Fast algorithms for automatic mapping with space–limited covariance functions

Ben Ingram[1], Dan Cornford[1], and David Evans[1]

Neural Computing Research Group, Aston University, Aston Triangle, Birmingham, B4 7ET. ingrambr@aston.ac.uk

1 Abstract

In this paper we discuss a fast Bayesian extension to kriging algorithms which has been used successfully for fast, automatic mapping in emergency conditions in the Spatial Interpolation Comparison 2004 (SIC2004) exercise. The application of kriging to automatic mapping raises several issues such as robustness, scalability, speed and parameter estimation. Various ad–hoc solutions have been proposed and used extensively but they lack a sound theoretical basis. In this paper we show how observations can be projected onto a representative subset of the data, without losing significant information. This allows the complexity of the algorithm to grow as $O\left(nm^2\right)$, where n is the total number of observations and m is the size of the subset of the observations retained for prediction. The main contribution of this paper is to further extend this projective method through the application of space–limited covariance functions, which can be used as an alternative to the commonly used covariance models. In many real world applications the correlation between observations essentially vanishes beyond a certain separation distance. Thus it makes sense to use a covariance model that encompasses this belief since this leads to sparse covariance matrices for which optimised sparse matrix techniques can be used. In the presence of extreme values we show that space–limited covariance functions offer an additional benefit, they maintain the smoothness locally but at the same time lead to a more robust, and compact, global model. We show the performance of this technique coupled with the sparse extension to the kriging algorithm on synthetic data and outline a number of computational benefits such an approach brings. To test the relevance to automatic mapping we apply the method to the data used in a recent comparison of interpolation techniques (SIC2004) to map the levels of background ambient gamma radiation.

2 Introduction

Spatial interpolation encompasses a large number of techniques that are used for prediction at spatial locations where data has not been observed. These techniques require that a model is constructed for how a given process behaves at locations where data has not been observed. Simple kriging, or best linear unbiased prediction (BLUP), is one method that has become very popular in geostatistics. By utilising covariance functions (or the variogram) to quantify spatial variability, maps can be produced from incomplete or noisy datasets using the kriging methodology. One aim of many automatic mapping systems is that the prediction is performed in (near) real–time [13]. Often spatial datasets are small since it can be expensive to obtain new observations and hence the goal of real time mapping can be achieved, because kriging can be performed efficiently with small datasets. However, in recent years the size of datasets has been increasing, due to the large increase of sensors on satellites sampling across the globe (eg. NEUROSAT [2]), aerial photography or large monitoring networks (eg. EURDEP[1]). It is not uncommon that the number of observations can run into the millions. To solve the kriging equations directly, a $(n \times n)$ covariance matrix, $\mathbf{\Sigma}$, for all the observations (n) needs to be inverted. Since the computation necessary to invert a matrix grows as $O\left(n^3\right)$, it can be seen that a naïve implementation of kriging is not feasible for large datasets. An additional problem that can arise with large datasets is due to instabilities associated with solving large systems of equations [10, 11] particularly when the ratio of the highest eigenvalue to the lowest eigenvalue of the covariance matrix becomes large (ill–conditioning). Concerns about the large linear systems involved in kriging and the instabilities in their solutions has motivated a search for better methods. It is unfortunate that the more data available for interpolation, the more numerically unstable the method may become [24]. This is fundamentally contrary to what one would like from an interpolation scheme.

In what follows we will comment on some of the common techniques that have been used to solve the above mentioned issues. Instead of inverting the covariance matrix directly, a number of researchers have proposed iterative methods for solving the kriging equations [17]. Although these methods don't have the poor scaling associated with direct methods ($O\left(n^2\right)$ per iteration), they don't yield exact results, although an improved estimate is obtained with each iteration. Solving the kriging equations exactly can be guaranteed, machine precision allowing, if the algorithm is run for n iterations. Hence the number of iterations should be significantly less than n to achieve any speed up advantages of using this method.

An alternative solution to the numerical instabilities that arise from solving large systems of equations is to use a sequential kriging algorithm [29, 5, 21]. Here each observation is considered individually and the inverse covariance

[1] http://eurdep.jrt.it

(or precision) matrix $\mathbf{\Sigma}^{-1}$ is updated iteratively using the Woodbury matrix inversion identity [19]. In this way, no large matrix inversions need to be performed. Furthermore, improvements in the stability of the system can be achieved by performing updates to the Cholesky factorisation of $\mathbf{\Sigma}^{-1}$ rather than using the Woodbury identity directly on $\mathbf{\Sigma}^{-1}$ which is known to be unstable particularly if insufficient numerical accuracy is retained or in the presence of round–off errors [14].

The problems arising from treating large datasets are not new, many techniques have been developed over the years to solve this computational bottleneck. In [9], an alternative solution was introduced which is called *moving window* or *local neighbourhood kriging* where a circular or elliptic moving window centred on the prediction location is used to select the observations within its boundaries that are to used for prediction. It is pointed out in [10] that local neighbourhoods can produce spurious behaviour, this happens particularly when we cross boundaries as new observations are introduced and removed from the moving window.

