Combining covariance tapering and lasso driven low rank decomposition for the kriging of large spatial datasets

Thomas Romary, Nicolas Desassis

Combining covariance tapering and lasso driven low rank decomposition for the kriging of large spatial datasets

Thomas Romary Nicolas Desassis

May 30, 2018

Abstract

Large spatial datasets are becoming ubiquitous in environmental sciences with the explosion in the amount of data produced by sensors that monitor and measure the Earth system. Consequently, the geostatistical analysis of these data requires adequate methods. Richer datasets lead to more complex modeling but may also prevent from using classical techniques. Indeed, the kriging predictor is not straightforwarldly available as it requires the inversion of the covariance matrix of the data. The challenge of handling such datasets is therefore to extract the maximum of information they contain while ensuring the numerical tractability of the associated inference and prediction algorithms. The different approaches that have been developed in the literature to address this problem can be classified into two families, both aiming at making the inversion of the covariance matrix computationally feasible. The covariance tapering approach circumvents the problem by enforcing the sparsity of the covariance matrix, making it invertible in a reasonable computation time. The second available approach assumes a low rank representation of the covariance function. While both approaches have their drawbacks, we propose a way to combine them and benefit from their advantages. The covariance model is assumed to have the form low rank plus sparse. The choice of the basis functions sustaining the low rank component is data driven and is achieved through a selection procedure, thus alleviating the computational burden of the low rank part. This model expresses as a spatial random effects model and the estimation of the parameters is conducted through a step by step approach treating each scale separately. The resulting model can account for second order non stationarity and handle large volumes of data.

1 Introduction

While a spatial datum was expensive to obtain in the traditional application fields of Geostatistics (e.g. drilling wells for oil reserve estimation), with the increasing deployment of remote sensing platforms and sensors networks, spatial data-base paradigms have moved from small to massive, often of the order of gigabytes per day. Therefore, methods for the geostatistical analysis of these kinds of data have to be developed. Indeed, richer datasets allows for more complex modeling but may also prevent from using straightforwardly classical techniques. The challenge of handling such datasets is to extract the maximum of information they contain while ensuring the numerical tractability of the associated inference and prediction algorithms. As an example, satellite image restoration is a challenging problem for geostatisticians, as it involves large amounts of data and possible non stationarity over the domain of interest, both in space and time.

The classical spatial predictor in geostatistics is the kriging, see e.g. (Matheron, 1970; Chilès and Delfiner, 2012). It applies when the phenomenon under study can be modeled by a second order random field. The first order moment accounts generally for large scale fluctuations of the phenomenon. It is described in a basis of known functions of the space-time, e.g. polynomials or exhaustively known covariates. They are generally called the drift functions. The covariance of the random field describes the regularity of the medium to small scale variability of the phenomenon.

The inference is generally conducted as follows: the drift functions are chosen. Then a parametric covariance model is selected and fitted. This step can also be straightforwardly automatized through the use of model selection and fitting algorithms. This step is conducted either through a method of moments (Desassis and Renard, 2013) or maximum-likelihood under a Gaussian hypothesis (Mardia and Marshall, 1984). This inference step is made more complex for several reasons in the context of large datasets. The phenomenon may exhibit a nonstationary covariance structure that has to be inferred (Fouedjio et al, 2015, 2016). The computation of the likelihood is also made impossible as it involves the computation of the determinant and the inverse of the covariance matrix of the data, whose size is $n \times n$, where n is the number of data. An approximation of the likelihood by compositing can then be necessary (Stein et al, 2004; Varin et al, 2011; Eidsvik et al, 2014).

Once the covariance model is known, the kriging interpolator is built as the best linear unbiased predictor of the random field at a new location. It coincides with the conditional expectation under Gaussian hypothesis. It consists of a weighted average of the data. The computation of the weights involves the inversion of the covariance matrix, whose size is $n \times n$. This approach is therefore intractable when n becomes large, as the computational complexity of the inversion scales with n^3.

Several approaches have been developed in the literature to address this problem. These approaches can be classified into two families, both aiming at making the inversion of the covariance matrix computationally feasible.

