Sequential, Sparse Learning in Gaussian Processes

Dan Cornford, Lehel Csato and Manfred Opper

Neural Computing Research Group, Aston University,
Birmingham B4 7ET, UK.
Telephone: +44 (0) 121 359 3611 x4667
Email: d.cornford@aston.ac.uk

Abstract

The application of Gaussian processes (or Gaussian random field models in the spatial context) has historically been limited to datasets of a small size. This limitation is imposed by the requirement to store and invert the covariance matrix of all the samples to obtain a predictive distribution at unsampled locations. Various ad-hoc approaches to solve this problem have been adopted, such as selecting a neighbourhood region and / or a small number of observation to use in the kriging process, but these have no sound theoretical basis and it is unclear what information is being lost. In this paper we present a recently developed Bayesian method for estimating the mean and covariance structures of a Gaussian process using a sequential learning algorithm which attempts to minimise the relative entropy between the true posterior process and the approximating Gaussian process. By imposing sparsity in a well defined framework, the algorithm retains a subset of *'basis vectors'* which best represent the *'true'* posterior Gaussian random field model in the relative entropy sense (that is both the mean and covariance are taken into account in the approximation). This allows a principled treatment of Gaussian processes on very large data sets, particularly when they are regarded as a latent variable model, which may be non-linearly related to the observations. We show the application of the sequential, sparse learning in Gaussian processes to wind field modelling and discuss the merits and draw-backs.

1 Introduction

The aim of this paper is not to give a review of traditional geostatistics, which is excellently covered in Cressie [1993], rather to introduce a new method for learning in Gaussian processes which has application to the processing of large data sets using a Bayesian geostatistical framework. Section 2 gives a brief review of geostatistics, using a slightly non-standard notation, in order to provide the context of this work. The parameterisation of the Gaussian process is discussed in Section 3, which is crucial for the novel learning algorithm that is developed in Section 4. The concept of sparsity is introduced in Section 5, while Section 6 shows how it is possible to learn hyper-parameters of the covariance model within the framework developed herein. We illustrate the application of the methods in Section 7 and discuss the limitations and potential of the algorithms in Section 8.

2 Background

Geostatistics has been around for many years, and is a well studied and frequently used branch of statistics [Cressie 1993]. It is based around an assumption that any finite collection of random variables (typically indexed by spatial location) is jointly Gaussian – that is $\boldsymbol{s}$ is a Gaussian process. Location is represented by the vector $\boldsymbol{x}$ and the variable of interest, referred to as the state variable, is represented by $\boldsymbol{s}(\boldsymbol{x})$. In what follows we will tend to suppress the explicit dependence on $\boldsymbol{x}$ for notational convenience, although this is still clearly present. We define a Gaussian process as:

$$\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}|K|^{1/2}} \exp\left(-0.5(\boldsymbol{s}-\boldsymbol{\mu})' K^{-1}(\boldsymbol{s}-\boldsymbol{\mu})\right) , \tag{1}$$

where $\boldsymbol{\mu}$ is the mean function of the process, which we shall assume without loss of generality is zero, K is the covariance and d is the dimension of $\boldsymbol{s}$. The parameters of the model, which we will refer to as *hyper-parameters*, are denoted by $\boldsymbol{\theta}$, and are regarded as parameterising the covariance function. If the mean were non-zero then $\boldsymbol{\theta}$ would include the parameters of the mean function. Thus $K = K(\boldsymbol{\theta})$ where in most cases the covariance function is chosen from some parametric family, such as an exponential, squared exponential (Gaussian) or spherical covariance model [Cressie 1993].

Geostatistics can be broken down into two main activities:

- determining the form of the covariance matrix (e.g. variogram estimation),
- and performing prediction (e.g. kriging).

In this work a general framework is assumed where observations of the process, $\boldsymbol{y}$, will not be directly of the state, $\boldsymbol{s}$, but rather they are indirect observations related to state by

$$\boldsymbol{y} = H(\boldsymbol{s}) + \boldsymbol{\epsilon} , \tag{2}$$

where H defines the known observation operator, and $\boldsymbol{\epsilon}$ defines the error on these observations, which is not necessarily Gaussian, but is assumed independent and identically distributed. Writing the model is this way is identical to the model based geostatistics of Diggle *et al.* [1998].

Observation operators, sometimes called *'forward models'*, map the state variable to the observations, and are particularly useful where the observation of the state is indirect, such as commonly occurs in remote sensing. Of course where we can directly observe the state the observation operator is simply the identity function, but this formalism remains useful since it is possible to deal with non-Gaussian noise in the direct observations in the same framework.

A Bayesian interpretation [Cressie 1993; Diggle *et al.* 1998] is adopted, where the aim is to infer the posterior distribution of the state, $\boldsymbol{s}$, given the all observations, $\boldsymbol{y}$:

$$\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{y}, \boldsymbol{\theta}, H) = \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{s}, H)\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{\theta})}{\int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{s}, H)\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{\theta}) d\boldsymbol{s}} . \tag{3}$$

This has the standard form of posterior = likelihood × prior ÷ evidence (or normalising constant)[1]. This framework for thinking about Gaussian processes can be very useful: the Gaussian process is seen as specifying a prior distribution over $\boldsymbol{s}$ (as a continuous function of $\boldsymbol{x}$), which is then updated

[1] Equation 3 is not a fully Bayesian model, since this would also treat the hyper-parameters as unknowns, which must also be integrated over for marginal inference on the state, but this adds an additional complexity, which necessitates sampling and is not pursued herein.

into the posterior given the observations $\boldsymbol{y}$. These process based ideas can help get away from some of the arbitrariness of the choices that must be made during a geostatistical analysis of data. For Gaussian noise and linear observations this posterior can be determined analytically, since the integral in the denominator is Gaussian and can be analytically evaluated. However numerically the evaluation of Equation 3 requires the inversion of a covariance matrix of dimension $nd \times nd$, where n is the number of samples and d is the dimension of the state space (often one, except in co-kriging), which is very computationally expensive and prohibits the treatment of large data sets with e.g. $nd > 1000$.