In many real world applications correlations between observations essentially vanish to zero beyond a certain separation distance, so it makes sense to use a covariance model that models this idea exactly since this leads to the covariance matrix $\mathbf{\Sigma}$ being sparsely populated. The advantage of $\mathbf{\Sigma}$ being sparse is that sparse matrix techniques can be used not only to solve the kriging equations with reduced computation, but also to reduce the amount of storage required for $\mathbf{\Sigma}$ [1]. In the same spirit of obtaining a sparse covariance matrix another approach given by [15] is based on using the Matérn covariance function and then inducing sparsity by tapering the covariance matrix. Utilising these methods gives varying results depending on the dataset, the reduction in computation and storage is connected to the sparsity in the data. Using the range parameter obtained from the variogram, the degree of sparsity can be gauged by the relative length of the range parameter with respect to the spatial extent and sampling density of the dataset.

Recent work by Cressie [3] has explored using reduced rank matrix approximations with satellite data which tends to be massive in size, but sparse in terms of prediction. The Frobenius norm between a fixed rank covariance matrix and the covariance matrix of the data is minimised to give an approximate covariance. Using the same tactic, other work has suggested using the Nyström method which replaces the covariance matrix $\mathbf{\Sigma}$ used in the kriging equations by a lower rank matrix $\tilde{\mathbf{\Sigma}}$ [32]. The quality of the approximation is good, however for small datasets it has been observed that the quality of the approximation can be poor [31].

In what follows we will employ a sequential simple kriging algorithm which is chosen not only for numerical stability, but also because we will show how the covariance matrix can be updated with new observations, but at the same time leaving its rank unchanged. This is achieved by projecting the effect, or weight of an observation onto a representative subset of the data (which we shall call *active points* from now on). In doing so the kriging weights of

the *active points* are updated in such a manner so as to minimise any loss of information (in the relative entropy sense [22])

$$\Delta H = -\frac{1}{2} log|\Sigma_t| + \frac{1}{2} log|\Sigma_{t-1}| \tag{1}$$

at each iteration t where Σ_t is the covariance matrix after t iterations and Σ_{t-1} is the covariance matrix calculated during the previous iteration.

Typically the number of active points can be reduced significantly before any measurable loss of information is observed. The number of active points we use can be selected *a–priori* or the algorithm can choose an appropriate active set size based on some error threshold not being exceeded. We will call this method *projected process kriging* (PPK) since the full kriging process is projected onto a reduced complexity kriging process. The underlying principles for this technique were first proposed in [6]. In this paper we will present the this method in the context of geostatistics.

The main contribution of this paper is to extend the PPK methodology to utilise space–limited covariance functions. In so doing, the covariance matrix $\mathbf{\Sigma}$ and it's inverse $\mathbf{\Sigma}^{-1}$ will become sparse, allowing the computational complexity to be reduced, not only in terms of speed but also in terms of data storage. This allows us to exploit two types of redundancy in the space–limited PPK method: redundancy in the observations caused by correlation (the projection on the *active subset*) and redundancy in the covariance matrix caused by the decay of correlation to almost zero in the sparse matrix representation using space limited covariance functions. Combined, these methods provide a principled algorithm which can optimally interpolate large datasets with minimal information loss. With the introduction of space–limited covariance functions, the data projection is performed only within the range of the support of the covariance lengthscale parameter, hence we will refer to this method as *local projected process kriging* (LPPK).

We show how this method is both fast and robust, and thus suitable to be applied to automatic mapping in near real–time. One particular challenging issue with automatic mapping is when extreme values are introduced into datasets. Since a basic assumption of kriging is that observed data closer together are more similar, when an extreme value is encountered this can often distort the fitted covariance model. By using space–limited covariance functions we are able to maintain the smoothness locally but at the same time produce a more robust, and compact, global model. Of course this is a pragmatic solution to the problem of extreme values; in future work we plan to models these as a separate population.

In Section 3 we will derive the equations that define our kriging methodology. Section 4 investigates the use of space–limited covariance functions for robust behaviour with extreme values. Section 5 describes the experimental set up and introduces the datasets that we use. Section 6 applies the methodology to the SIC2004 data showing that fast automatic mapping is possible,

together with a discussion of the results. Section 7 gives our conclusions about the application of this method.

3 Kriging Methodology

In this section we present the notation, equations and definition of our kriging extension (LPPK). Exact computation in kriging tends to become impossible when the number of observations exceeds several thousand. Local kriging methods show how moving windows of a subset of the dataset can be used. Either a specified number of observations or a search radius is used to determine which observations should be used. As the window moves new observations are included and discarded. As these boundaries are crossed discontinuities tend to occur. The longer the range parameter in the variogram the larger the number of observations that are needed to reduce this discontinuous effect. Despite these problems, kriging algorithms are almost always implemented in this manner. As a statistically principled alternative to this we use the entire dataset for prediction and show how the effect of observations can be sequentially projected onto a representative subset of the data called *active points* with a minimal loss of information. In doing so the complexity of the system is significantly reduced.

