

Sparse, Sequential Bayesian Geostatistics

Dan Cornford, Lehel Csato, Manfred Opper

Neural Computing Research Group, School of Engineering and Applied Science,
Aston University, Birmingham B4 7ET. Tel. +44 (0)121 359 3611 x4667; Fax. +44
(0)121 333 4586; Email d.cornford@aston.ac.uk

Biography

Dr. Dan Cornford is a lecture in Computer Science and works in the Neural Computing Research Group at Aston University. Research interests are in the field of spatial statistics, space-time modelling and data assimilation.

Lehel Csato is a post-doc in the same group working on an EPSRC grant (GR/R61857/01) looking at applying sparse sequential Gaussian processes to data assimilation.

Manfred Opper is a Reader in the same group, with research interests in statistical physics, sequential learning and Bayesian learning.

Introduction

A limiting factor to the applicability of geostatistics (random field models / Gaussian processes) is the size of data set that can be analysed. This limit is imposed by computer memory constraints on the storage of the covariance matrix and the computational burden of inverting this matrix. While there are matrix inversion lemmas and computational tricks which can help here, the $O(n^3)$ scaling remains. In this paper we propose an alternative approach to geostatistical prediction based on Bayesian learning, that allows a *Kalman Filter like* sequential algorithm that can also use a sparsity heuristic and learn maximum likelihood estimates of the covariance parameters. This algorithm allows us to apply geostatistics to very large problems, particularly when the resulting process can be represented sparsely. The framework also provides a framework for dealing with non-Gaussian noise models and non-linear observation processes.

The Gaussian Process (GP), or random field model defines a probability distribution over the variables of interest, which we call the state variables. For most applications the key advantage of using a GP model is that it allows the user to estimate not just the mean value, but also the uncertainty in the state variables. In the Sparse, Sequential Bayesian (SSB) framework a GP prior distribution is placed over the state variables. The exact form of this prior distribution (that is the choice of the covariance function and the parameters of these) are set by the user, however we have developed an algorithm which allows the computation of the maximum likelihood covariance function parameters as part of the learning process.

Theoretical Framework

If we denote the state variable as $z(\mathbf{x})$, where \mathbf{x} is the spatial index which we drop from now on, the GP prior distribution as $p_0(z)$, then learning in GP's (for a single data point) can be viewed as corresponding to the following Bayesian update:

$$P_1(z|d) = 1/\text{norm} * p(d|z)p_0(z) .$$

Thus the updated posterior distribution is equal to a product of an unknown normalising constant, the likelihood, $p(d|z)$, and the prior $p_0(z)$. This definition of

geostatistics may not seem natural to those used to the kriging equations. In our work we define a new parameterisation of the GP in terms of a vector \mathbf{a} , and a matrix \mathbf{Q} , which are not the familiar means, and covariances of the standard GP. These parameters, \mathbf{a} and \mathbf{Q} , completely represent the Gaussian process, and details are given elsewhere (Csato and Opper, 2002).

The algorithm proceeds by iteratively computing the posterior $p_{t+1}(z|d)$, given the new data point, by minimising the KL divergence, or relative entropy, between the approximating GP posterior and the exact (possibly non-Gaussian) posterior given by the product of the likelihood $p(d_{t+1}|z)$ and the prior, which is the predictive posterior from the previous step, $p_t(z)$. This means that \mathbf{a} and \mathbf{Q} grow in size as each data point is considered. The prior is always Gaussian, thus when the likelihood is Gaussian, so is the posterior, and everything is exact. In this case we retrieve the kriging solution, but with an iterative learning strategy.