The first one circumvents the problem by enforcing the sparsity of the covariance matrix (Furrer et al, 2006), making it invertible in a reasonable computation time ($O(n)$). This implies however that the covariance model that is fitted to the data is compactly supported with a short range (Gneiting, 2002), neglecting possible medium scale variability. In other words, the fitted covariance is constrained to this particular structure and may not reflect the whole complexity of the phenomenon fluctuations. More precisely, while accurate in densely sampled areas, this method generally fails at providing correct predictions in scarcely sampled areas, as the correlation between remote locations decreases quickly towards zero, especially at the edges of the domain (Stein, 2013). This approach simplifies the inference. In particular, it makes the maximum-likelihood approach tractable as the inverse covariance matrix is made accessible (Kaufman et al, 2008).

The second available approach assumes, on the contrary, a low rank representation of the covariance function. Many low rank approaches have been proposed over the years, among which we can cite the fixed rank kriging (Cressie and Johannesson, 2008), the predictive processes (Banerjee et al, 2008) and also more recent works (Katzfuss, 2017; Nychka et al, 2015; Romary, 2013). It assumes that the covariance function can be represented on a limited number, say p, of basis functions. Adding a nugget effect, i.e. a local unstructured noise, to the modeling makes the inversion of the covariance matrix feasible through the Woodbury matrix inversion formula with a complexity $O(n \times p^3)$, where p is small, relatively to n. The latter approach tends to overestimate the regularity of the spatial phenomenon, failing to capture its local behavior. Again, the whole complexity of the phenomenon fluctuations may not be correctly reproduced in this framework, as the small scale fluctuatione are considered unstructured and modelled by a nugget effect. The generated predictions are generally too smooth, failing at reproducing the small scale fluctuations, see (Stein, 2014). It is worth noting that the low rank approach provides a non stationary covariance model by construction. Concerning the inference, chiefly two approaches are available: either a conventional model is fitted and a low rank approximation is consequently computed (Banerjee et al, 2008; Romary, 2013), or the inference is conducted through an expectation-maximization (EM) algorithm (Katzfuss and Cressie, 2011). Direct method of moments approach is also available (Cressie and Johannesson, 2008).

Finally, Markov random field models depend on the observation locations, and realignment to a much finer grid with missing values is required for irregular locations (Lindgren et al, 2011; Sun and Stein, 2016), which induces an additional cost for both inference and prediction.

In this work, we propose to use covariance models that combines the advantages of both approaches.

Indeed, using additively a compactly supported short range covariance and a low rank component in the modeling of the covariance allows capturing both local and medium scale fluctuations of the phenomenon and therefore improving the prediction accuracy. In this settings, the covariance matrix is still invertible through the Woodbury formula at the same computational cost as the low rank approach ($O(n \times p^3)$). This is not a new idea. It has been already proposed in (Stein, 2008), where a few Legendre polynomials are used on the sphere, in (Sang and Huang, 2012) where the remainder of a predictive process approximation is tapered and more recently in (Ma and Kang, 2017). The originality of our work is to take advantage of a scale separation between the sparse and low rank terms. Consequently, the inference of both terms can be conducted almost separately, making it computationally efficient.

Further more, in low rank approximations, the choice of the basis functions is generally taken arbitrarily. The set of basis functions is either built from a fixed family of functions (Cressie and Johannesson, 2008; Nychka et al, 2002; Stein, 2008) or from colums of stationary isotropic covariance matrices (Banerjee et al, 2008; Katzfuss, 2017). Here, the choice of the basis functions sustaining the low rank component is data driven and achieved through a selection procedure, which results in practice in a low number of functions, hence decreasing the associated computational burden. Moreover, this provides an important flexibility to the model allowing it to capture a large variety of possible fluctuations including varying anisotropy and smoothness, as illustrated in the examples.

The paper is organized as follows. The first part reviews briefly existing approaches and describes the proposed one. The second part details the developed inference procedure. Finally, the performances of the approach are assessed on synthetic examples and a real dataset.