3.1 The kriging equations

We assume that $Z(\mathbf{x})$ is a random spatial process with the covariance function $k(\mathbf{x}_a, \mathbf{x}_b)$ where the process $Z(\mathbf{x})$ is known only at n spatial locations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. We define the vector of available data as:

$$\mathbf{Z} \equiv (Z(\mathbf{x}_1) \ldots Z(\mathbf{x}_n)) \tag{2}$$

Simple kriging is an optimal spatial predictor in that it minimises the mean–squared prediction error. To guarantee that the kriging predictor is optimal, we need to assume that Z is a stationary process and that $Z(\mathbf{x})$ has zero mean, or that the mean function has already been removed from the dataset. Note that in this paper we emphasise simple kriging; for other forms of kriging see [4]. The best linear unbiased predictor at an unobserved location $\mathbf{x}_{n+1}$ is given by

$$\hat{Z}(\mathbf{x}_{n+1}) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{x}_i) \tag{3}$$

The predictor is simply a weighted sum of all the observations. Each weight λ_i is the corresponding weight for the observation $Z(\mathbf{x}_i)$. The weights are calculated by a decreasing function of the distance between each observation location and the prediction location. This scaled distance is based on choosing a covariance model which describes the variation in the process. The covariance function will be referred to as $k(x_a, x_b)$ where x_a and x_b are two spatial

locations and the covariance function returns the covariance between the two locations. The kriging weights are thus calculated by

$$\boldsymbol{\lambda} = \boldsymbol{\Sigma}^{-1}\mathbf{k} \tag{4}$$

where $\boldsymbol{\Sigma}$ is the square $n \times n$ matrix of the covariance between each of the observations given by:

$$\boldsymbol{\Sigma} = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{bmatrix} \tag{5}$$

and $\mathbf{k}$ is the vector of covariances between the observation locations $\mathbf{x}$ and the prediction location x_{n+1}:

$$\mathbf{k} = \begin{bmatrix} k(x_1, x_{n+1}) \\ \vdots \\ k(x_n, x_{n+1}) \end{bmatrix} \tag{6}$$

This yields the equation for predictive mean of the process at location x_{n+1}:

$$\hat{Z}(\mathbf{x}_{n+1}) = \mathbf{k}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z} \tag{7}$$

and the predictive variance is:

$$\boldsymbol{\sigma}^2_{n+1}(\mathbf{x}_{n+1}) = k^* - \mathbf{k}^T\boldsymbol{\Sigma}^{-1}\mathbf{k} \tag{8}$$

where

$$k^* = k(x_{n+1}, x_{n+1}) \tag{9}$$

is the total sill variance of the process.

The complexity of solving the linear system $\mathbf{b} = \boldsymbol{\Sigma}^{-1}\mathbf{Z}$ directly in equation (7) is $O(n^3)$ in computation and $O(n^2)$ in storage. This basically prohibits straightforward kriging for large datasets of more than a few thousand, and also raises issues for smaller datasets that we wish to treat in (near) real–time. Furthermore, in an automatic mapping system the parameters of the process need to be reestimated without human intervention. In a Bayesian setting or if we were to use a maximum likelihood approach to estimate the process parameters we cannot avoid the need to invert the covariance matrix.

3.2 Sequential kriging

We now introduce a sequential kriging algorithm. Most kriging implementations are batch algorithms whereby all the observations (or sometimes a smaller subset) are processed in a single iteration. Here we will show a sequential algorithm whereby the model is updated as each observation is considered individually. To do so we will show the partitioned covariance matrix

$$\mathbf{\Sigma}_{n+1} = \begin{bmatrix} \mathbf{\Sigma}_n & \mathbf{k} \\ \mathbf{k}^T & k^s \end{bmatrix} \tag{10}$$

and it's inverse, which can be derived from the Sherman–Morrison–Woodbury formula [25],

$$\mathbf{\Sigma}_{n+1}^{-1} = \begin{bmatrix} \mathbf{\Sigma}_n^{-1} + \sigma_{n+1}^2 \mathbf{m}\mathbf{m}^T & \mathbf{m} \\ \mathbf{m}^T & (\sigma_{n+1}^2)^{-1} \end{bmatrix} . \tag{11}$$

Equation (10) shows how the covariance matrix $\mathbf{\Sigma}$ is partitioned and gives an intuition about the constituent parts of a covariance matrix where $\mathbf{k} = k(\mathbf{x}_{n+1}, \mathbf{x}_{1\ldots n})$ which is the covariance evaluated between the new observation and the observations already included in the model at that iteration. Equation (11) shows the partitioned inverse $\mathbf{\Sigma}^{-1}$. We see that the matrix $\mathbf{\Sigma}_{n+1}$ does not need to be inverted directly, the matrix inverse can be expanded successively by extending with an extra row and column for each new observation given calculations of $\mathbf{m} = -\sigma_{n+1}^{-2}\boldsymbol{\lambda}$ where $\boldsymbol{\lambda} = \mathbf{\Sigma}_n^{-1}\mathbf{k}$ which the reader will recognise from equation (4) and $\sigma_{n+1}^2 = k^* - \mathbf{k}^T\mathbf{\Sigma}_n^{-1}\mathbf{k}$ which is also the predictive variance at the new location [25]. Looking at the constituent parts individually an intuition can be gained about the process that is taking place. The vector $\mathbf{m}$ is the vector of weights given by $\mathbf{\Sigma}_n^{-1}\mathbf{k}$, of the existing system given the new observation that is to be added in the current iteration, this is then scaled by the inverse predictive variance $\left(\sigma_{n+1}^2\right)^{-1}$ (or predictive precision [26]). The existing matrix inverse, $\mathbf{\Sigma}_n^{-1}$, is updated by the outer product $\mathbf{m}\mathbf{m}^T$ scaled by the predictive variance of the new observation, which thus cancels.

