac{\sqrt{3}r}{l_M}} \operatorname{erf}\left(\lambda + \frac{r}{l_{SE}}\right) - e^{-\frac{\sqrt{3}r}{l_M}} \operatorname{erf}\left(\lambda - \frac{r}{l_{SE}}\right)\right] \qquad (14)$$

where $\lambda = \frac{\sqrt{3}}{2}\frac{l_{SE}}{l_M}$, $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t$.

5.2 Matern 3/2 × Matern 3/2

A combination of two Matérn 3/2 kernels results in the following cross covariance term:

$$k_{M_1\times M_2}(r; l_1, l_2) = \sigma_{12}\left(l_1 e^{-\sqrt{3}\frac{r}{l_1}} - l_2 e^{-\sqrt{3}\frac{r}{l_2}}\right) \qquad (15)$$

where $\sigma_{12} = 2\sqrt{l_1 l_2}/\left(l_1^2 - l_2^2\right)$, and l_1 and l_2 are the length scales of the first and second Matérn 3/2 kernels, respectively.

5.3 Sparse × Sparse

One of the main challenges in multi-task learning with Gaussian processes is the computational burden of inverting a large matrix of size $\sum_{i=1}^M N_i$ where M is the number of tasks, and N_i is the number of points in task i. To tackle this problem in the case of single task GPs, an intrinsically sparse covariance function was proposed in [Melkumyan and Ramos, 2009]. This covariance function has the property of vanishing to zero after a certain distance – represented as a hyperparameter estimated during the learning phase. As the produced covariance matrix is sparse, significant speed ups can be obtained during inversion. Here, we derive an extension of that covariance function for the multi-task case, in combination with itself and other covariance functions. Note that the resulting multi-task sparse×sparse covariance function has compact support.

Combination of two sparse kernels with characteristic length-scales l_1 and l_2 results in

$$k_{S_1\times S_2}(r; l_1, l_2) = \frac{2}{3\sqrt{l_1 l_2}}\left[l_{\min} + \frac{1}{\pi}\frac{l_{\max}^3}{l_{\max}^2 - l_{\min}^2} \times\right.$$
$$\left. \times \sin\left(\pi\frac{l_{\min}}{l_{\max}}\right)\cos\left(\frac{2\pi r}{l_{\max}}\right)\right] \quad \text{if} \quad r \le \frac{|l_2 - l_1|}{2} \qquad (16)$$

$$k_{S_1\times S_2}(r; l_1, l_2) = \frac{2}{3\sqrt{l_1 l_2}}\left[\bar{l} - r + \frac{l_1^3 \sin\left(\pi\frac{l_2 - 2r}{l_1}\right)}{2\pi\left(l_1^2 - l_2^2\right)} - \right.$$
$$\left. - \frac{l_2^3 \sin\left(\pi\frac{l_1 - 2r}{l_2}\right)}{2\pi\left(l_1^2 - l_2^2\right)}\right] \quad \text{if} \quad \frac{|l_2 - l_1|}{2} \le r \le \frac{l_1 + l_2}{2} \qquad (17)$$

and

$$k_{S_1\times S_2}(r; l_1, l_2) = 0 \quad \text{if} \quad r \ge \frac{l_1 + l_2}{2} \qquad (18)$$

where $H(x)$ is the Heaviside unit step function, $\bar{l} = (l_1 + l_2)/2$, $l_{\min} = \min(l_1, l_2)$ and $l_{\max} = \max(l_1, l_2)$.

5.4 Squared Exponential × Sparse

The cross covariance term between squared exponential and sparse kernels can be calculated analytically resulting in

$$k_{SE\times S}\left(r;l_{SE},l_S\right)=\sigma_{SE\times S}\left[\operatorname{erf}\left(\frac{l_S}{2l_{SE}}-\frac{r}{l_{SE}}\right)+\right.$$

$$\left.+\operatorname{erf}\left(\frac{l_S}{2l_{SE}}+\frac{r}{l_{SE}}\right)+e^{-\left(\frac{l_{SE}}{l_S}\pi\right)^2}\operatorname{Re}\left[\gamma\left(r;l_{SE},l_S\right)\right]\right] \quad (19)$$

where Re $[\gamma]$ stands for the real part of γ and

$$\gamma\left(r;l_{SE},l_S\right)=e^{\frac{2\pi r}{l_S}\mathrm{i}}\left[\operatorname{erf}\left(\frac{l_S}{2l_{SE}}-\frac{r}{l_{SE}}-\mathrm{i}\frac{l_{SE}}{l_S}\pi\right)+\right.$$

$$\left.+\operatorname{erf}\left(\frac{l_S}{2l_{SE}}+\frac{r}{l_{SE}}+\mathrm{i}\frac{l_{SE}}{l_S}\pi\right)\right],$$