Several approaches exist to mediate the numerical problems that arise in large data sets, such as using moment based estimators to estimate the covariance function, then using local neighbourhoods to solve the prediction equations. This has the problem of generating artificial boundary effects as samples are included and excluded by the neighbourhood, which is chosen on an arbitrary basis.

When H in non-linear, or the observation noise is non-Gaussian the solution to Equation 3 is no longer analytic and optimisation methods can be used to provide maximum *a posteriori* probability estimates, or sampling can be used to provide a complete non-parametric estimation of the posterior distribution. Sampling based methods are very numerically intensive and suffer from issues of convergence detection, that is, when have we taken enough samples to provide a stable estimate of Equation 3? Optimisation methods are feasible, but only produce a single estimate of the state, without any uncertainty measure, although in principle such measures could be estimated by determining the Hessian matrix of Equation 3 at the maximum *a posteriori* probability value. This would also be rather expensive; using the Hessian, which is a local measure, might be rather sensitive to multiple minima in Equation 3 or slow convergence of the optimiser.

2.1 Covariance estimation

The most common approach to covariance estimation is to assume (and possibly even check) strict or second order stationarity [Cressie 1993], and then use this assumption to allow inference of a covariance function or variogram using an ergodic assumption. The stationarity assumption is crucial to the existence and inference of a stationary covariance function.

In method of moments based approaches, a non-parametric estimator is used to compute sample covariance functions by computing binned estimates of the sample covariance as a function of separation distance. To make this model continuous it is then common practice to fit a covariance function to this non-parametric estimator – a method of moments estimator. Often the form of the covariance function can be chosen on the basis of arguments about the physical processes which generated the data, or more data driven methods such as cross validation can be used. It can be very difficult to estimate properties of the process, such as differentiability using observations unless the observations sample the process very densely.

Alternatively, and with more statistical rigour, it is possible to estimate the covariance function directly from the data using a maximum likelihood method. This is straight-forward where H is linear, but is more tricky for non-linear H, or non-Gaussian noise, $\boldsymbol{\epsilon}$. Given the model Equation 1 and the observation equation Equation 2, the likelihood of the observations is dependent on the hyper-parameters, $\boldsymbol{\theta}$. These parameters are typically length scales and variance scales, although some covariance functions also possess smoothness parameters - e.g. the Bessel function based covariances, or Matérn covariance functions [Cressie 1993].

The standard practice is to minimise the negative log likelihood with respect to $\boldsymbol{\theta}$. The likelihood will be a non-linear function of the $\boldsymbol{\theta}$ even under the linear Gaussian model, thus optimisation algorithms must be used. Each step in the optimisation **requires** computation of the inverse covariance matrix for the whole data set, something which is again computationally very expensive for large data sets.

In principle, where prior knowledge is available (which we would argue is in almost every case) the hyper-parameters, $\boldsymbol{\theta}$, should be given prior distributions and estimation should compute the maximum *a posteriori* probability values of these parameters. The optimal solution would be to compute the joint posterior over the state and hyper-parameters and then integrate over the hyper-parameters to compute the marginal distribution of the state. This in not numerically practical for most situations, so a maximum *a posteriori* probability estimate is generally sought. For very large data sets it is reasonable to expect (if the stationarity assumption is valid) that the posterior distribution of the hyper-parameters is strongly peaked, and thus the impact of a maximum *a posteriori* probability assumption will be minimal.

2.2 Prediction

Once the hyper-parameters of the model have been estimated, it is then possible to make predictions, either where there is data, or where there is none, using the fitted stationarity covariance function to estimate the covariances between locations. This activity is generally referred to as kriging (in its many forms), and can be considered as the best linear unbiased estimator, given the assumptions made. As well as a prediction of the mean, a prediction of the covariances is also provided, which is important because this predictive *distribution* is necessary to make optimal use of the data in decisions and further processing. In the linear Gaussian case (classical geostatistics) the prediction equation for the mean is

$$\boldsymbol{s}_p = \boldsymbol{k}' K^{-1} \boldsymbol{y} \ , \tag{4}$$

where $\boldsymbol{k}$ is the covariance between the observation locations and the prediction location, K is the covariance between all observations, $\boldsymbol{y} = \boldsymbol{s}$ since $H = 1$, and the equation for the prediction covariance is

$$K_p = K(0,0) - \boldsymbol{k}' K^{-1} \boldsymbol{k} \ . \tag{5}$$

Both equations require the computation of the inverse of the covariance of all the observations, although as noted earlier, in practice these equations are often solved over neighbourhoods by using a small subset of local points.

In the next section we introduce a parameterisation of the posterior Gaussian process which makes it possible to introduce a sparse, sequential variational learning algorithm for Gaussian processes, which enables treatment of very large data sets in a principled manner and allows for consistent estimation of hyper-parameters using all the data.

3 Parameterisation

In order to enable the general model defined in Equation 1, Equation 2 and Equation 3 to be treated numerically a universal parameterisation of the posterior of a Gaussian processes is developed. This allows the development of learning algorithms which can be modified to include sequential processing of observations, sparsity, and hyper-parameter estimation.

