



Proposal/Contract no.: 033811  
Project start: September 1, 2006  
Project end: August 31, 2009



# INTAMAP

*Interoperability and Automated Mapping*

SIXTH FRAMEWORK PROGRAMME

PRIORITY IST-2005-2.5.12

ICT for Environmental Risk Management

## Deliverable 2.3

Improved methods for hyper-parameter identification in the sparse, sequential framework

Title of Deliverable	Improved methods for hyper-parameter identification in the sparse, sequential framework
Deliverable reference number	INTAMAP D2.3
Related WP and Tasks	WP2, Task 2.4
Type of Document	Internal deliverable, Public
Authors	AST
Date	8 October 2008
Version	1.1

**Project coordinator**

Dr. Edzer J. Pebesma

Utrecht University, The Netherlands

E-mail: [e.pebesma@geo.uu.nl](mailto:e.pebesma@geo.uu.nl)

<http://www.intamap.org/>

## Revision History

Version	Date	Changes	Authors
1.0	6-10-2008	First draft	Ben Ingram
1.1	8-10-2008	Reviewed and ammended for submission	Dan Cornford

## Related task

### Task 2.4 Incorporating the CTI method in the Bayesian framework of spatial analysis

This task will focus on integrating the anisotropy analysis with the Bayesian framework of spatial analysis. In particular, the sparse sequential approach to geostatistics (Cornford et al, 2005), covariance parameters, are estimated from the data using a maximum likelihood type II estimate, which is computed in a computationally efficient manner, and can be used even in the case of very large data sets. However recent experience in the SIC2004 exercise has shown that reliable estimation of such hyperparameters requires further research. In this task we will integrate the work done in TUC with the work at AST to improve the reliability of both methods.

Active partners: AST, TUC

### **Legal Notices**

The information in this document is subject to change without notice. The Members of the INTAMAP Consortium make no warranty of any kind with regard to this document, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The Members of the INTAMAP Consortium shall not be held liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.

# Contents

1	Introduction . . . . .	2
2	Methodology . . . . .	3
	2.1 Improving stability of parameter estimation . . . . .	4
	2.2 Experiments . . . . .	5
3	Results . . . . .	5
	3.1 Simulated dataset . . . . .	5
	3.2 Radiation monitoring scenario . . . . .	7
4	Conclusions . . . . .	8

## **Executive Summary**

This internal deliverable shows how we can combine the developments in D2.2 with the developments in D3.3 and D4.4 to improve the stability and efficiency of the sparse sequential Gaussian process method developed at Aston. The particular contribution of this report is to demonstrate the improved stability of the sparse sequential method on both simulated and real data from WP5. The primary conclusion is that the performance of the sparse sequential method can be significantly speeded up, while also increasing the accuracy of the resulting inference. This means that the method can be employed in a more robust manner in the INTAMAP processing chain, something that is rather important in emergency mapping scenarios. The other significant benefit is a further speed up of the algorithm which means potentially huge data sets can now be treated in near real-time.

# 1 Introduction

Determining the optimal model parameters given a dataset using the sparse, sequential framework (SSGP) introduced in Deliverables 3.3 & 4.4 poses a number of potential problems. An approximation to the marginal log-likelihood is used during parameter estimation. This approximate marginal likelihood function is represented by a subset of the available data (which we call active points) and these are selected based on how informative they are about the Gaussian process posterior distribution. During SSGP parameter estimation two distinct processes take place: selecting active points while estimating the posterior distribution and estimating covariance parameters. Selecting the active points depends on the covariance parameters and selecting the covariance parameters depends on the active points. If initial estimates for the covariance parameters are poor then the SSGP algorithm may need to run for a number of iterations until the parameters convergence to reasonable estimates.

In this deliverable we build on the earlier work of Csató and Opper [2002] by introducing a stable and reliable method for determining covariance parameters. We use the methods developed in Deliverable 2.2 to identify anisotropy in the dataset and correct this as part of a preprocessing stage before we process the data using the SSGP algorithm. To determine an estimate for the correlation range of the process we use a method-of-moments estimator which can be computed in a very short time. An estimate for the process correlation range enables the active points to be selected more reliably during the first iteration of the algorithm.