2 Modeling

We assume that the phenomenon under study can be modeled by a second order random field of the form:

$$Z(x) = \mu(x) + Y(x),\ x \in \mathcal{X} \subset \mathbb{R}^d, \tag{1}$$

where $\mu(x)$ is a deterministic trend and $Y(x)$ is a centered second order random field with covariance function $\mathrm{Cov}(Y(x), Y(y)) = \mathrm{C}(x, y)$. More specifically, the trend can be expressed as a weighted sum of drift functions:

$$\mu(x) = \sum_{l=1}^{L} \beta_l f_l(x), \tag{2}$$

where $(\beta_l)_{l=1,\dots,L}$ and $(f_l)_{l=1,\dots,L}$ are deterministic and L is fixed. The drift functions are functions of the space, such as polynomials, and/or auxiliary variables accounting for a systematic behaviour of the phenomenon under study. The first drift function is generally the indicator of $\mathcal{X}$, accounting for a constant mean. A selection of the relevant functions can be performed preliminarily using standard variable selection techniques such as forward, backward, stepwise methods (see e.g. Saporta, 2006) or penalization approaches (see e.g. Hastie et al, 2009) if the candidates are many. This modelling is common in Geostatistics where the trend accounts for the large scale fluctuations while the second term accounts for smaller scale fluctuations occurring around the drift.

2.1 Reminder on kriging

Under these settings, given a set of observations $z = (z(x_1), \dots, z(x_n))^t$ from a realization of Z sampled at locations $(x_1, \dots, x_n) \in \mathcal{X}^n$, the universal kriging predictor provides the best linear unbiased predictor in the sense of the mean squared error at a new location x_0. It is a weighted sum of the observations:

$$Z^* = \Lambda^t z, \tag{3}$$

whose weights $\Lambda = (\lambda_1, \ldots, \lambda_n)$ are solution of the universal kriging system, straightforwardly obtained minimizing the mean squared prediction error under unbiasedness constraints:

$$\begin{pmatrix} C & F \\ F^t & 0 \end{pmatrix} \begin{pmatrix} \Lambda \\ \mu \end{pmatrix} = \begin{pmatrix} C_0 \\ f_0 \end{pmatrix} \tag{4}$$

where C is the covariance matrix associated to the observations, i.e. with entries $C_{ij} = \mathrm{C}(x_i, x_j)$, $(i,j) \in \{1, \ldots, n\}^2$, C_0 is the vector of covariances between $Z(x_0)$ and the observations, $f_0 = (f_1(x_0), \ldots, f_l(x_0))^t$,

$$F = \begin{pmatrix} f_1(x_1) & \ldots & f_L(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_n) & \ldots & f_L(x_n) \end{pmatrix},$$

and $\mu = (\mu_1, \ldots, \mu_L)^t$ is the vector of Lagrange multipliers accounting for the unbiasedness constraints.

The solution to this system is obtained by block inversion:

$$Z^* = C_0^t C^{-1} z + (f_0^t - C_0^t C^{-1} F)(F^t C^{-1} F)^{-1} F^t C^{-1} z, \tag{5}$$

where it can be seen that one matrix-vector and one matrix-matrix products involving the inverse covariance matrix are required. When n is large, as the number of operations involved in the computation of C^{-1} scales with n^3, the computation of Z^* is made intractable, unless a particular form for C is adopted. The prediction variance or kriging variance can also be computed as:

$$\begin{aligned} \mathbb{V}(Z(x_0) - Z^*) =& C(0) - C_0^t C^{-1} C_0 + \\ & (f_0 - F^t C^{-1} C_0)^t (F^t C^{-1} F)^{-1} (f_0 - F^t C^{-1} C_0), \end{aligned} \tag{6}$$

where $C(0) = \mathbb{V}(Z(x))$, corresponding to the punctual variance. We remind that, (5) and (6) are identical to the conditional expectation and conditional variance under the Gaussian hypothesis.

Two main families of approaches have been designed to overcome this limit and are desribed in the following subsections. We then develop an approach where they are combined.

2.2 Covariance tapering

Introduced in Furrer et al. Furrer et al (2006), covariance tapering is an approximation to the standard linear spatial predictor. The idea is to taper the spatial covariance function to zero beyond a certain range using a positive definite but compactly supported function. This terms to deliberately introduce zeros into the matrix C to make it sparse. The linear system (4) with a sparse (enough) covariance matrix can be solved efficiently.