A natural representation of the posterior Gaussian process, Equation 3, can be derived, which is related to the representer theorem, often used with spline models [Wahba 1991]. The Gaussian process posterior mean is parameterised as

$$\langle \boldsymbol{s}(\boldsymbol{x}) \rangle_{\text{post}} = \langle \boldsymbol{s}(\boldsymbol{x}) \rangle_{\text{prior}} + \sum_{i=1}^{n} q_i K(\boldsymbol{x}, \boldsymbol{x}_i) \ , \tag{6}$$

where $\langle \boldsymbol{s}(\boldsymbol{x}) \rangle_{\text{prior}} = \boldsymbol{\mu}$ is mean function with respect to the prior (which is generally zero), and $K(\boldsymbol{x}, \boldsymbol{x}_i)$ is the covariance between the point $\boldsymbol{x}$ and the points $\boldsymbol{x}_i$ used in the approximation. The covariance function, parameterised by $\boldsymbol{\theta}$, is assumed known from the prior, although the hyper-parameters can be re-estimated, as is shown later in Section 6. The scalar q_i's thus define the mean value of the posterior process, which is very unlikely to be zero at any given point, even though the prior generally is. The covariance of the Gaussian process posterior is parameterised as

$$K_{\text{post}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) + \sum_{i,j=1}^{n} K(\boldsymbol{x}, \boldsymbol{x}_i) R_{i,j} K(\boldsymbol{x}_j, \tilde{\boldsymbol{x}}) \ , \tag{7}$$

where $K(\boldsymbol{x}_i, \tilde{\boldsymbol{x}})$ is the covariance of the prior Gaussian process between locations $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}}$ and $R_{i,j}$ contains the information about the posterior covariance. This is a redefinition of the way to approximate the posterior distribution Equation 3 in terms of a finite set of parameters, $\boldsymbol{q}$ and R. For notational convenience we will write $\boldsymbol{\alpha} = \{q_i, R_{ij}, \forall i, j = 1 \ldots n\}$.

The parameterisation using $\boldsymbol{\alpha}$ is exact in the linear, Gaussian case. In the non-linear and / or non-Gaussian case the posterior in Equation 3 is no longer Gaussian; methods to address this are discussed in the next section. It is also possible to retain the same form of parameterisation, but only retain a subset of the observations, which we refer to as *'basis vectors'* that makes it is possible to control the growth in the number of parameters retained in the parameterised posterior, as discussed in Section 5.

4 Learning framework

The aim is to produce a framework for learning the parameters, $\boldsymbol{\alpha}$, of the representation Equation 6 and Equation 7 using the Bayesian formulation from Equation 3. In the case of H linear and $\boldsymbol{\epsilon}$ Gaussian, this posterior can be computed exactly (solving a Gaussian integral) and an algorithm which allows sequential processing of the data can be developed to update the posterior Gaussian process, by an update to $\boldsymbol{\alpha}$. This is similar in spirit to the Kalman filter, and the result is exact, so it is possible to process the data in arbitrary order. Actually the algorithm is rather like the application of the iterative Woodbury matrix inversion formula [Press *et al.* 1992], and still scales computationally as $O(n^3)$, so it is equivalent to the process of inverting the matrix in Equation 4 and Equation 5. Details of the algorithm can be found elsewhere [Csató 2002; Csató and Opper 2002] – the intention here is to give a higher level overview; details of the computation are very involved.

When the observations are non-linearly related to the state, or the noise is non-Gaussian, then in general the posterior Equation 3 will no longer be Gaussian. This presents a problem, since the parameterisation proposed above can only represent Gaussian posteriors. There are two choices:

- accept the non-Gaussian posterior and attempt to sample from this very high dimensional distribution;
- accept that although the exact posterior is not Gaussian, a Gaussian process approximation to this posterior is the only feasible solution available in reasonable time.

The approach of sampling is prohibitively expensive for even moderately large data sets, and is completely unsuitable for real time applications. Note, it is the posterior distribution of the unobserved state, $\boldsymbol{s}$, that is assumed to be Gaussian, **not** the observations.

The approach adopted to learning the parameters, $\boldsymbol{\alpha}$, is a variational one. This involves defining an approximating distribution $\mathrm{q}(\boldsymbol{s})$ to the true posterior given by Equation 3, which is now denoted $\mathrm{p}(\boldsymbol{s})$ for notational convenience, although it is clear that this is still conditioned on the observations and hyper-parameters. The aim in variational learning is to determine the $\mathrm{q}(\boldsymbol{s})$ which best fits the true distribution $\mathrm{p}(\boldsymbol{s})$. Best is defined here to mean the Gaussian distribution which has minimal *Kullback-Leibler (KL) distance* to the true (non-Gaussian) posterior.

4.1 Kullback-Leibler distance – a measure of distances between distributions

The KL-distance, which is sometimes referred to as the relative entropy, for reasons which will soon be illustrated, measures the *'distance'* between two probability distributions. It is non-symmetric; the KL-distance between $\mathrm{p}(\boldsymbol{s})$ and $\mathrm{q}(\boldsymbol{s})$ is given by:

$$\begin{aligned} \mathrm{KL}(\mathrm{p}(\boldsymbol{s}), \mathrm{q}(\boldsymbol{s})) &= \int \ln\left[\frac{\mathrm{p}(\boldsymbol{s})}{\mathrm{q}(\boldsymbol{s})}\right] \mathrm{p}(\boldsymbol{s}) d\boldsymbol{s} && (8) \\ &= \int \ln\left[\mathrm{p}(\boldsymbol{s})\right] \mathrm{p}(\boldsymbol{s}) d\boldsymbol{s} - \int \ln\left[\mathrm{q}(\boldsymbol{s})\right] \mathrm{p}(\boldsymbol{s}) d\boldsymbol{s} \; . && (9) \end{aligned}$$

The first term in Equation 9 is the entropy of the true posterior which is an (unknown) constant, thus only the second term need be considered when minimising Equation 9 with respect to the parameters, $\boldsymbol{\alpha}$. This order in the KL-distance is appropriate because the average is over the true posterior; minimising this is equivalent to matching moments when $\mathrm{q}(\boldsymbol{s})$ is Gaussian. So it turns out that after all the complex maths the optimal approximation (in the sense described above) is the $\mathrm{q}(\boldsymbol{s})$ which matches the moments of the true posterior $\mathrm{p}(\boldsymbol{s})$. These moments are very simple to compute in the Gaussian linear case, but are far from simple in the non-Gaussian or non-linear H case. In practice it is necessary to compute derivatives (with respect to the parameters $\boldsymbol{\alpha}$) of integrals of the likelihood over the predictive Gaussian process prior approximation at the previous step.