## 2 Methodology

As with previous work, the SSGP method assumes that any finite collection of random variables is jointly Gaussian. We assume that the data is of the form:

$$(\mathbf{x}_i, y_i) : i = 1, \dots, n, \quad (1)$$

where  $\mathbf{x}_i$  refers to a spatial location and  $y_i$  is an observation at the location  $\mathbf{x}_i$ . Each observation,  $y_i$ , is assumed to be a realisation of a random variable  $Y_i$  which is dependent on the value of an unobserved random process.

As detailed in Deliverables 3.3 & 4.4 we adopt a Bayesian framework for the iterative algorithm. The aim is to infer the posterior distribution of the underlying random process  $S(\mathbf{x})$  given the observed data,  $Y = \{y_i\}_{i=1..n}$ . This has the standard form:

$$p(S(\mathbf{x})|Y, \theta) = \frac{[\prod_i p(y_i|S(\mathbf{x}))] p(S(\mathbf{x})|\theta)}{p(Y|\theta)} \quad (2)$$

where the posterior is the product of the (potentially arbitrary) likelihood terms and the Gaussian process prior term, divided by a normalising constant, often called the marginal likelihood defined as:

$$p(Y|\theta) = \int \left[ \prod_i p(y_i|S(\mathbf{x})) \right] p(S(\mathbf{x})|\theta) dS(\mathbf{x}). \quad (3)$$

It is this normalising constant or *marginal likelihood* that concerns us in this deliverable. The term marginal likelihood is a reference to the marginalisation over the latent process  $S(\mathbf{x})$ . Although the SSGP framework treats datasets with non-Gaussian likelihood functions, the posterior process (and hence prior distribution for each iteration) is always Gaussian (due to a projection onto the best approximating Gaussian as specified in D4.4). In the standard Gaussian process case (i.e. simple kriging), assuming that the prior is a Gaussian process and that the likelihood is a factorized Gaussian<sup>1</sup>, identities for products of Gaussians can be used to give the log marginal likelihood:

$$\log p(Y|\theta) = -\frac{1}{2} \mathbf{Y}^T \Sigma \mathbf{Y} - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log 2\pi. \quad (4)$$

In previous Deliverables we showed how the posterior moments of the Gaussian process were parameterised. The posterior mean is parameterised as:

$$\mu_{posterior}(\mathbf{x}) = \mu_{prior}(\mathbf{x}) + \sum_i^m \alpha_i c(\mathbf{x}, \mathbf{x}_i), \quad (5)$$

where  $c(\mathbf{x}, \mathbf{x}_i)$  is the covariance function between the location  $\mathbf{x}$  and the active point location  $\mathbf{x}_i$  used in the approximation. We refer to  $m$  as the active set size, but in this instance  $m = n$ . Later in this report when sparsity is introduced into the model,  $m$  can be reduced so that  $m \ll n$ . We write the covariance between two spatial locations as  $c(\mathbf{x}, \mathbf{x}_i) = \sigma^2 \rho(\mathbf{x}, \mathbf{x}_i) = \text{cov}(\mathbf{x}, \mathbf{x}_i)$ . We assume that the process variance,  $\sigma^2$ , and the parameters of the correlation function,  $\rho(\cdot)$  are known at this stage.  $\boldsymbol{\alpha}$  are the parameters of the posterior mean of the

<sup>1</sup>In the SSGP method a similar factorisation arises from the projection of the non-Gaussian posterior to the optimal Gaussian posterior and is central to the expectation propagation algorithm employed in practice.

process. The posterior variance is parameterised as:

$$c_{posterior}(\mathbf{x}, \mathbf{x}') = c_{prior}(\mathbf{x}, \mathbf{x}') + \sum_{i,j=1}^m c(\mathbf{x}, \mathbf{x}_i) \mathbf{C}(ij) c(\mathbf{x}_j, \mathbf{x}') \quad (6)$$

where  $\mathbf{C}$  is a matrix of parameters and is used to represent the posterior process variance.