How the zeros are introduced is crucial. In particular, the positive definiteness of any sparse modification of the covariance matrix must be maintained. Let C_τ be a covariance function that is identically zero outside a particular range described by τ. The tapered covariance function is the product of C_τ and the initial covariance function C:

$$\mathrm{C}_{tap}(x, y) = \mathrm{C}(x, y)\mathrm{C}_\tau(x, y) \tag{7}$$

The intuition behind this choice is both that the product C_{tap} preserves *some* of the shape of C (in particular, its behaviour near the origin) and that it is identically zero outside of a fixed range. C_{tap} is a valid covariance function, as the product of two positive definite functions. Finally, C_{tap} allows to build a sparse covariance matrix that can be inverted using sparse linear algebra techniques (e.g. sparse Choleski decomposition), to compute (4) and (6).

The theoretical justification of tapering relies on a result for kriging with misspecified covariance (Stein, 1993), stating roughly that as far the behavior at the origin of the covariance is well represented then the kriging predictor reaches optimality under infill asymptotics settings, that is when the prediction is performed on a bounded domain with increasingly denser data locations.

The main drawback of this approach is that it neglects the long range structure that the data may present. While this has negligible effect in densely sampled areas the effect can be significant in less sampled areas. Further more, the range of the taper τ is set such that the resulting covariance is sparse enough and needs to be reduced as the number of data increases. Consequently, the prediction accuracy can suffer from gaps between data locations when the sampling is not uniformly dense.

2.3 Low rank approaches

Many low rank approaches have been proposed over the past few years, among which we can cite the fixed rank approach (Cressie and Johannesson, 2008), predictive processes (Banerjee et al, 2008) and also more recent works (Katzfuss, 2017; Nychka et al, 2015). Precisely, under this framework the covariance matrix of the observed random vector can be written:

$$C = \sigma^2 I + P\,B\,P^t, \tag{8}$$

where I is the identity, P a known $n \times p$ matrix, B a $p \times p$ covariance matrix. Such a model naturally arises by considering a spatial process Z such that

$$Z(x) = \sum_{i=1}^{p} \eta_i\, P_i(x),$$

where the P_i are known basis functions and $(\eta_1, \ldots, \eta_p)^T$ is multivariate normal with mean 0 and covariance B.

p must be sufficiently small, so that the inverse of C can be computed. Indeed, C^{-1}, via the Woodbury-Morrison formula, takes the following form:

$$C^{-1} = \sigma^{-2} I - \sigma^{-2} P\{\sigma^2 B^{-1} + P^t P\}^{-1} P^t, \tag{9}$$

which only involves inverting the $p \times p$ matrices, ensuring a fast computation of (4) and (6), as far as p is small enough.

The inference the parameters σ^2 and B can be conducted by a method of moments (Cressie and Johannesson, 2008) or by maximum likelihood estimation through an expectation-maximization (EM) algorithm (Katzfuss and Cressie, 2011; Braham et al, 2017).

An important point is the choice of the basis functions P_i. A wide set of options is available: Cressie and Johannesson (Cressie and Johannesson, 2008) propose using bi-square functions, the W-wavelet basis functions are considered in (Nychka et al, 2002), Stein (Stein, 2008) uses a normalized Legendre polynomials basis on the sphere. In the proposed combination approach described below, we show that the choice of the basis functions can be data driven, using penalized least squares.

Finally, it is important to emphasize that this approach will neglect the small scale fluctuations of the signal of interest. Indeed, the fine scale fluctuations are represented by an unstructured noise. This generally results in a loss of prediction accuracy and oversmooth prediction maps.

2.4 Combination

Knowing the advantages and drawbacks of theses two approaches, it seems natural to try to combine them, so that the variability of the random field can be captured at all scales. This has been proposed in (Sang and Huang, 2012) where the remainder of a predictive process approximation is tapered.

We suggest here to proceed in the opposite way, first considering the small scale fluctuations before investigating the modeling of the larger scales.

The predictor can be computed for large datasets as the covariance matrix is of the same form as in (8), where the identity is replaced by a sparse matrix. More specifically, the form $C = A + PBP^t$, still allows to use the Woodbury-Morrison matrix identity that takes then the following form:

$$C^{-1} = A^{-1} - A^{-1}P(B^{-1} + P^tA^{-1}P)^{-1}P^tA^{-1}. \tag{10}$$

Furthermore, if A is sparse, P is $n \times p$ and B is $p \times p$, with $p \ll n$, the numerical complexity of (10) scales with $p^3 \times n$. Indeed, (10) involves only the inversion of the sparse matrix A and the $p \times p$ matrices B and $(B^{-1} + P^tA^{-1}P)$.