It has been shown that the Sherman–Morrison–Woodbury formula, described above, can be numerically unstable [14] particularly in cases when sufficient numerical accuracy is not retained or in the presence of round–off errors. Indeed in [27] it is stated that the formula should only be used as a symbolic rewriting tool, and that for actual computations the Cholesky decomposition should be used since this is well known to be numerical stable and efficient. Therefore for stable implementations we recommend retaining the Cholesky factor of the inverse rather than calculating the inverse directly. By using the Cholesky factor we also increase performance. Further details of the Cholesky factor derivations can be found in [27].

3.3 Projected Process Kriging

The PPK algorithm makes a refinement to the sequential kriging algorithm. Instead of blindly increasing the size of the covariance matrix with each new observation, the PPK algorithm calculates how informative each new observation is. This information measure is then used to decide whether an observation is sufficiently important enough to warrant being added to the covariance matrix as described by the sequential kriging algorithm. If the observation adds little or no new information then the observation effect can instead be

projected onto the covariance matrix without having to increase the size of the covariance matrix. This process is repeated until all observations have been considered.

We will now show how to reduce the complexity of the algorithm, while retaining important features in the data. This is done in the framework of a sequential algorithm. We refer to the representative observations that are retained for prediction as the *active set* which we will denote by I and the r will represent the active set size. At each iteration t, an observation is added to the model, or active set, until a maximum number of observations (i.e. the model complexity, or active set size r, we desire) is reached. It is possible to update the model exactly with a new input, without increasing the size of the active set if

$$k\left(\mathbf{x}, \mathbf{x}_{t+1}\right) = \sum_{i=1}^{t} \lambda_{t+1}(i) k(\mathbf{x}, \mathbf{x}_i) \tag{12}$$

for all $\mathbf{x}$ [8]. In such a case we can exactly update the model without increasing the complexity. If equation 12 does not hold then an approximation is made by minimising the Kullback-Leibler (KL) divergence measure. We will not explain the details in this paper, but instead we direct the reader to [6] for a more rigourous treatment.

Rewriting the predictive mean and variance equations as:

$$\hat{Z}(\mathbf{x}_{t+1}) = \overrightarrow{\boldsymbol{\alpha}}\mathbf{k}\ , \tag{13}$$

$$\boldsymbol{\sigma}^2_{t+1}\left(\mathbf{x}_{t+1}\right) = k^* - \mathbf{k}^T \overrightarrow{\boldsymbol{\Sigma}}^{-1}\mathbf{k}\ , \tag{14}$$

gives an alternative parameterisation of the kriging equations. Notice that we have used introduced the projection operator, $\rightarrow$, notation to refer to the parameters of the projected model to distinguish them from the parameters of the traditional kriging model. The difference between the two inverse covariance matrices in equations (8) and (14) is that $\boldsymbol{\Sigma}^{-1}$ is the inverse covariance matrix between the observations in the active set only. $\overrightarrow{\boldsymbol{\Sigma}}^{-1}$ is also the inverse covariance matrix between the observations in the active set, however, additionally the effect of the inactive observations has been projected on to this matrix. We have introduced the notation $\boldsymbol{\alpha} = \overrightarrow{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_I$ where $\mathbf{Z}_I$ corresponds to the observations in the active set *prior to any projection*. However after observation projections, $\overrightarrow{\boldsymbol{\alpha}} \neq \overrightarrow{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}_I$. Thus $\overrightarrow{\boldsymbol{\alpha}}, \overrightarrow{\boldsymbol{\Sigma}}^{-1}$ parameterise the projected process kriging equations.

In the cases where a residual error occurs in trying to satisfy equation (12) we have to decide how informative the new observation is. The goal is to select the active set so that prediction error is minimised. A simple scoring heuristic is used to score the active points to determine which active point location least well represents the process. A number of scoring methods have been proposed

which measure scores such as relative entropy gain [21] and the information gain criterion [28]. Upon deciding whether the new observation is sufficiently informative, we either add it to the active set and remove the least informative point from the active set, or alternatively if it isn't sufficiently informative, we simply project the effect of this new observation onto the active points in the active set.