4.2 Variational Bayes

As mentioned, minimising the KL-distance between the true distribution and the approximating posterior is equivalent to matching the (first two) moments of the two distributions. It may be helpful to think of this operation as a projection of the non-Gaussian posterior from Equation 3 to the Gaussian approximation which is closest as measured by the KL-distance metric. This projection step is key to the algorithm, and is carried out at each observation sequentially. Computational details are given in Csató and Opper [2002].

In practice this update requires the minimisation (and thus computation) of the integral of likelihood over the predictive Gaussian prior (which comes from the Gaussian posterior approximation having ingested all observations up to the current one). For some models this integral can be done analytically, such as when H is linear and the noise, $\boldsymbol{\epsilon}$, is exponential (see Figure 1), or where the likelihood is given by a Gaussian mixture model, but for general forward models it is necessary to either linearise or use sampling based methods. This approximation for non-linear H introduces an important factor into the framework – the method is approximate, due to the projection and thus the order of processing the observations may become important. This can be explained by noting that with

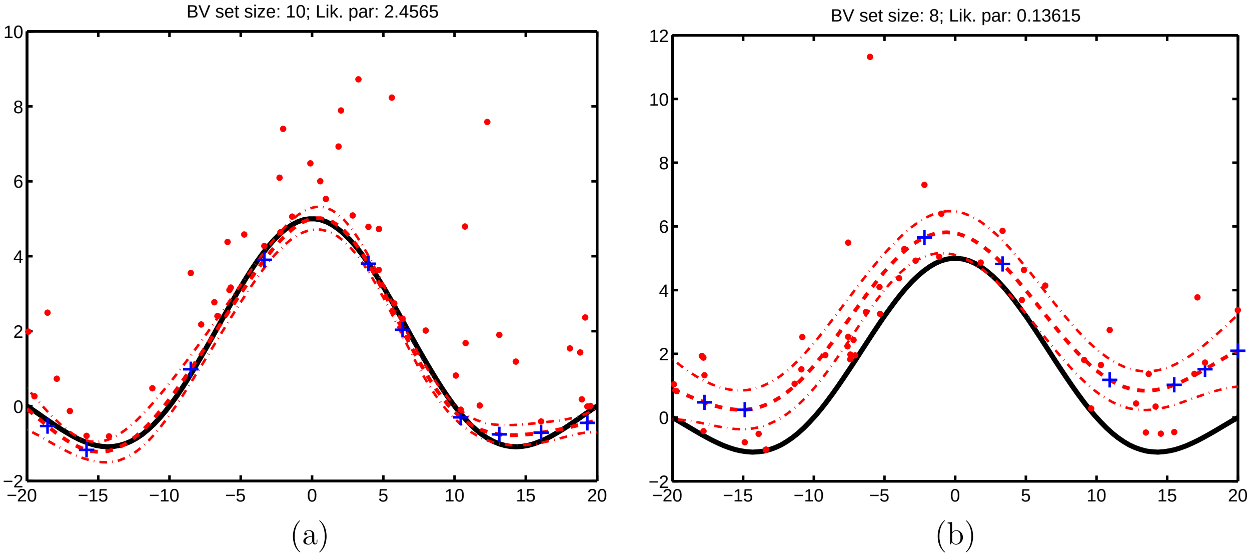


Figure 1: An example showing how the method is able to learn the posterior distribution of the state (mean given by the dashed red line, +/- one standard deviation given by the dash-dot red line), given observations (red dots) with very non-Gaussian (single sided exponential) noise. The true generating process is given by the solid (black) line, and the *basis vectors* are given by the blue crosses. In a) the correct noise model is applied in the learning, while in b) Gaussian noise is assumed. Not only is the estimate of the state biased, but the estimated uncertainty is also too high. The confidence intervals are plotted in latent (state) space, not observation space.

non-linear observation operators the log posterior is no longer a quadratic form (Gaussian), rather it may have several local optima, and an unlucky choice of the ordering of the data may result in the algorithm finding such a local minima. This algorithm is however far less sensitive to these type of effects compared to optimisation approaches because the likelihood is integrated (i.e. smoothed) over the prior (see the second term in Equation 9) and then extremised. To reduce the impact of data ordering on the approximation a data recycling method is introduced that helps minimise the problems of local optima, or poor convergence.

4.3 Data recycling

Inference in the online approximation in Csató and Opper [2002] is based on a single sweep through the data, but as noted above this might produce a rather poor approximation to the true posterior. Unfortunately, using further sweeps through the data with the same sequential algorithm in order to achieve a refinement of the approximation would violate the inherent assumption of the data independence and would lead to an unprincipled approximation.