It is interesting to notice that the constraints imposed on the covariance model by its scalability with the number of data lead to the definition of a particular form of covariance function where the fluctuations of the random field are decomposed among several components:

$$\mathrm{C}(x, y) = \mathrm{A}(x, y) + P(x)^tBP(y) + \sigma_x\sigma_y\delta_{xy}, \tag{11}$$

where A is a continuous, compactly supported covariance function (it vanishes when $|x - y|$ is larger than a given, fixed range), B is a symmetric $p \times p$ positive definite covariance matrix, P is a set of basis functions and σ_x is a, possibly nonstationary, nugget effect.

Specifically, the A covariance function is built using the construction used in covariance tapering, namely as the product of any covariance function with a compactly supported one (e.g. the spherical model):

$$\mathrm{A}(x, y) = \mathrm{C}_\theta(x, y)\mathrm{C}_\tau(x, y), \tag{12}$$

where C_θ is any covariance function, possibly non stationary (Paciorek and Schervish, 2006; Fouedjio et al, 2016; Porcu et al, 2009) and C_τ is a compactly supported correlation function with a short fixed range (Gneiting, 2002). This construction induces the corresponding covariance matrix A to be sparse, with a density controlled by the fixed scale parameter of C_τ. This construction is the same as tapering (Furrer et al, 2006) but its use is somehow different: rather than bulding a compactly supported approximation of a non compactly supported covariance function, we only use a compactly supported covariance function to model the small scale fluctuation.

The first and last term of (11) allow to model the short scale fluctuations of the phenomenon under study which is necessary to interpolate accurately the data in densely sampled areas. Indeed, as will be seen in section 4, the low rank term only in (11) fails to capture the fine scale structure. Moreover, for most classes of covariance models that is when the sampling becomes denser and denser in a fixed domain, the quality of the kriging interpolator is essentially guided by the short scale behaviour (Stein, 1999).

The second term of (11) allows to capture the intermediate scale fluctuations, between those modeled by A and by the drift. This representation is non stationary by construction. The scale of the fluctuations modeled by this term is determined by the choice of the basis functions constituting P. Therefore, they must be chosen with care and not too numerous so as to keep the inversion numerically attractive.

Finally, the nugget component of (11) is classically used to model a measurement error, for instance due to the internal variability of the sensor used, or to model a non-continuous behaviour of the regionalized variable, hence the term "nugget". We can allow the value of the nugget to vary smoothly across the domain of interest and filter it out at observation locations in prediction.

2.5 Random field model

Assuming the covariance model of (11), the random field model can be rewritten as follows:

$$Z(x) = \mu(x) + \sum_{i=1}^{p} \eta_i P_i(x) + S(x) + \sigma_x \varepsilon(x), \tag{13}$$

where $\eta = (\eta_1, \ldots, \eta_p)^t$ is a vector of centered, square-integrable random variables with covariance matrix B, $(P_i)_{i=1,\ldots,p}$ is a set of basis functions, S is a centered second order random field with covariance A (12), and $\varepsilon(x)$ is a centered, square-integrable, non structured random function with variance 1. Moreover, to respect the model (11) S, η and ε have to be uncorrelated. Z can hence be decomposed into four terms, each corresponding to a different scale of fluctuations.

2.6 Which basis functions?

In low rank approaches, the basis functions are usually set in advance. As said in the introduction, there is a wide set of options for the basis functions to be used but the choice remains arbitrary. We propose a data driven way to select the basis functions among a vast family of candidates, possibly including anisotropies.

First note that equation (13) can be rewritten as

$$Z(x) - \mu(x) = \sum_{i=1}^{p} \eta_i P_i(x) + Y(x), \tag{14}$$

where $Y(x) = S(x) + \sigma_x \varepsilon(x)$ has only short spatial correlations. In other words, we can express the detrended data as a regression model with a random effect. We propose therefore to operate a data driven selection of basis functions from a large set of candidates, using variable selection methods currently used in variable selection problems such as the LASSO (Tibshirani, 1996). This point will be further detailed in the next section.