The mechanics of updating these parameters  $\boldsymbol{\alpha}$  and  $\mathbf{C}$  is avoided in this report. Full details can be found in Deliverables 3.3 & 4.4. As can be seen from Equations (5) & (6), the model parameters,  $\boldsymbol{\alpha}$  and  $\mathbf{C}$ , depend on the covariance structure of the model. Using this parameterisation, Csató and Opper [2002] show that the log marginal likelihood is given by:

$$\log p(Y|\theta) = -\frac{1}{2} \mathbf{Y}^T \boldsymbol{\alpha} - \frac{1}{2} \log |C| - \frac{n}{2} \log 2\pi. \quad (7)$$

We maximise the log marginal likelihood with respect to the covariance parameters. Typically a gradient-based local optimisation algorithm is used, but if computation time is not an issue then slower global optimisation algorithms such as simulated annealing can be employed. The derivative with respect to  $\theta$  of Equation (7) can be simplified and written as:

$$\frac{\partial}{\partial \theta} \log p(\mathbf{Y}|\theta) = \frac{1}{2} \text{trace} \left( (\boldsymbol{\alpha} \boldsymbol{\alpha}^T - \mathbf{C}^{-1}) \frac{\partial \Sigma}{\partial \theta} \right) \quad (8)$$

where the derivative of the covariance function with respect to a parameter in the model is denoted by  $\frac{\partial \Sigma}{\partial \theta}$ .

When local optimisation algorithms are used, it is important that the initial values for the parameters are set. In the SSGP framework, these initial parameters will largely determine the active point selection process for the first iteration. If the initial parameters are a very bad estimate, convergence to reasonable estimates could take a large number of iterations of the algorithm. This can be seen by two extreme examples: imagine the length scales are estimated too short - this means that all points will be incorporated into the posterior distribution (if this is permitted) and this will slow down the algorithm considerably on the first step, although on subsequent steps it is likely to recover well. The alternative is that the length scales are set too long, and a very small number of active points are chosen. These could not capture the small scale features and thus the marginal likelihood based estimates would either converge slowly or more seriously not at all.

## 2.1 Improving stability of parameter estimation

As already noted, parameter estimation within the SSGP framework raises potential problems, particularly in the case of selecting starting parameters. Here we apply the developments of Deliverable 2.2 to the SSGP framework. We use the anisotropy corrections determined in D2.2 to rotate and rescale the data so that isotropy is restored allowing a simple isotropic covariance model to be estimated. Additionally we use a method-of-moments variogram estimator to determine initial estimates for the covariance parameters.

Determining the covariance parameters, particularly the correlation range parameter decreases the number of iterations of the SSGP algorithm. A *good* range parameter estimate leads to the active point locations being more appropriate for estimating the given model. Should the active point locations separation distances be too large, then the likelihood ap-

proximation will be very poor. If the active point location separation distances is too small, then the likelihood approximation will be very slow to compute, or if the active set size is fixed, then the likelihood approximation will be also very poor.

In this report we change the SSGP algorithm to ensure stable and reliable estimation of the covariance parameters. Firstly, we use the anisotropy correction code to transform the data. We then compute a method-of-moments variogram to determine the correlation range. Using this estimate of the correlation range, we select the active point locations. Using these active point locations we perform maximum marginal likelihood parameter estimation to determine new estimates for the covariance parameters. With these updated parameters, new active point locations can be determined. These are then fixed for the rest of the algorithm. Then 2-3 iterations of the expectation propagation algorithm are performed to produce a final posterior estimate.

The approximation of the log marginal likelihood depends not only on the data and covariance parameters, but also on the active set. As a consequence, the marginal likelihood value varies depending on the active set selected. In the existing SSGP parameter estimation, a rare, but occasionally observed issue would be the log marginal likelihood function fluctuating slightly. This behaviour is due to the active set changing during each iteration of the algorithm. By fixing the active set after the first iteration, this problematic issue disappears.