The projection step that updates $\overrightarrow{\boldsymbol{\alpha}}$ requires calculating a vector of weight innovations $\boldsymbol{\beta}$ for each new observation that is to be projected onto the existing active set

$$\boldsymbol{\beta} = \boldsymbol{\lambda} - \left(\overrightarrow{\boldsymbol{\Sigma}}^{-1} \mathbf{k} \right) , \tag{15}$$

where $\boldsymbol{\lambda} = \boldsymbol{\Sigma}^{-1}\mathbf{k}$ which is the weighting of the active set locations with respect to the new observation location. $\boldsymbol{\beta}$ essentially computes the difference in weights between the traditional kriging weights and the projected process kriging weights as each observation is incorporated. The update equations for the model are now

$$\overrightarrow{\boldsymbol{\alpha}}_{t+1} = \overrightarrow{\boldsymbol{\alpha}}_t + q_{t+1}\boldsymbol{\beta} \tag{16}$$

$$\overrightarrow{\boldsymbol{\Sigma}}^{-1}_{t+1} = \overrightarrow{\boldsymbol{\Sigma}}^{-1}_{t} + r_{t+1}\boldsymbol{\beta}\boldsymbol{\beta}^T \tag{17}$$

where for a Gaussian noise model on the observations, the projection parameters are

$$q_{t+1} = \frac{\hat{Z}\left(\mathbf{x}_{t+1}\right) - Z\left(\mathbf{x}_{t+1}\right)}{\sigma^2_{t+1}} \tag{18}$$

which measures the scaled difference, at the current location ($\mathbf{x}_{t+1}$), between the model prediction after t iterations $\hat{Z}\left(\mathbf{x}_{t+1}\right)$ (ie. the previous iteration and given an *active set* $\mathbf{I}$) and the observed value at the location $Z\left(\mathbf{x}_{t+1}\right)$, and

$$r_{t+1} = \frac{1}{\sigma^2_{t+1}} \tag{19}$$

which correctly scales the weights for the covariance updates. There are a number of things to note. Firstly, we make mention of the update parameters. In this paper we have assumed a Gaussian likelihood function although the update equations can be derived for arbitrary noise/likelihood models [7]. These update equations were calculated by minimising the KL divergence between the true model, Σ_t, and the approximating model, $\overrightarrow{\Sigma}_t$ at each iteration t. Secondly the reader should note that the updates for equation (17) are similar to the updates shown by the partitioned matrix in equation (10) but do not increase the size of the matrix.

We have just discussed the process of adding new observations to the model representation whether by increasing the complexity of the model or by projecting the observation's effect onto a representative subset of the data. We now note that another feature of this method is removal of active points from the active set. This is basically the reverse process of adding active points.

We direct the reader to [5] where full derivations can be found. To optimally apply this algorithm in practice, it has proven useful to discourage the use of active point deletion. A better approach is to select the active set *a–priori* (selecting the most informative locations rather than randomly) and then project the inactive observations onto this set. We have noted in experiments that a dynamic active set can lead to active point flip–flop behaviour whereby newly inserted active points are removed on the next iteration.

3.4 Relation to Bayesian frameworks

It is common practice within geostatistics that explicit stochastic models are rarely declared and as a result little use is made of the likelihood–based methods of inference which are central to modern statistics [12]. The phrase *model–based geostatistics* has been used to describe an approach to geostatistical problems based on using formal statistical methods under an explicitly assumed stochastic model. For our experiments we assume a model–based Bayesian framework. Many statistically principled extensions to kriging models have been proposed but have yet to find common place in the geostatistical community due in part to the additional complexity that they bring. In a Bayesian framework the parameters of the covariance model are considered random variables also. During the PPK algorithm the parameters are learned using a maximum likelihood type II approach [5] which uses the marginal likelihood to calculate the best parameters for the model. We believe the variogram remains a useful tool, particularly in helping to justify assumptions about the data and in previous work we have shown how this can aid the modeller in determining the covariance model [20].

4 Space–limited Covariance functions

We will now briefly introduce space–limited covariance functions. Is it often realistic to assume that correlations essentially disappear beyond a certain separation distance. However, many popular covariance functions (e.g. Gaussian, exponential) assume correlations exist, albeit small, at infinite spatial separations. There are a number of reasons why one might want to use a space–limited covariance function, however one cannot use these arbitrarily. The essential feature of compact covariances is to exploit the range of the variation within the data, with respect to the overall spatial range of the dataset. Firstly by using sparse matrix methods the space needed to store the covariance can be greatly reduced. Direct and iterative sparse matrix methods are available that result in increased performance. It should be noted that matrix reordering algorithms become particularly useful in increasing computational efficiency. We will use the Reverse Cuthill–McKee algorithm as this reduces the bandwidth of the sparse matrix [23].

Unfortunately, to construct a space–limited covariance function the covariance function cannot be simply truncated beyond a certain lag distance, since this can destroy the positive definitiveness of the covariance matrix [16]. For our experiments we will use a squared exponential covariance function along with a space–limited approximation proposed in [18] which is visually indistinguishable. It should be noted that space–limited covariance models are already commonly used in geostatistics in the forms of the circular and spherical covariance models. There are a plethora of space–limited covariance functions that have been used, particularly in Atmospheric sciences. There are a variety of ways to construct space–limited covariance functions from existing covariance functions. One example, the Wendland construction shows how the differentiability at the origin can be specified and still maintain it's positive definiteness [30]. Covariance validity strongly relies on the dimensionality of the data. Since we are working with spatial data such problems do not pose any serious problems within our work.