Thinking about the Karhunen-Loève expansion (Loève, 1955) makes this idea natural. Indeed, any centered second order random field admits the following representation:

$$\sum_{i \in I} \xi_i \lambda_i \phi_i(x), \tag{15}$$

where I is at most countable, $(\xi_i)_{i \in I}$ are standard uncorrelated random variables, $(\lambda_i)_{i \in I}$ is the set of the of the covariance function of the field and the $(\phi_i)_{i \in I}$ is the set of the associated eigenvectors. The rate of decay of the eigenvalues is fast for covariances with smooth behavior, see e.g. (Romary and Hu, 2007; Romary, 2009). Conversely, when considering a realization of a Gaussian random field, most of the ξ_i take values close to zero in the representation (15), as standard Gaussian random variables. Consequently, only a small number of components are "relevant" in the representation of the realization. Consequently, modeling this scale of variation with a low number of carefully selected basis functions seems to make sense.

3 Inference and prediction

In this section, we will assume that $\mu(x) = 0$ without loss of generality. Indeed, the ordinary least square (OLS) estimate of the trend, can be computed and subtracted from the data to get a centered dataset. In the context of a large spatial dataset, the additional variance due to the use of a non efficient estimator is generally negligible.

Table 1: Inference steps

1.	Fit the parameters of the small scale component (range and variance) by variogram fitting
2.	Select the basis functions by the LASSO
3.	Estimate B and σ^2 with fixed P by ML
4.	Change the penalty in 2. and update the selection
5.	Estimate B with fixed P and σ^2 by ML

The parameters to be inferred are the covariance matrix B, the parameters of the covariance function A and the nugget σ_x. We describe here a step by step inference method. We also propose an approach to select the most relevant basis functions P among a large dictionary of candidates.

We will first consider that the covariance $\sigma^2 \mathrm{A}_\theta(x,y) = \sigma_\phi^2 \mathrm{R}(x,y) + \sigma_x \sigma_y \delta_{xy} = \mathrm{Cov}(S(x)+\varepsilon(x), S(y)+\varepsilon(y))$, where θ is the set of small scale covariance parameters, including σ_ϕ^2, the parameters the correlation function R and σ_x. In the stationary case, $\sigma_x = \sigma_\varepsilon$ is constant over the domain. In other words, we will group the nugget and the short scale fluctuations term, as the inference related to these two terms will be conducted conjointly. The special case of an observation error modeled by a nugget effect will be dealt with in a dedicated paragraph.

We consider a large vector of data $z = (z(x_1), \ldots, z(x_n))$. Under Gaussian hypothesis, the full log-likelihood writes

$$\begin{aligned} L_z(\sigma^2, \theta, B) = &-\frac{1}{2} \ln \det(\sigma^2 A_\theta + P^t B P) \\ &+ \frac{1}{2\sigma^2} z^t \left(A_\theta^{-1} - \frac{1}{\sigma}^2 A_\theta^{-1} P (B^{-1} + \frac{1}{\sigma}^2 P^t A_\theta^{-1} P)^{-1} P^t A_\theta^{-1} \right) z. \end{aligned} \tag{16}$$

Maximizing directly the log-likelihood function is not straightforward and cannot be computed analytically. Making appear the coefficients η of the basis functions as latent variables allows to decompose the likelihood, indeed:

$$\begin{aligned} \eta &\sim \mathcal{N}(0, B) \\ Z|\eta &\sim \mathcal{N}(P^t \eta, \sigma^2 A_\theta). \end{aligned}$$

Consequently, an EM algorithm can be implemented to estimate the parameters. Nevertheless, to do this, the basis functions need to be known beforehand. One solution could be to add a penalty term in (16) but this would only have add more complexity. To simplify the parameter inference, we therefore rely on a step by step algorithm where we treat the different scales separately, thus lowering the computational burden. The several consecutive steps are presented in table 1 and detailed in the following paragraphs.

We could have stopped at step 3. in table 1 while keeping a fixed σ^2 while estimating B. Estimating both σ^2 and B within the EM step aims at balancing the explained variations between the small scale component and the basis functions, while mitigating the possible parameter estimation errors caused by variogram fitting. Indeed, the fitted variogram of the short scale component try to explain the whole observed variability, at the considered lags, without taking into account the additional basis functions.

Consequently, the penalty is generally set at a too large value at first guess and the number of basis functions selected is too low. Updating the value of σ^2 within the EM algorithm allows assigning a better weight to the short scale variations. A new selection of basis functions can then be performed in accordance with this new weight and the covariance matrix of its random coefficients is then estimated. It is possible to iterate the process, updating σ^2 and redoing a selection, but this does not prove useful in practice.