## 2.2 Experiments

To demonstrate the improvements in computational efficiency, we consider two datasets. First we simulation a random dataset and then we look at a radiation monitoring scenario across Europe for WP5. The aim of these experiments is to gauge the performance of the SSGP algorithm in terms of computation time and parameter estimation accuracy. We consider the three configurations of the SSGP algorithm:

- default SSGP
- SSGP with starting values initialised by method-of-moments variogram estimator
- SSGP with anisotropy correction and method-of-moment starting value estimation

Two outcomes are expected. Firstly, we expect that the number of iterations of the algorithm should be reduced with our extended algorithm. This is because for the first iteration of the algorithm, the correlation range parameters are estimated well leading to the active set selection being improved. The second aspect that we expect to improve is that if the anisotropy parameters have been determined correctly, then the number of active points needed to represent the model should be reduced.

## 3 Results

### 3.1 Simulated dataset

To simulate a dataset, we used the Cholesky decomposition method [Cressie, 1993] on a Gaussian process with a covariance function with a N-S range of 4 and a E-W range of 1.

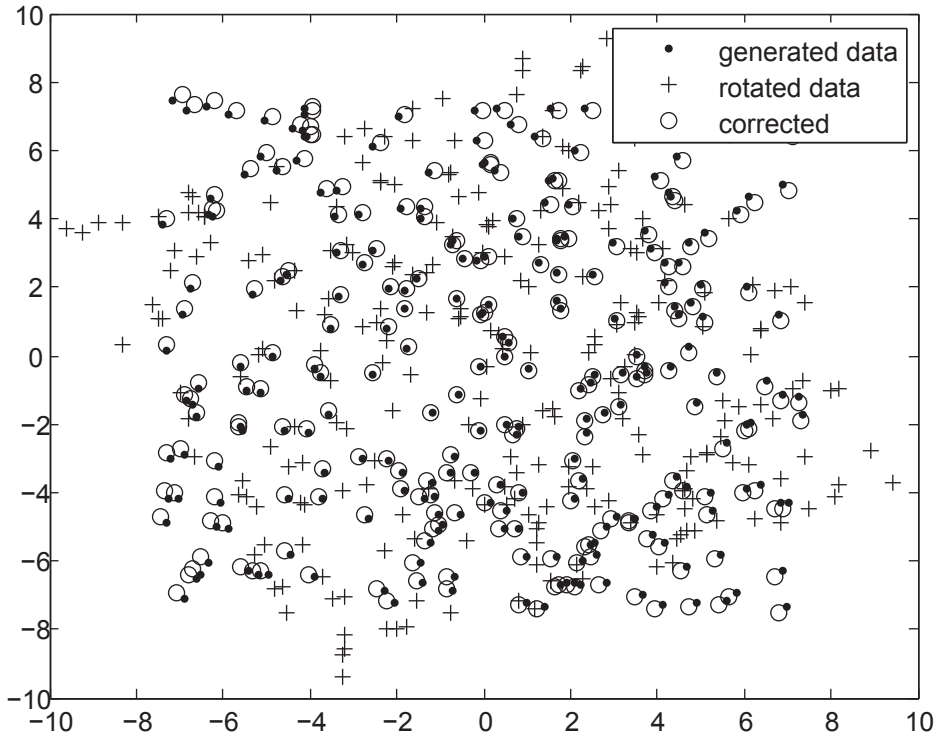


Figure 1: Locations for simulated dataset. Solid dots indicate the simulated data. Pluses indicate the simulated data after a rotation has been applied. Circles indicate data locations after the anisotropy angle has been determined and corrected for.

The sill variance is set to 1 and the nugget variance to 0.1. We simulated at 250 different locations. We further rotated the dataset by  $25^\circ$  to produce a fully anisotropic model.

Figure 1 shows the simulated dataset, the rotation and then the dataset after the anisotropy correction has been applied. It is worth noting that the dots almost return to their original locations, hence it can be seen how this is a particularly effective method for anisotropy correction.

In the experiments, the SSGP algorithm was run until the estimated parameters had converged. Convergence in this instance was determined if there was no change in the estimated parameters from the previous iteration.

Table 1 shows how correcting the data for anisotropy makes a significant difference when estimating covariance parameters. Using the default SSGP or the SSGP with MoM causes the parameters to be estimated poorly, largely because we have used a model which requires axis aligned isotropy in the covariance parameterisation. Furthermore, it can be seen that when the correct parameters are determined, the number of active points is reduced in this example. Because of the nature of the SSGP algorithm, generally speaking, the number of active points is related to the spatial variability in the data, more variation requires more active points to capture this behaviour.