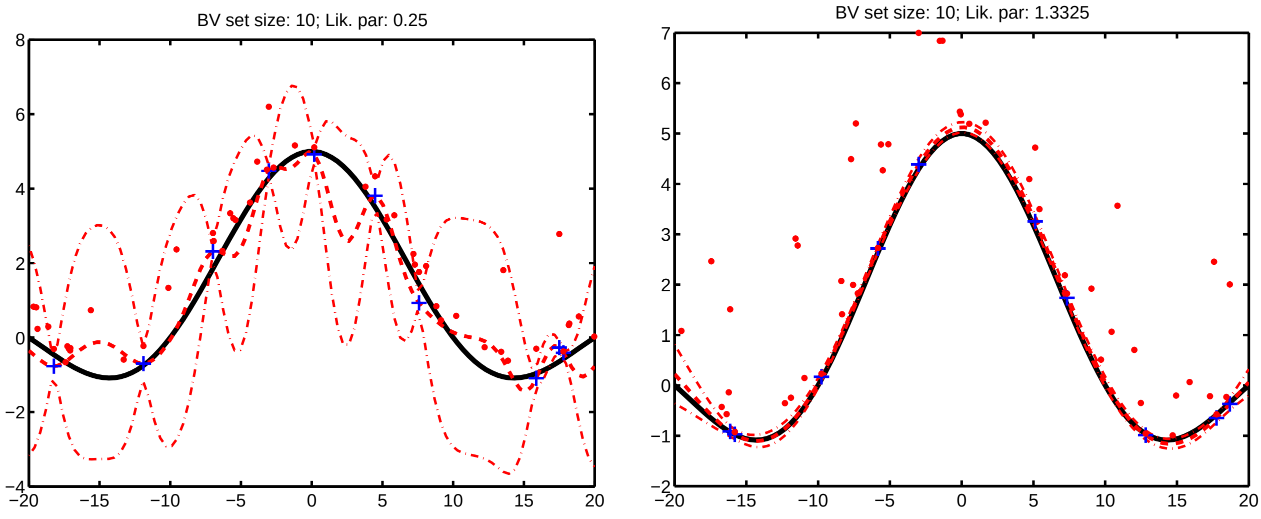
The problem of data recycling is overcome using the recently presented *expectation propagation* framework Minka [2000]. A principled improvement of the sequential approximation is achieved by altering the Gaussian process posterior, Equation 3, in a way that, although having seen the data once already, second and subsequent online inclusions are possible. Intuitively, the effect of the data to be processed is first approximately *'deleted'* from the solution and only then is included for a second time. Details of this method can be found in Csató [2002].

5 Sparsity

The next step is to ask whether the computational complexity of the algorithm can be reduced, but important features in the data retained. The answer is yes, but as ever there is a price. The parameterisation used for the Gaussian process posterior enables the posterior to be written, not in terms of the data, as is done in Equation 4 and Equation 5, but rather in terms of a set of parameters, $\boldsymbol{\alpha}$. These retained $\boldsymbol{\alpha}$'s are stored at a set of points that we call *'basis vectors'*, which can be the observation locations, but they don't have to be – they could be a grid, or locations at which we want to make predictions (e.g. see Figure 1). Learning involves estimation of the parameters, $\boldsymbol{\alpha}$, having seen each data point in turn. It is possible to decide at each step whether it is *'useful'* to increase the number of basis vectors and thus increase the size of $\boldsymbol{\alpha}$ and update these to take account of the observation. Alternatively the size of $\boldsymbol{\alpha}$ can be left unchanged, **but** the effect of the observation **is taken into account** through changes to $\boldsymbol{\alpha}$. In this way it is possible to control the complexity of the algorithm to $O(nm^2)$, where m is the number of basis vectors retained, and n is the number of observations as before. Alternatively it is possible to specify the minimal loss of information (in the KL-distance sense) that is acceptable, and only add basis vectors where this is exceeded. When approached in this way the sparsity ensures a compact representation appropriate to the complexity of the data.

In practice, at each step in the algorithm (that is as each observation is processed) the dimension of the posterior process is increased. To determine the optimal representation of this posterior in terms of $\boldsymbol{\alpha}$ (at the basis vectors) the posterior is projected on to the Gaussian process which is most close in the KL-distance sense. A basis vector will be added only if this is necessary according to whatever criteria are chosen to control the growth in complexity.

6 Hyper-parameter estimation



(a)

(b)

Figure 2: The same example set up as in Figure 1 with one sided exponential noise. In a) we have fixed the hyperparameters to reasonable (but still inappropriate) values and then learnt the posterior with these fixed values, while in b) the hyper-parameter values are learnt by maximising the evidence after each sweep through the data, as described in the text.

Using the sparse Gaussian process framework, it is also possible to perform an approximate estimation of any *hyper-parameters* contained in the likelihood or in the Gaussian process covariance functions using the *maximum likelihood II* method. In this procedure, the total probability or *evidence* of the data (the denominator in Equation 3) given by

$$\mathrm{p}(\boldsymbol{y} \mid H, \boldsymbol{\theta}) = \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{s}, H)\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{\theta}) d\boldsymbol{s} \ , \tag{10}$$

is maximised with respect to the collection of hyper-parameters. An *'expectation maximisation'* algorithm for an iterative minimisation of the evidence can be applied. The nontrivial *'E-Step'* of this algorithm requires the computation of posterior expectations which are consistently approximated using the sparse GP posterior provided by our method. Experiments on highly non-smooth data models such as regression with one sided exponential noise show a rather robust estimation of hyperparameters with this approach, as exemplified in Figure 2.

6.1 The complete algorithm

The overall algorithm can be summarised as:

1. **Specify** the prior Gaussian process distribution $\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{\theta})$
2. FOR $i = 1 : n$ LOOP
   (a) **Update** the current prior using Bayes rule (Equation 3) and a single observation to give $\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{y}_i, \boldsymbol{\theta})$
   (b) By matching the moments of $\mathrm{p}(\boldsymbol{s} \,|\, \boldsymbol{y}_i, \boldsymbol{\theta})$ to the approximating Gaussian process $\mathrm{q}(\boldsymbol{s} \,|\, \boldsymbol{y}_i, \boldsymbol{\theta})$, project this potentially non-Gaussian posterior to the closest Gaussian process in the KL-distance sense which becomes the prior in the next iteration.
      i. If necessary increase the size and **update $\boldsymbol{\alpha}$**, or simply update the existing values (**sparsity**)

   END LOOP
3. If desired **re-estimate** the hyper-parameters, $\boldsymbol{\theta}$, of the prior
4. If necessary **recycle** the data (always true if hyper-parameters have been re-estimated) – repeat to step 2.