$$\sigma_{SE\times S}=\left(2\pi\right)^{1/4}\sqrt{l_{SE}/(6l_S)}.$$

5.5 Matern 3/2 × Sparse

The cross covariance term between Matérn 3/2 and sparse kernels can be calculated analytically resulting in

$$k_{M\times S}\left(r;l_M,l_S\right)=\sigma_{M\times S}\exp\left(-\sqrt{3}r/l_M\right) \quad (20)$$

where

$$\sigma_{M\times S}=\frac{2\pi^2}{\pi^2+\lambda_{M\times S}^2}\frac{\sinh\lambda_{M\times S}}{\sqrt{3\lambda_{M\times S}}},\quad \lambda_{M\times S}=\frac{\sqrt{3}}{2}\frac{l_S}{l_M},$$

and l_M and l_S are the length scales of the corresponding covariance functions.

5.6 Squared Exponential × Squared Exponential

The cross covariance term between two squared exponential kernels has an analytical form given by

$$k_{SE_1\times SE_2}\left(r;l_1,l_2\right)=\sqrt{\frac{2l_1l_2}{l_1^2+l_2^2}}\exp\left(-\frac{r^2}{l_1^2+l_2^2}\right). \quad (21)$$

For the general anisotropic case:

$$k_{SE_1}\left(\mathbf{x},\mathbf{x}';\Omega_1\right)=\exp\left[-\frac{\left(\mathbf{x}-\mathbf{x}'\right)^T\Omega_1^{-1}\left(\mathbf{x}-\mathbf{x}'\right)}{2}\right] \quad (22)$$

$$k_{SE_2}\left(\mathbf{x},\mathbf{x}';\Omega_2\right)=\exp\left[-\frac{\left(\mathbf{x}-\mathbf{x}'\right)^T\Omega_2^{-1}\left(\mathbf{x}-\mathbf{x}'\right)}{2}\right] \quad (23)$$

$$k_{SE_1\times SE_2}\left(\mathbf{x},\mathbf{x}';\Omega_1,\Omega_2\right)=2^{D/2}\frac{\left|\Omega_1\right|^{1/4}\left|\Omega_2\right|^{1/4}}{\sqrt{\left|\Omega_1+\Omega_2\right|}}\times$$

$$\times\exp\left[-\left(\mathbf{x}-\mathbf{x}'\right)^T\left(\Omega_1+\Omega_2\right)^{-1}\left(\mathbf{x}-\mathbf{x}'\right)\right] \quad (24)$$

Multidimensional and anisotropic extensions of the other models are possible by taking the product of the cross covariance terms defined for each input dimension.

Note that the examples above do not consider parameters for the amplitude (signal variance) of the covariance functions. This, however, can be added by multiplying blocks of the multi-task covariance matrix by coefficients from a PSD matrix as in [Bonilla *et al.*, 2008].

6 Experiments

6.1 1D Simulation

The first experiment demonstrates the benefits of using the multi-kernel methodology in an artificial 1D problem for two dependent tasks. The observations for the first task are generated from a minus sine function corrupted with Gaussian noise. Only the observations for the second part of the function are used and the objective is to infer the first part from observations of the second task. Observations for the second task were generated from a sine function with some additional complexity to make the function less smooth and corrupted by Gaussian noise. A comparison between independent GP predictions, multi-task GP with squared exponential kernel for both tasks, and the multi-kernel GP (squared exponential kernel for the first task and Matérn 3/2 for the second) is presented in Figure 1. It can be observed in Figure 1(c) that the multi-kernel GP models the second function more accurately. This helps in providing a better prediction for the first task.

Despite the simplicity of this experiment it simulates a very common phenomenon in grade estimation for mining. Some elements have a much higher concentration variability but follow the same trend as others. Being able to aptly model these dependencies from noisy x-ray lab samples is essential for an accurate final product. This is empirically demonstrated in the second experiment.