5 Experimental Setup

5.1 SIC2004

The problems of real–time automatic mapping were recently discussed in the Spatial Interpolation Comparison 2004 exercise and entrants presented various solutions [13]. The datasets supplied as part of the exercise were collected from an automatic radiation monitoring networking across Germany. In the contest only 200 observations were given to contest entrants, with a further 808 spatial locations with known values withheld for each entrant to submit the predictions of their model for comparison. To test the robustness of the algorithm a further dataset was given to participants in the form of a joker dataset where some extreme values were introduced in the form of a simulated radioactive contaminant escape.

5.2 Dataset

To demonstrate our method we will choose a larger subset of the available data. In doing so, the advantages of projecting the data onto a smaller subset will become more obvious. We selected 650 observations at random, and then used the remaining 358 observations for cross validation. We decided to fix the number of active points, but still allowed the algorithm the flexibility to choose which observations were more informative. Deciding how many active points should be retained is a complex problem and one with no clear theoretical answer. We opt for inspecting a curve as shown in Figure 1 which shows the cross–validation RMSE as a function of the number of active points. Before any clear loss of information is noticed, it can be seen that the number of active points can be reduced to about 90 at which point the RMSE starts to

rise. Figure 1 also shows the behaviour for the joker dataset. The reader will notice that the quality of the approximation is poorer for the joker dataset and that the approximation deteriorates around the 200 active points mark. Not having the joker dataset *a–priori*, means we have to set the number of active points based on our observations from the routine dataset. Knowing that an automatic mapping network has to deal with unpredictable events as well as routine situations, we opt to set the number of active points for our experiments to 120. This gives the algorithm some flexibility should unpredictable events be monitored. Determining the number of active point in the dataset is still something that requires much work. It is difficult to know *a–priori* the complexity inherent in the data. In selecting a value for the active set size we must allow for increased complexity in the observed data. We have taken this into account by setting the active set at 120 when during routine conditions it is clear that only 90 observations need be used. Although selecting 120 active points is well below the estimated 200 active points that are needed for the prediction with the joker dataset with negligible loss of information, the model predictions will still be appropriate for determining when emergency conditions arise.

6 Discussion

Table 1. Summary statistics for routine dataset

Covariance	**MAE**	**ME**	**RMSE**	**R**
Gaussian	0.4846	-0.0015	0.7016	0.793
Gneiting	0.4841	-0.0014	0.7012	0.794

Table 2. Summary statistics for joker dataset

Covariance	**MAE**	**ME**	**RMSE**	**R**
Gaussian	0.9136	-0.1380	2.2113	0.767
Gneiting	0.8996	-0.1195	2.2025	0.769

For the routine dataset the results (Table 1) using the different covariance functions are virtually identical. There seems to be a very slight improvement gained by using the compactly supported Gneiting covariance function. Figure 2 shows maps plotted using the two covariance functions. The features of each map are virtually identical as might be expected, although the active points selected seem quite different. One important difference is the density of

the covariance matrix. Figure 4 shows how the covariance matrix is sparsely populated in the case of the compactly supported covariance.

The joker data shows quite different results. Firstly, examining Figure 4, it can be seen that in this case the covariance matrix is much more sparsely populated. The sparsity is a function of the range parameter. During the algorithm the optimal parameters are inferred using a maximum likelihood type II approach having processes all the data. Clearly due to the extreme values, the range parameters have been significantly reduced. Looking at the maps in Figure 3 it is clear the horizontal range parameter has far exceeded the length of the vertical range parameter. Table 2 shows the summary statistics for the joker dataset. There is now a greater improvement in the quality of the compact Gneiting covariance based prediction than with the routine dataset, although the improvement is only marginal with little operational significance.

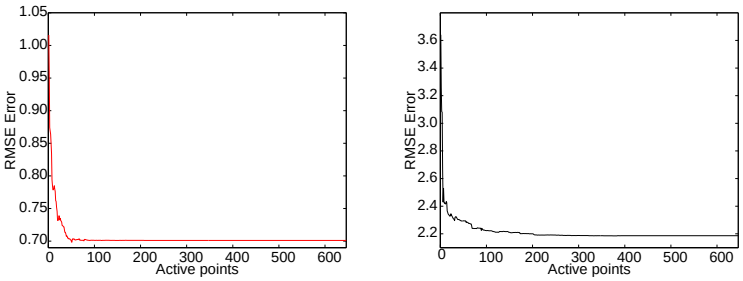

Fig. 1. Error curves showing how the crossvalidation error as a function of the number of active points returned for (left) the routine dataset and (right) the joker dataset.