If the initial estimates for the covariance parameters are good, the results support the intuition that the number of iterations of the algorithm should be reduced. By reducing the number of iterations, the execution time is also reduced, even though there are more active points.

Table 1: Results for simulated dataset. Note the true values of the N-S and E-W length scales should be 4.00 and 1.00 respectively.

Method	N-S	E-W	Sill	Nugget	Active Pts	SSGP Itr	Time(s)
SSGP	1.90	0.95	1.01	0.10	40	5	5.19
SSGP+MoM	1.89	0.93	1.00	0.10	42	4	3.51
SSGP+MoM+Aniso	3.97	0.96	1.04	0.10	25	2	1.74

Table 2: Results for BfS scenario dataset.

Method	N-S	E-W	Sill	Nugget	Active Pts	SSGP Itr	Time(s)
SSGP	6.04	1.08	58.13	7.24	37	5	4.97
SSGP+MoM	5.99	1.07	57.81	7.24	37	5	5.01
SSGP+MoM+Aniso	6.86	1.01	59.98	7.21	34	2	1.88

### 3.2 Radiation monitoring scenario

To test the method with a real-world dataset, a scenario generated by the German Federal Office for Radiation Protection (BfS) in WP5 was used. The aim, as with the simulated dataset was to reduce the number of iterations of the SSGP algorithm. A total of 261 observations were used to determine the parameters. The same procedure applied to the simulated dataset was also used here.

Table 2 shows the estimated parameters using the 3 different methods. The first thing to notice is that the number of active points is similar for the three different methods. One reason for this is that the anisotropy detection algorithm found a rotation of only  $4^\circ$ . Only two iterations were required for the SSGP+MoM+Aniso method as was the case with the simulated dataset. This shows how our improved method for determining the covariance parameters improves not only the estimation accuracy as shown with the simulated dataset, but also it shows how the efficiency of the algorithm has been improved. The range, sill and nugget parameters for the three methods all seem to be around the same values which is to be expected, given the axis aligned nature of the anisotropy.

## 4 Conclusions

In this deliverable we have shown how the SSGP algorithm, developed in earlier deliverables, can estimate covariance parameters more efficiently when coupled with other developments within the INTAMAP project. Firstly, we have found that correcting for anisotropy reduces the number of active points needed in the representation of a model. Secondly, we found by using a method-of-moments based variogram estimator that the number of iterations of the SSGP algorithm can be reduced. This improvement required a change to the existing algorithm whereby the active set was fixed after the first iteration of the algorithm. Furthermore, the experiments did not show any signs of fluctuations when computing the approximation to the log marginal likelihood as can sometimes occur. For larger datasets, where the number observations runs into the thousands, this would show much larger speed increases as a significant part of the computational time in the SSGP algorithm is the swapping in and out of locations into the active set. By fixing the active set after the first iteration, computation is significantly reduced, but this can only be done when one is confident of the initial covariance parameter values.

Further work could consider other estimates for the correlation range parameters above the method-of-moments techniques used here. A further improvement to the implementation could be an efficient algorithm for managing memory for the swapping-based refinement which this algorithm utilises. One other avenue of research could assume that active points were not a subset of the observed data, but rather they can be located anywhere. Snelson [2007] treat active point selection and parameter estimation as the same process and jointly learn active point locations and covariance parameters. This is prone to getting stuck in sub-optimal local minima and is also very slow to converge, however with a good initialisation, as we have here, the method might be more robust.

# Bibliography

- Noel A.C. Cressie. *Statistics for Spatial Data*. John Wiley and Sons, New York, 1993.
- L. Csató and M. Opper. Sparse online Gaussian processes. *Neural Computation*, 14(3): 641–669, 2002.
- E. L. Snelson. *Flexible and efficient Gaussian process models for machine learning*. PhD thesis, Gatsby Computational Neuroscience Unit, <http://www.gatsby.ucl.ac.uk/~snelson/pub.html>, 2007.