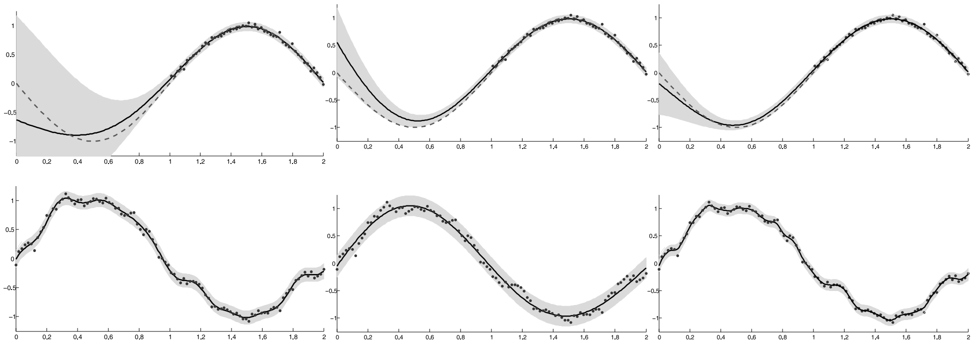
6.2 Iron Ore

1363 samples from an iron ore mine were collected and analyzed in laboratory with x-ray instruments to determine the concentration of three components: iron, silica and alumina. Iron is the main product but equally important is to asses the concentration of the contaminants silica and alumina. The samples were collected from exploration holes of about 200m deep, distributed in an area of 6 km^2. Each hole is divided into 2 meter sections for laboratory assessment, the lab result for each section is then an observation in our dataset. The final dataset consists of 4089 data points representing 31 exploration holes. We separate two holes to use as testing data. For these holes we predict the concentration of silica given iron and alumina. The experiment is repeated employing different multi-task covariance functions with either squared exponential or Matérn kernel for each task combined with the cross-covariance terms presented in Section 5. The results are summarized in Table 1 which demonstrates that the dependencies between iron, silica and alumina are better captured by the Matérn 3/2 × Matérn 3/2 × SqExp multi-kernel covariance function.

6.3 Jura

In the third experiment GPs with different multi-kernel covariance functions were applied to the Jura dataset. The Jura dataset is a benchmark dataset in geostatistics[2]. It consists of a training set with 259 samples in an area of 14.5km^2 and a testing set with 100 samples. The task is to predict the concentration of cadmium (Cd), lead (Pb) and zinc (Zn) at

[2] The Jura dataset is available at:
http://goovaerts.pierre.googlepages.com/



(a) Independent GPs with SqExp and SqExp (b) Multi-task GPs with SqExp and SqExp (c) Multi-Kernel GPs with SqExp (top) and Mat3/2 (bottom)

Figure 1: Predictive mean and variance for independent GPs, multi-task GPs and multi-kernel GPs. SqExp indicates the squared exponential covariance functions while Mat3/2 indicates the Matérn 3/2 covariance function. The dots represent the observations and the dashed (red) line represents the ground truth for task 1.

Kernel for Fe	Kernel for SiO_2	Kernel for Al_2O_3	Absolute Error
SqExp	SqExp	SqExp	2.7995±2.5561
Matérn 3/2	**Matérn 3/2**	**SqExp**	**2.2293±2.1041**
Matérn 3/2	SqExp	Matérn 3/2	2.8393±2.6962
SqExp	Matérn 3/2	Matérn 3/2	3.0569±2.9340
Matérn 3/2	Matérn 3/2	Matérn 3/2	2.6181±2.3871

Table 1: Mean and standard deviation of absolute error for iron grade

new locations. The multi-kernel covariance functions proposed in this paper enable considering different kernels for each of the materials thus maximizing the predictive qualities of the GP. The 259 training samples were used at the learning stage and the 100 testing samples were used to evaluate the predictive qualities of the models. The square root mean square error (SMSE) for all possible triplet combinations of SqExp and Matérn 3/2 kernels are presented in Table 2. The results demonstrate that the dependencies between cadmium, lead and zinc are better captured by the Matérn 3/2 × SqExp × SqExp triplet-kernel.

6.4 Concrete Slump

In the last experiment the concrete slump dataset[3] is considered. This dataset contains 103 data points with seven input dimensions and 3 outputs describing the influence of the constituent parts of concrete on the overall properties of the concrete. The seven input dimensions are cement, slag, fly ash, water, SP, coarse aggr. and fine aggr., and the outputs are slump, flow and 28-day compressive strength of concrete. 83 data points are used for learning and 20 data points are used for testing. The square root mean square error (SMSE) for all possible triplet combinations of SqExp and Matérn 3/2 kernels for this dataset are presented in Table 3. The results demonstrate that the dependencies between slump, flow and 28-day compressive strength of concrete are better captured by the SqExp × Matérn 3/2 × Matérn 3/2 triplet-kernel.

7 Conclusions

This paper presented a novel methodology to construct cross covariance terms for multi-task Gaussian process. This methodology allows the use of multiple covariance functions for the same multi-task prediction problem. We prove that if a stationary covariance function can be written as a convolution of two identical basis functions, a cross covariance term can always be defined resulting in a positive semi-definite multi-task covariance matrix. A general methodology to find the basis function is then developed based on Fourier analysis.

We provide analytical solutions for six combinations of covariance functions, three of them combining different covariance functions. The analytical forms for the cross covariance terms can be directly applied to GP prediction problems but are useful for other kernel machines.

