
Multi-Kernel Gaussian Processes

Arman Melkumyan **Fabio Ramos**
Australian Centre for Field Robotics
The University of Sydney
Sydney, NSW 2006, Australia
{a.melkumyan, f.ramos}@acfr.usyd.edu.au

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 [6, 4], its formulation as a non-parametric Bayesian regression technique boosted the application of these models to problems beyond spatial stochastic process modeling [5, 9].

Although Gaussian process inference is usually formulated for a single output, in many machine learning problems the objective is to infer multiple tasks jointly, possibly exploring 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 more than one actuator and the understanding and accurate modeling of the dependencies between the control outputs can significantly improve the controller.

The main challenge for multi-task Gaussian processes is to define valid cross-covariance functions that are both positive semi-definite and informative [4]. In this paper we generalize the multi-task Gaussian process 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 Sparse, and two different Sparse covariance functions. The Sparse covariance function was recently proposed for exact GP inference in large datasets [8]. This property of sparsity can be naturally incorporated in the multiple output case with the definition of valid cross sparse terms as described in this paper.

Our model can be seen as a mathematical procedure to derive multi-kernel covariance functions for the convolution process of two smoothing kernels (basis functions), assuming the influence of one latent function [3]. It can also incorporate the extensions proposed in [1] where multiple latent functions influence the output.

2 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 $\mathbf{x}_{11}, \dots, \mathbf{x}_{N_1 1}, \mathbf{x}_{12}, \dots, \mathbf{x}_{N_2 2}, \dots, \mathbf{x}_{1M}, \dots, \mathbf{x}_{N_M M}$ and corresponding noisy outputs $\mathbf{y} = (y_{11}, \dots, y_{N_1 1}, y_{12}, \dots, y_{N_2 2}, \dots, y_{1M}, \dots, y_{N_M M})^T$, where \mathbf{x}_{il} and y_{il} correspond to the i th 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 l [2]. Assuming zero mean for the outputs we define a covariance matrix over all latent functions in order to explore the dependencies between different tasks

$$\text{cov} [f_l(\mathbf{x}), f_k(\mathbf{x}')] = K_{lk}(\mathbf{x}, \mathbf{x}'), \quad (1)$$

where K_{lk} is a positive semi-definite (PSD) matrix. 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.

3 Constructing Multi-Kernel Covariance Functions

3.1 General Form

To construct valid cross covariance terms between M covariance functions $k_{11}(x, x')$, $k_{22}(x, x')$, ..., $k_{MM}(x, x')$ we need to go back to their basis functions. The proposition below states that if the M covariance functions $k_{ii}(x, x')$, $i = 1 : M$ can be written as convolution of their basis functions 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 $k_{ii}(x, x')$, $i = 1 : M$ are single-task stationary covariance functions and can be written in the following form:*

$$k_{ii}(x, x') = \int_{-\infty}^{\infty} g_i(x-u) g_i(x'-u) du, \quad i = 1 : M. \quad (2)$$

The M task covariance function defined as

$$K(x, x') = \int_{-\infty}^{\infty} g_i(x-u) g_j(x'-u) du \quad \text{for } x \in T_i, x' \in T_j, i, j = 1 : M \quad (3)$$

where T_i and T_j stand for tasks i and j , respectively, is a PSD multi-task covariance function.

Proof. For any M set of points $X_i = (x_{1i}, x_{2i}, \dots, x_{N_i i})$ and arbitrary real numbers $A_i = (a_{1i}, a_{2i}, \dots, a_{N_i i})$, $i = 1 : M$, via basic algebraic manipulations it can be shown that the quadratic form generated by the function K is PSD which proves that K is a M task PSD covariance function. ■

The covariance functions k_{ii} , $i = 1 : M$ can have the same form with different hyper-parameters or can have completely different forms. When the covariance functions can be written as in Eq. (2) the cross covariance terms can be calculated as in Eq. (3). The main difficulty is finding g_i for popular covariance functions. The following section demonstrates how this can be performed for stationary covariance functions through the Fourier analysis.

3.2 Constructing Cross Covariance Terms with Fourier Analysis

If $k(\boldsymbol{\tau})$ is a stationary covariance function in \mathbb{R}^D with a spectral density $S(\mathbf{s})$, then $k(\boldsymbol{\tau})$ and $S(\mathbf{s})$ are Fourier duals of each other (Wiener-Khinchine theorem), i.e.

$$k(\boldsymbol{\tau}) = \mathbf{F}_{\mathbf{s} \rightarrow \boldsymbol{\tau}}^{-1} [S(\mathbf{s})] (\boldsymbol{\tau}), \quad S(\mathbf{s}) = \mathbf{F}_{\boldsymbol{\tau} \rightarrow \mathbf{s}} [k(\boldsymbol{\tau})] (\mathbf{s}), \quad \boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'. \quad (4)$$

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

$$k(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^D} g(\mathbf{x} - \mathbf{u}) g(\mathbf{u} - \mathbf{x}') du \quad (5)$$

where $g(\mathbf{u}) \equiv g(-\mathbf{u})$ and changing the variable of integration we obtain

$$k(\mathbf{x}, \mathbf{x}') = (g * g)(\boldsymbol{\tau}) \quad (6)$$

where $*$ stands for convolution.