7 Conclusions

We have presented a kriging extension (PPK) where the complexity of the model can be controlled by minimising the information loss from the dataset. We have exploited the redundancy in the data (or sampling density) by projecting the observations onto a representative subset of the data. We have further extended this by the use of space–limited covariance functions (LPPK) ensuring that a reduced number of the most relevant observations are projected and hence further improving stability and speed. By doing so we have exploited the two types of redundancy in the data thus giving an efficient method for automatic mapping. Our results indicate that prediction accuracy is not being lost, but rather a more efficient representation of the model has been found. Selecting the number of active points is still an unsolved problem, although our algorithm is sufficiently flexible to dynamically increase

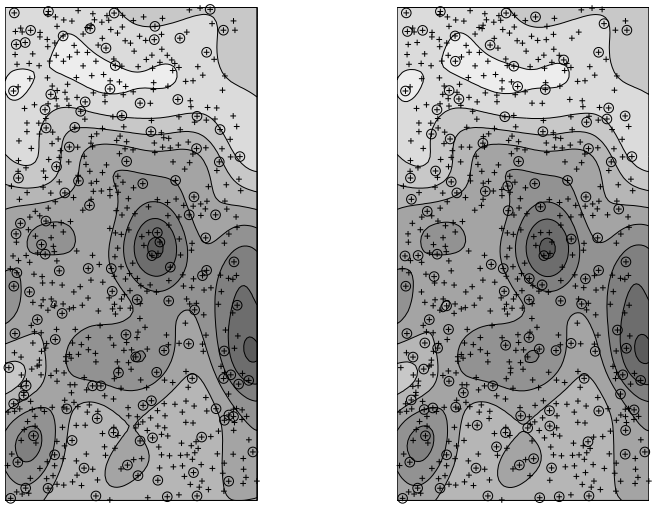

Fig. 2. Maps created from the routine dataset. Data are marked by crosses and active points by circles. (left) Gaussian covariance function and (right) Gneiting covariance function

the active set size based on ensuring that an error measure is not exceeded. This dynamic behaviour however reduces the computational efficiency and therefore we recommend fixing the active set.

In this paper we have used a fixed active set size since this leads to faster performance. We have seen how in emergency conditions that the complexity can vary particularly when emergencies arise. We note that it is possible with this algorithm to adaptively select the size of the active set. This is done based on measuring the residual error associated with each projection. If the residual error exceeds a certain threshold then the complexity of the model can be increased.

Having a more compact representation of the model is advantageous in multiple ways. We note that reduced bandwidth for model transmission and computation are both important properties for algorithms for use in low power systems such as satellites or in hand held devices.

In utilising a sequential algorithm we have ensured that a) the covariance matrix does not become poorly conditioned and b) large datasets need not fit into computer memory. With these two properties coupled with the projection of the data onto reduced rank representation we have opened up numerous avenues for dealing with large datasets in a principled way. Coupled with the use of sparse covariance matrices and space–limited covariance functions we have shown how we can further reduce the storage and computation burden.

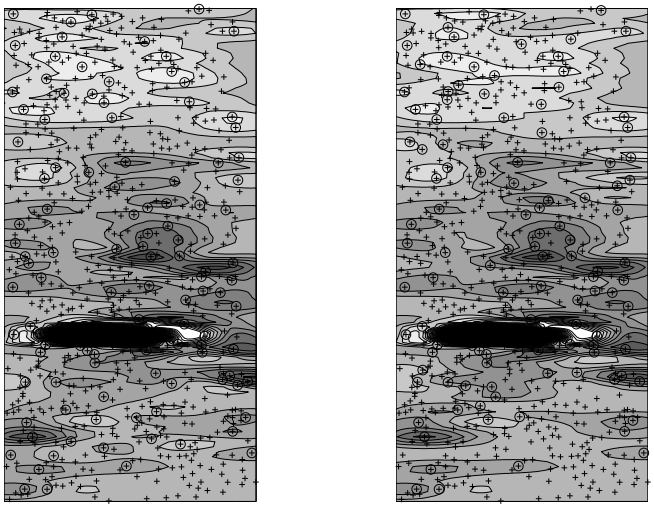

Fig. 3. Maps created from the joker dataset. Data are marked by crosses and active points by circles. (left) Gaussian covariance function and (right) Gneiting covariance function

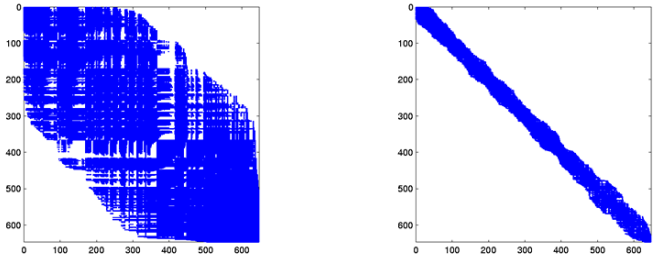

Fig. 4. Structure of covariance matrix after Cuthill–McKee reordering.(left) Routine dataset, 49.4% sparsity and (right) Joker dataset 7.8% sparsity

The problem of treating outliers has been a topic of much discussion. We have shown that by using a space–limited covariance function a slight improvement in robustness can be achieved. Although no significant improvements in robustness have been achieved, we feel that in a spatial setting with large datasets that space–limited covariance functions are appropriate due to their effective use of storage and reduction in computation.