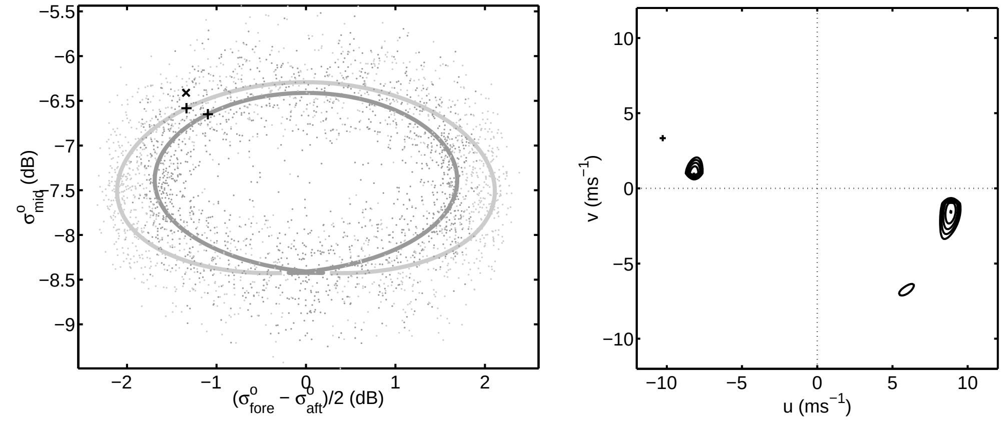
Many of the steps involved in this algorithm are non-trivial to implement, however we have developed a MATLAB toolbox which is freely available under GPL from:

`http://www.ncrg.aston.ac.uk/Projects/SSGP/`

This toolbox includes a number of demonstrations, but has been designed with a quite functional user interface, which could certainly be improved for applying these methods to practical problems. It is also true that MATLAB is not the ideal environment for implementing some of these methods, since it is not possible to completely vectorise the method.

7 Application to scatterometer data

Obtaining wind vectors, $\boldsymbol{v}$, over the ocean is important to numerical weather prediction (NWP) since the ability to produce a forecast of the future state of the atmosphere depends critically on



(a) (b)

Figure 3: (a) An approximate cross section through the 3D cone which defines the forward model. The lighter part is the upwind segment, the darker the downwind segment. A satellite observation, $\boldsymbol{y}$, is plotted as x, and corresponding closest points on the model manifold, are marked +. Also shown are samples from the noise distribution along the manifold (small grey dots), this giving an idea of instrument noise. (b) A contour plot of the local conditional probability density function of $\boldsymbol{y}$, as a function of $\boldsymbol{s}$. The NWP wind vector is marked by the cross and the two humps correspond to the two projections in (a).

knowing the current state accurately [Haltiner and Williams 1980]. However, the observation network over the oceans (particularly in the Southern Hemisphere) is very limited [Daley 1991]. Thus it is hoped that the global coverage of ocean wind vectors provided by satellite-borne scatterometers [Offiler 1994] will improve the accuracy of weather forecasts by providing better initial conditions for NWP models [Lorenc *et al.* 1993]. The scatterometer data also offer the potential of improved wind climatologies over the oceans [Levy 1994] and the possibility of studying, at high resolution, interesting meteorological features such as cyclones [Dickinson and Brown 1996].

7.1 Scatterometers

The illustration uses scatterometer data from the ERS-2 satellite; the on-board vertically polarised microwave radar operates at 5.3 GHz and measures the backscatter from gravity-capillary waves on the ocean surface of ∼5 *cm* wavelength. Backscatter from the ocean surface is measured by the normalised radar cross section, generally denoted by σ^o, and has units of decibels. A 500-km wide swathe is swept by the satellite to the right of the track of its polar orbit. There are 19 cells sampled across the swathe, and each cell has dimensions of roughly 50 by 50 *km*, which implies that there is some overlap between cells.

Each cell is sampled from three different directions by the 'fore', 'mid', and 'aft' beams, giving a triplet, $\boldsymbol{y} = (\sigma^o_f, \sigma^o_m, \sigma^o_a)$. This triplet $\boldsymbol{y}$, together with the incidence and azimuth angles of the beams (which vary across the swathe), is related to the average wind vector $\boldsymbol{v}$ within the cell [Offiler 1994]. Here $\boldsymbol{v}$ is our state vector $\boldsymbol{s}$. We assume that any unmodelled effects are largely related to wind speed and thus their impact is implicitly included in the empirical models which have been developed [Cornford *et al.* 2001; Bullen *et al.* 2003].

7.2 Forward models

In recent work [Bullen *et al.* 2003] we developed a scatterometer forward model based on a combination of a radial basis function network and a truncated Fourier series:

$$\boldsymbol{y} = a_0 + a_1 \cos(\chi) + a_2 \cos(2\chi) + a_3 \cos(3\chi) + a_4 \cos(4\chi) \ , \tag{11}$$

where $a_0, a_1 \ldots a_4$ are the outputs from the radial basis function network with inputs wind speed and beam incidence angle, and χ is the wind direction relative to the satellite azimuth angle. The non-linear forward model provides a local estimate of $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{s})$, and can be *locally* inverted (using non-linear optimisation) to retrieve $\boldsymbol{s}$, however care must be taken due to the presence of multiple local solutions, resulting from the Fourier form of the forward model and the observation noise, as illustrated in Fig. 3.

For practical applications where many thousands of scatterometer observations must be processed quickly, thus non-linear optimisation from multiple starts is prohibitively expensive, and only provides the most probable $\boldsymbol{s}$, not the posterior distribution, which is desired.

7.3 Vector Gaussian process priors

A stationary vector Gaussian process model is used to represent wind fields, based on the decomposition of a vector field into purely divergent and purely rotational flow, known as Helmholtz's theorem [Daley 1991]. We do not give details here, but the full development can be found in [Cornford 1998]. This allows control over the ratio of divergence to vorticity in the resulting vector field and automatically produces valid, positive definite, joint covariance matrices K for the wind vector components.