Applying the Fourier transformation to Eq. (6) and using the well-known equality

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

where $*$ in superscript stands for the Fourier transformation, one has that

$$k^*(\mathbf{s}) = \sqrt{2\pi} (g^*(\mathbf{s}))^2. \quad (7)$$

Using Eq. (7) one can calculate the basis function using the covariance function as follows:

$$g(\boldsymbol{\tau}) = \frac{1}{(2\pi)^{1/4}} \mathbf{F}_{\mathbf{s} \rightarrow \boldsymbol{\tau}}^{-1} \left[\sqrt{\mathbf{F}_{\boldsymbol{\tau} \rightarrow \mathbf{s}} [k(\boldsymbol{\tau})]} \right]. \quad (8)$$

4 Examples

Using the framework described above multiple single-task covariance functions can be integrated through a multi-task covariance function. In this section we provide analytical solutions for combining the Matérn ($\nu = 3/2$ see [9], p.85) covariance function with the Sparse covariance function [8] and for combining two Sparse covariance functions with different characteristic length-scales. We have included the Sparse covariance function as this provides an elegant solution to handle large datasets. The idea is to produce intrinsically sparse matrices which can be inverted efficiently.

For example, the analytical calculations of the cross covariance term between the Matérn 3/2 and Sparse covariance functions results in

$$k_{M \times S}(r; l_M, l_S) = \frac{8\sqrt{2}}{3^{3/4}} \sqrt{\frac{l_M}{l_S}} \frac{\pi^2 l_M^2}{4\pi^2 l_M^2 + 3l_S^2} \sinh\left(\frac{\sqrt{3}}{2} \frac{l_S}{l_M} r\right) \exp\left(-\sqrt{3} \frac{r}{l_M}\right), \quad (9)$$

where l_M and l_S are the length scales for the Matérn 3/2 and Sparse covariance functions respectively, and $r = |x - x'|$.

A combination of two Sparse kernels with characteristic length-scales l_1 and l_2 results in

$$k_{S_1 \times S_2}(r; l_1, l_2) = \frac{2}{3\sqrt{l_1 l_2}} \begin{cases} l_{\min} + \frac{1}{\pi} \frac{l_{\max}^3}{l_{\max}^2 - l_{\min}^2} \sin\left(\pi \frac{l_{\min}}{l_{\max}}\right) \cos\left(\frac{2\pi r}{l_{\max}}\right) & \text{if } r \leq \frac{|l_2 - l_1|}{2} \\ \frac{l_1 + l_2}{2} - r + \frac{1}{2\pi(l_1^2 - l_2^2)} \left[l_1^3 \sin\left(\pi \frac{l_2 - 2r}{l_1}\right) - l_2^3 \sin\left(\pi \frac{l_1 - 2r}{l_2}\right) \right] & \text{if } \frac{|l_2 - l_1|}{2} \leq r \leq \frac{l_1 + l_2}{2} \\ 0 & \text{if } \frac{l_1 + l_2}{2} \leq r \end{cases} \quad (10)$$

where $H(x)$ is the Heaviside unit step function, $l_{\min} = \min(l_1, l_2)$, $l_{\max} = \max(l_1, l_2)$, and $r = |x - x'|$.

Multidimensional and anisotropic extensions of these covariance functions can be constructed by taking the product of the cross covariance terms defined for each input dimension and applying linear transformations as in [7].

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 [2]. We have also derived analytical solutions for the combination of square exponential with sparse, square exponential with Matérn and two different Matérn covariance functions.

5 Experiments

For this experiment 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 assess 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 km². 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 analytically derived corresponding cross-covariance terms. The results are summarized in Table 1 which demonstrates that the dependences between iron, silica and alumina are better captured by the Matérn 3/2 \times Matérn 3/2 \times SqExp multi-kernel covariance function.

6 Discussion

This paper presented a novel methodology for constructing cross covariance terms for multi-task Gaussian processes. This methodology allows the use of multiple covariance functions for the same multi-task prediction problem. A general methodology for calculating the basis functions of stationary covariance functions using the techniques of Fourier analysis is proposed. A general method-

Kernel for Fe	Kernel for SiO ₂	Kernel for Al ₂ O ₃	Absolute Error
SqExp	SqExp	SqExp	2.7995±2.5561
Matern 3/2	Matern 3/2	SqExp	2.2293±2.1041
Matern 3/2	SqExp	Matern 3/2	2.8393±2.6962
SqExp	Matern 3/2	Matern 3/2	3.0569±2.9340
Matern 3/2	Matern 3/2	Matern 3/2	2.6181±2.3871

Table 1: Mean and standard deviation of absolute error

ology for constructing cross covariance terms using these basis functions is also proposed and the resulting multi-task covariance function is proved to be positive semi-definite.

As an example we provided analytical solutions for cross covariance functions combining the Matérn and Sparse as well as two Sparse covariance functions. The presented multi-task Sparse covariance function provides computationally efficient (and exact) way of performing inference in large datasets [8]. Note however that approximate techniques such as [10, 12] can also be used.

The presented analytical forms for the cross covariance terms can be directly applied to multi-task GP prediction problems and are also useful for other kernel machines.

Acknowledgments

This work has been supported by the Rio Tinto Centre for Mine Automation and the ARC Centre of Excellence programme, funded by the Australian Research Council (ARC) and the New South Wales State Government.

References

- [1] 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.
- [2] 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.
- [3] P. Boyle and M. Frean. Dependent gaussian processes. In L. Saul, Y. Weiss, and L. Bottou, editors, *NIPS*, volume 17, pages 217–224. MIT Press, 2005.
- [4] N. Cressie. *Statistics for Spatial Data*. Wiley, 1993.
- [5] D. MacKay. Gaussian processes: A replacement for supervised neural networks? In *NIPS97 Tutorial*, 1997.
- [6] G. Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963.
- [7] A. Melkumyan and E. Nettleton. An observation angle dependent nonstationary covariance function for Gaussian process regression. In *16th International Conference on Neural Information Processing*, 2009.
- [8] A. Melkumyan and F. Ramos. A sparse covariance function for exact gaussian process inference in large datasets. In *The 21st International Joint Conference on Artificial Intelligence*, 2009.
- [9] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [10] 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.
- [11] 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.
- [12] 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.