Sparsity

The method has several advantages of standard kriging methods. First we will consider sparsity. When we applying kriging, especially to data sets obtained from remote sensing (or indeed any other variable that is well sampled with respect to the process length scale) we often find that we have more observations than we can easily deal with. Several methods have been proposed to overcome this, the most commonly used being the selection of a kriging neighbourhood. In our framework it is possible to define the difference between on GP and another in terms of their KL distance, and by defining a heuristic for the minimal KL distance gain required to include a new data point and thus increase the number of observations being used we can control the growth in complexity of the algorithm. It is important to note that even those points which are not included in the basis vector set affect the GP and it's predictions: that is they change \mathbf{a} and \mathbf{Q} , but do not increase the size of \mathbf{a} and \mathbf{Q} . The term basis vector is used to denote those data points used by the algorithm.

When we do not include all the data, through the application of a sparsity heuristic, the learning algorithm is no longer exact, indeed the solution can depend on the order the data is presented. It was previously thought that this reason would preclude a Kalman filter being developed for spatial data, however we have been able to exploit a recent idea from the machine learning community; that of Expectation Propagation (EP) (Minka, 2000). This allows data to be re-used by taking into account its previous effect on learning. Thus we can cycle through the data set many times, to ensure that the solution we obtain is independent of the order of the presentation of the data to the algorithm.

The ability to recycle the data means we have also been able to learn the parameters of the covariance function (the variance and length scales typically), although the form of the covariance function must be specified by the user. We are exploring whether it is possible to use the Bayesian model evidence to select the most appropriate model. At present it is only possible to retrieve maximum likelihood (point) estimates of the covariance function parameters, although we could quite easily add priors over these and retrieve the maximum a posteriori probability values.

General Error Distributions

Another advantage of the method we propose is that it can be used when the likelihood is non-Gaussian. The likelihood is the probability of the observations given the GP state variable. Traditionally in kriging the observations and the state variables have often been the same thing (modulus measurement noise), however it is very often the case that we can only make indirect measurements of the variable of interest. For instance in remote sensing we may be interested in sea surface temperature, by we measure a radiance: these are linked, but not in a simple way. When the observation process is non-linear, or the errors are known to be non-Gaussian, then the posterior distribution, given a GP prior will in general be non-Gaussian. The only general way to address inference under this model is to attempt a Monte Carlo sampling based analysis (Diggle and Ribiero, 2001). This has several problems, such as the time taken and issues of convergence.

With our method we need to compute the update when adding observations one at a time. This means that the integral we need to approximate is much lower dimensional, and can be approximated quickly by methods such as quadrature, or often, calculated analytically. In effect the integral we wish to compute will allow us to project the non-Gaussian posterior each sample to the best approximating posterior GP, where best is taken in the KL divergence sense. This means we can very efficiently learn the posterior GP with a variety of noise models, and with general non-linear observation operators, although this often necessitates approximations (e.g. the linearisation of the observation operator) which mean the data must be recycled many times to achieve a stable solution.

Applications



Figure 1. An example of sparse learning using a non-linear forward model and a vector GP prior. The figure shows the posterior distribution and the associated

uncertainty, together with the basis vectors (in black) that make up the approximation.

We will illustrate the application of this method on simulated data, to show the consistence and real data to show the utility and potential benefits. The real data will involve surface temperature interpolation to show the method working on a scalar GP with a Gaussian noise model, and wind vector retrieval from remotely sensed scatterometer observations (Figure 1), to show the method working on large data sets, with a non-linear observation operator, and non-Gaussian likelihoods. We will show the effect sparsity has and demonstrate the learning of covariance parameters. Software for Matlab will be shortly available.

References

Csato, L. and Opper, M., 2002. Sparse on-line Gaussian Processes. *Neural Computation*, **14**, 641-669.

Minka, T. P., 2000. *Expectation Propagation for Approximate Bayesian Inference*. Ph.D. Thesis, MIT, USA. (<http://vismod.www.media.mit.edu/~tpminka>)

Diggle, P. J. and Ribiero J. R., 2001. Bayesian inference in Gaussian model-based geostatistics. *Geographical and Environmental Modelling* (to appear).

The preferred presentation format is oral.