The maximum *a posteriori* probability values of the vector Gaussian process hyper-parameters, learnt on a set of wind fields obtained from numerical weather prediction models are used in the specification of the prior Gaussian process. These values suggested that the wind fields are once differentiable. Thus the modified Bessel covariance function simplifies to a polynomial-exponential covariance function which has the form:

$$C(r) = E^2 \left(1 + \frac{r}{L} + \frac{r^2}{3L^2}\right) \exp\left(-\frac{r}{L}\right) + \eta^2 \tag{12}$$

where r is the separation distance of two points, L is a characteristic length scale parameter, E^2 is the process variance and η^2 is the noise variance. This form is much quicker to compute and is used for this reason. We treat the maximum *a posteriori* probability hyper-parameters, $\boldsymbol{\theta} = \{r, L, E^2, \eta^2\}$ of the prior Gaussian process as known and fixed and do not re-estimate them.

7.4 Variational retrieval using sparse, sequential Gaussian processes

The variational approximation to learning is Gaussian processes is based on minimising a KL-distance, thus it requires the computation of certain Gaussian integrals over the result of propagating a Gaussian distribution through the forward model (the computation of the likelihood) – see Equation 3. For many problems of interest (such as most remotely sensed measurements) the state vector is non-linearly related to the observations, and the resulting combination of the Gaussian process prior and the likelihood produce in general non-Gaussian posteriors. Using the sequential character of the approximation, these integrals must be performed over the latent variable at a *single* location only

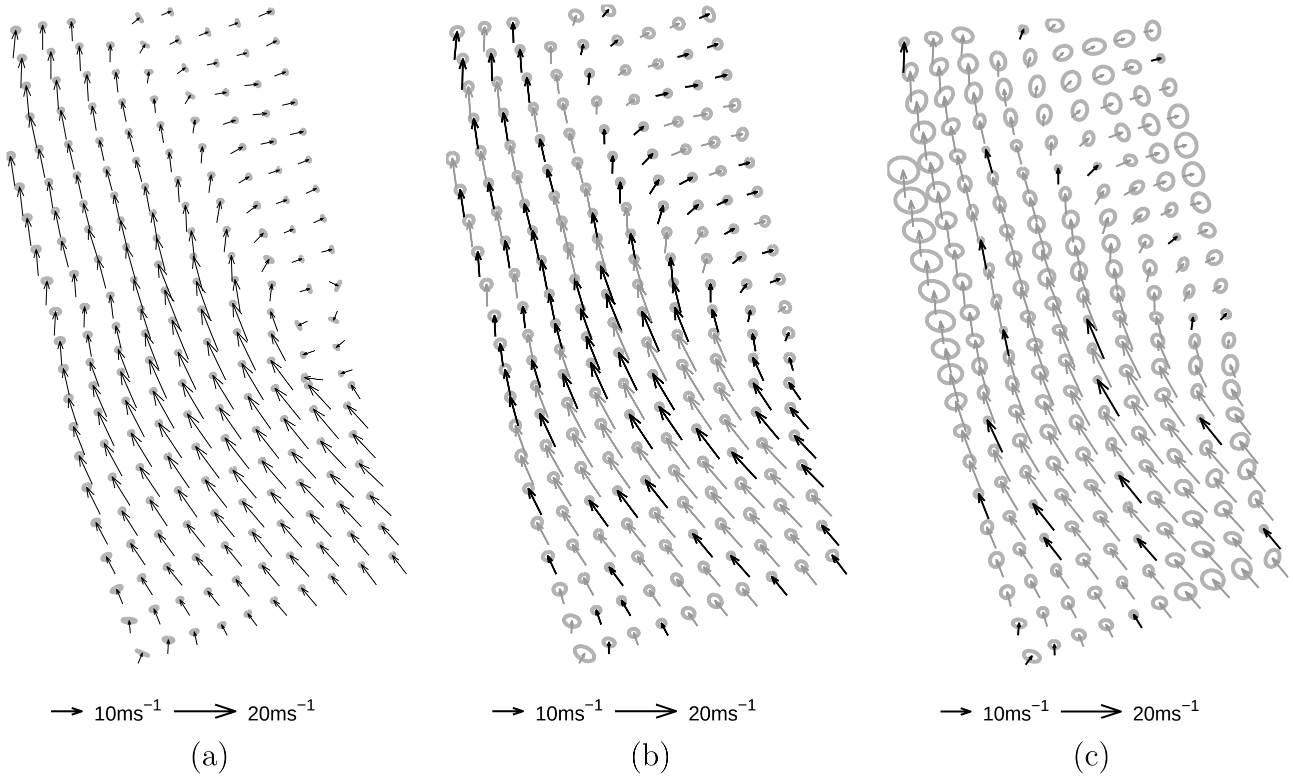


Figure 4: Wind field retrieval from scatterometer observations. (a) shows the results of Markov Chain Monte Carlo sampling which took over 12 hours of CPU time. In (b) the results of the parse sequential Bayesian learning algorithm which took less than 2 minutes on the same CPU with only the (100) basis vectors are drawn in black, the remaining wind vectors are interpolated, shown in grey. (c) shows the same data with only 25 basis vectors retained, which took even less time to compute.

and are thus typically low dimensional (2 dimensional for the wind field problem, where a vector Gaussian process is defined).

While this integral is *analytically tractable* for the case of the simple parametric inverse model [Csató *et al.* 2001], the integrals using general non-linear forward models require numerical methods. We have implemented several approximations to the necessary integrals. The most simple approach is to linearise the model about the prior mean (i.e. predictive mean of the current state given the observations which have so far been introduced into the approximation). This method is slower than the direct inverse model approach where the integrals can be solved analytically, but as shown in [Cornford *et al.* 2003] the results remain excellent, albeit that more iterations of the expectation propagation algorithm are required, since the linearisation means we are only using local approximations at each sequential update.

Figure 4 shows how the combination of the modified learning algorithm including expectation propagation for data recycling and linearisation applied to the forward models to compute the required integrals compares to an exhaustive sampling approach (for which convergence is not guaranteed, but has been visually inspected). The speed of processing using the sparse, sequential methods means that fully probabilistic retrieval of scatterometer winds can now be undertaken in real time using more accurate forward models.