In future work we will investigate using arbitrary likelihood and algorithm robustness to improve prediction accuracy. For larger datasets we plan to use

parallel computation algorithms to distribute computation power and memory needed across a number of processors.

Acknowledgements

This work was partially supported by the BBSRC contract 92/EGM17737. The SIC2004 data was obtained from [http://www.ai-geostats.org]. This work is partially funded by the European Commission, under the Sixth Framework Programme, by the Contract N. 033811 with DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. The views expressed herein are those of the authors and are not necessarily those of the European Commission.

References

1. R. Barry and R. K. Pace. Kriging with large data sets using sparse matrix techniques. *Communications in Statistics: Computation and Simulation*, 26(2):619–629, 1997.
2. D. Cornford, L. Csató, D. J. Evans, and M. Opper. Bayesian analysis of the scatterometer wind retieval inverse problems: some new approaches. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):609–626, 2004.
3. N. A Cressie. Spatial prediction for massive datasets. In *Mastering the Data Explosion in the Earth and Environmental Sciences: Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference*, 2006.
4. Noel A.C. Cressie. *Statistics for Spatial Data*. John Wiley and Sons, New York, 1993.
5. Lehel Csató. *Gaussian Processes – Iterative Sparse Approximation*. PhD thesis, Neural Computing Research Group, www.ncrg.aston.ac.uk/Papers, March 2002.
6. Lehel Csató and Manfred Opper. Sparse representation for Gaussian process models. In NIPS13ed, editor, *NIPS*, volume 13, pages 444–450. MIT, 2001.
7. Lehel Csató and Manfred Opper. Sparse on-line Gaussian Processes. *Neural Computation*, 14(3):641–669, 2002.
8. Lehel Csató and Manfred Opper. Greedy sparse approximation to Gaussian Processes by relative entropy projection. Technical report, Neural Computing Research Group, prep.
9. M. W. Davis. The practice of kriging. *Advanced Geostatistics in the Mining Industry*, 31, 1976.
10. M. W. Davis and C. Grivet. Kriging in a global neighborhood. *Mathematical Geology*, 16:249–265, 1984.
11. M. W. Dietrich and G. N. Newsam. A stability analysis of the geostatistical approach to aquifer transmissivity identification. *Stochastic Environmental Research and Risk Assessment*, 3:293–316, 1989.
12. P J Diggle, J A Tawn, and R A Moyeed. Model-based geostatistics. *Applied Statistics*, 47:299–350, 1998.

13. G. Dubois and S. Galmarini. Spatial interpolation comparison (SIC) 2004: introduction to the exercise and overview of results. In *Automatic Mapping Algorithms for Routine and Emergency Monitoring Data*, 2005.
14. S Fine and K Scheinberf. Effcient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2002.
15. R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15:502–523, 2006.
16. G. Gaspari and S. Cohn. Construction of correlation functions in two and three dimensions, 1996.
17. M. N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, Cambridge University, 1997. `http://www.inference.phy.cam.ac.uk/mng10/`.
18. T. Gneiting. Correlation functions for atmospheric data analysis. *Quarterly Journal of the Royal meteorological Society*, 125:2449–2464, 1999.
19. G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, second edition, 1989.
20. Ben Ingram, Lehel Csató, and David Evans. Fast spatial interpolation using sparse gaussian processes. *Applied GIS*, 1(2), 2005.
21. N. Lawrence and R. Herbrich. A sparse bayesian compression scheme — the informative vector machine. In *Advances in Neural Information Processing Systems*, 2001.
22. Neil D. Lawrence, Matthias Seeger, and Ralf Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems*, 2002.
23. Wai-Hung Liu and Andrew H. Sherman. Comparative analysis of the Cuthill-McKee and the reverse Cuthill-McKee ordering algorithms for sparse matrices. *SIAM Journal on Numerical Analysis*, 13(2):198–213, April 1976.
24. A. E. Long. *Cokriging, Kernels, And The SVD: Toward Better Geostatistical Analysis*. PhD thesis, The University of Arizona, 1994.
25. D. J. C. MacKay. Introduction to Gaussian processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, NATO ASI Series, pages 133–166. Kluwer Academic Press, 1998.
26. U Menzefricke. On the performance of the Gibbs sampler for the multivariate normal distribution. *Communications in Statistics - Theory and Methods*, 24:191–213, 1995.
27. M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, July 2003.
28. M. Seeger and C. Williams. Fast forward selection to speed up sparse gaussian process regression, 2003.
29. J A Vargas-Guzmán and T C Jim Yeh. Sequential kriging and cokriging: Two powerful geostatistical approaches. *Stochastic Environmental Research and Risk Assessment*, 13:416–435, 1999.
30. H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4:389–396, 1995.
31. C. K. I. Williams, C. E. Rasmussen, A. Schwaighofer, and V. Tresp. Observations on the Nyström method for gaussian process prediction. Technical report, Edinburgh University, 2002.

32. C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems*, 13:682–688, 2001.