We presented a multi-task sparse covariance function which provides computationally efficient (and exact) way of performing inference in large datasets [Melkumyan and Ramos, 2009]. Note however that approximate techniques such as [Williams and Seeger, 2001; Lawrence *et al.*, 2003; Snelson and Ghahramani, 2006] can also be used.

As a future work we plan to extend the approach to non-stationary covariance functions, possibly combining non-stationary and stationary kernels. This can be useful in applications involving space and time domains such as pollution estimation or weather forecast.

[3] The concrete slump dataset is available at: http://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test

Kernel for Cd	Kernel for Pb	Kernel for Zn	SMSE for Cd	SMSE for Pb	SMSE for Zn
SqExp	SqExp	SqExp	1.0231	13.7199	42.4945
Matérn 3/2	Matérn 3/2	Matérn 3/2	0.9456	11.9542	38.7402
Matérn 3/2	Matérn 3/2	SqExp	0.9079	11.4786	42.1452
Matérn 3/2	SqExp	Matérn 3/2	0.8239	9.7757	36.2846
SqExp	Matérn 3/2	Matérn 3/2	1.0375	12.4937	39.6459
SqExp	SqExp	Matérn 3/2	0.8214	9.9625	37.8670
SqExp	Matérn 3/2	SqExp	1.0269	12.087	42.6403
Matérn 3/2	**SqExp**	**SqExp**	**0.7883**	**9.7403**	**34.4978**

Table 2: Square root mean square error for cadmium (Cd), lead (Pb) and zinc (Zn) for all possible triplet-kernels combining SqExp and Matérn 3/2

Kernel for Slump	Kernel for Flow	Kernel for Strength	SMSE for Slump	SMSE for Flow	SMSE for Strength
SqExp	SqExp	SqExp	13.8776	820.4181	733.1642
Matérn 3/2	Matérn 3/2	Matérn 3/2	13.6224	820.6727	733.5744
Matérn 3/2	Matérn 3/2	SqExp	14.7709	821.8064	733.0741
Matérn 3/2	SqExp	Matérn 3/2	14.2670	822.7529	733.5768
SqExp	**Matérn 3/2**	**Matérn 3/2**	**13.5690**	**820.3678**	**732.7032**
SqExp	SqExp	Matérn 3/2	15.3459	821.1577	733.6685
SqExp	Matérn 3/2	SqExp	16.2332	824.4468	733.7083
Matérn 3/2	SqExp	SqExp	13.7503	845.5608	741.3144

Table 3: Square root mean square error for slump, flow and strength of concrete for all possible triplet-kernels combining SqExp and Matérn 3/2

Acknowledgements

This work was supported by the Rio Tinto Centre for Mine Automation and the Australian Centre for Field Robotics.

References

[Alvarez and Lawrence, 2009] M. Alvarez and N. D. Lawrence. Sparse convolved gaussian processes for multi-output regression. In D. Koller, Y. Bengio, D. Schuurmans, and L. Bottou, editors, *NIPS*. MIT Press, 2009.

[Bonilla *et al.*, 2008] E. V. Bonilla, K. M. Chai, and C. K. I. Williams. Multi-task gaussian process prediction. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NIPS*, pages 153–160. MIT Press, 2008.

[Boyle and Frean, 2005] P. Boyle and M. Frean. Dependent gaussian processes. In L. Saul, Y. Weiss, and L. Bouttou, editors, *NIPS*, volume 17, pages 217–224. MIT Press, 2005.

[Caruana, 1997] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[Cressie, 1993] N. Cressie. *Statistics for Spatial Data*. Wiley, 1993.

[Evgeniou *et al.*, 2005] T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:515–537, 2005.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[Lawrence *et al.*, 2003] N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse gaussian process methods: The information vector machine. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 625–632. MIT Press, 2003.

[MacKay, 1997] D. MacKay. Gaussian processes: A replacement for supervised neural networks? In *NIPS97 Tutorial*, 1997.

[Melkumyan and Ramos, 2009] A. Melkumyan and F. Ramos. A sparse covariance function for exact gaussian process inference in large datasets. In *The 21st IJCAI*, 2009.

[Rasmussen and Williams, 2006] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[Snelson and Ghahramani, 2006] E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT press, 2006.

[Teh *et al.*, 2005] Y. W. Teh, M. Seeger, and M. I. Jordan. Semiparametric latent factor models. In R. G. Cowell and Z. Ghahramani, editors, *AISTATS 10*, pages 333–340, 2005.

[Wackernagel, 2003] H. Wackernagel. *Multivariate Geostatistics*. Springer, 2003.

[Williams and Seeger, 2001] C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.