We were also able to apply the non-linear forward model using an exhaustive sampling method, and an importance sampling method, based on samples from the *predictive prior* from the previous iteration.

These have been implemented, but not yet documented. We expect that the local sampling methods will present an important step to enlarge the applicability of sparse Gaussian process algorithms to realistic data models by allowing us to retain their full non-linearity without significantly slowing down the processing of the data.

8 Conclusions

This paper shows how a principled, Bayesian approach to geostatistics can be adopted without the need to restrict the applicability to small problems, or require very long computing times. The algorithm shows that it is possible to develop a Kalman filter like algorithm for kriging, and that we can incorporate sparsity and hyper-parameter estimation (variogram modelling) into the same consistent framework. The use of KL-distance to minimise the discrepancy between the approximating and true posterior provides a principled framework for sparsity and dealing with non-linear H and on-Gaussian errors. This allows us to undertake probabilistic inference in these situations with relatively small computational expense. Of course the method suffers from several problems, including:

- it is necessary to assume (check) stationarity;
- convergence of the algorithm can only be guaranteed in the linear H case;
- interpreting the KL-distances in not trivial – this is not a very intuitive measure for many people;
- implementing the method for any model H requires sampling (relatively slow, but no extra coding) or linearised version of H (generally faster, but may give a worse approximation and requires coding);
- at present hyper-parameter estimation is likelihood based, however an extension allowing priors over the hyper-parameters is possible;
- if the true posterior in Equation 3 is very non-Gaussian, then the Gaussian approximation may be rather poor and thus not particularly useful – this must be judged by the context;
- the algorithm involves some quite complex manipulations of the parameters $\boldsymbol{\alpha}$, and this, together with the use of the expectation propagation method means that it can be rather unstable in cases where the likelihood is very peaked (corresponding to a small nugget effect in geostatistics) – this could be helped by a different implementation, but requires considerable numerical analysis;
- using the method can be rather tricky because of the novelty of the approach;

Most of these drawbacks are either innate to any geostatistical method or are to do with the way the algorithm is implemented, however the most serious weakness of the algorithm is that although we can cope with non-linear H, in these circumstances the approximation may not be very good. That being said, this may be the best we can hope for in large systems; the only viable alternative is a sampling based approach (possibly a particle filter like extension), which will be computationally rather unappealing.

In future work we would like to extend the model to space-time phenomena with application to data assimilation in numerical weather prediction, better understand the impact of the way we treat the integral over the non-linear H, incorporate a Bayesian prior for the hyper-parameters, $\boldsymbol{\theta}$, in the re-estimation step and address the issue of inference in non-stationary and non-Gaussian processes, using a variety of methods.

Acknowledgements

This work was supported by EPSRC grant GR/R61857/01.

References

Bullen, R. J., D. Cornford, and I. T. Nabney 2003. Outlier detection in scatterometer data: Neural network approaches. *Neural Networks* **16**, 419–426.

Cornford, D. 1998. Flexible Gaussian Process wind field models. Technical Report NCRG/98/017, Neural Computing Research Group, Aston University, Aston Triangle, Birmingham, UK. URL: `http://www.ncrg.aston.ac.uk/~cornfosd/`.

Cornford, D., L. Csató, D. J. Evans, and M. Opper 2003. Bayesian analysis of the scatterometer wind retrieval inverse problem: some new approaches. *Journal of the Royal Statistical Society* **C**. accepted.

Cornford, D., I. T. Nabney, and G. Ramage 2001. Improved neural network scatterometer forward models. *Journal of Geophysical Research - Oceans* **106**, 22331–22338.

Cressie, N. A. C. 1993. *Statistics for Spatial Data*. New York: John Wiley and Sons.

Csató, L. 2002. *Gaussian Processes – Iterative Sparse Approximation*. Ph.D. thesis, Neural Computing Research Group, Aston Univeristy, http://www.ncrg.aston.ac.uk/Papers.

Csató, L., D. Cornford, and M. Opper 2001. Online learning of wind-field models. In *International Conference on Artificial Neural Networks*, pp. 300–307.

Csató, L. and M. Opper 2002. Sparse on-line Gaussian processes. *Neural Computation* **14**, 641–669.

Daley, R. 1991. *Atmospheric Data Analysis*. Cambridge: Cambridge University Press.

Dickinson, S. and R. A. Brown 1996. A study of near-surface winds in marine cyclones using multiple satellite sensors. *Journal of Applied Meteorology* **35**, 769–781.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed 1998. Model-based geostatistics. *Applied Statistics* **47**, 299–350.

Haltiner, G. J. and R. T. Williams 1980. *Numerical Prediction and Dynamic Meteorology*. Chichester: John Wiley.

Levy, G. 1994. Southern-hemisphere low-level wind circulation statistics from the SeaSat scatterometer. *Annales Geophysicae - Atmospheres, Hydroshperes and Space Sciences* **12**, 65–79.

Lorenc, A. C., R. S. Bell, S. J. Foreman, C. D. Hall, D. L. Harrison, M. W. Holt, D. Offiler, and S. G. Smith 1993. The use of ERS-1 products in operational meteorology. *Advances in Space Research* **13**, 19–27.

Minka, T. P. 2000. *Expectation Propagation for Approximate Bayesian Inference*. Ph.D. thesis, Dep. of El. Eng. & Comp. Sci.; MIT, vismod.www.media.mit.edu/∼tpminka.

Offiler, D. 1994. The calibration of ERS-1 satellite scatterometer winds. *Journal of Atmospheric and Oceanic Technology* **11**, 1002–1017.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery 1992. *Numerical Recipes in C: The Art of Scientific Computing* (2nd ed.). Cambridge University Press.

Wahba, G. 1991. *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.