3.1 Inference of the small scale structure

The first step of the inference approach consists in fitting the, possibly non stationary, small scale covariance structure A_θ.

When considering a stationary compactly supported covariance to model the small scale structure, several options are available for the estimation of the parameters. Maximum likelihood can be considered as well as variogram fitting. The latter proves to be less computer demanding and will therefore be used in the application section.

In presence of short scale non stationarity, a practical way to infer non stationary covariance structures of the form given in (Paciorek and Schervish, 2006) consists in fitting stationary models locally as proposed in (Fouedjio et al, 2016) or in (Parker et al, 2016). Interpolating the fitted value of the parameters over the domain of interest allows to compute the value of the covariance function between any two points of the domain, which is necessary to compute the kriging interpolator. This approach only requires that the parameters governing the non-stationarity vary regularly over the domain, precisely the associated random function, S in our model, can be considered to be locally second order stationary. This approach can however be time consuming, though an embarrassingly parallel workload, and the benefits offered by considering a non stationary small scale structure may not be worth the trouble.

Recalling that the short scale structure should provides a sparse matrix, the range of the compactly supported correlation function, the taper, has to be small enough. We refer the interested reader to (Furrer et al, 2006) for the choice of the taper. Basically, it has to be chosen at least as smooth at the origin as the original model.

Once the parameters of this structure have been inferred, the remaining second order variability will be modeled using basis functions. Indeed, the short scale structure explains a given part of the variations of the phenomenon under study and this will guide the choice of the basis functions to be used.

3.2 Selection of the basis functions

To estimate the matrix B, the basis functions constituting P have to be known. Numerous families of functions are available: local bi-square functions (Cressie and Johannesson, 2008), wavelets (Nychka et al, 2002), Legendre polynomials (Stein, 2008), smoothing splines, radial basis functions, cosines, etc.

No clue exists however to choose the correct one for the problem at hand, unless a covariance model has already been fitted and predictive process can be considered (Banerjee et al, 2008; Katzfuss, 2017).

To alleviate this choice, we propose to select the most relevant basis functions among a large choice of candidates. This approach has already been proposed in (Hsu et al, 2012), considering however that the random variables $(\eta_i)_{i=1,\dots,p}$ associated to each basis function are independent, resulting in a diagonal B, which seems too strong an assumption. The estimation of B, the covariance matrix of the $(\eta_i)_{i=1,\dots,p}$, is treated in the next paragraph.

Here we propose as basis functions a collection of kernels anchored at a set of knots homogeneously distributed over the domain. We choose to use compactly supported covariance kernels for several reasons:

- we can build a multiresolution representation by considering several scale parameter values,

- we can easily introduce anisotropic basis functions,
- we can consider several levels of regularity at the origin (e.g. spherical, wendland, cubic, etc. kernels),
- compact support makes the storage lighter and the prediction faster.

Practically, we consider a handful of values for the scale parameter and orientation of the anisotropy. To each value of the scale parameter corresponds a grid where the kernel is anchored. Scale parameter values and orientations are crossed so as to generate anisotropies. The scale parameter values are meant to correspond to ranges greater than the one considered in the small scale structure, so as to achieve scale separation between the two components.

Going back to (14), the problem of selecting the relevant basis functions among a dictionary of candidates terms to solve the following penalized regression problem:

$$\arg\min_{\beta} \|(z - \widehat{\mu}) - P\beta\|_2^2 + \lambda\|\beta\|_1, \tag{17}$$

where β is a vector of estimates of the $(\eta_i)_{i=1,\dots,p}$ and λ is a penalty term. The rationale for this is that if the covariance of the random field admits a low rank representation, then the random field can be expressed as a random linear combination of the basis functions. Solving (17) makes it possible to select the active components in this representation. We choose to use a L_1 penalty so as to proceed to a more drastic selection than that would have been obtained with the L_2 norm (Hastie et al, 2009).

Practically, (17) is minimized though a coordinate descent algorithm such as implemented in the `R` package `glmnet` (Friedman et al, 2010). This type of algorithm has proven to be numerically efficient for solving convex optimization problems such as (17). The value of the penalty term is set by cross-validation. As we already know, at this step, the variance σ^2 explained by the small scale structure, we propose to set the penalty at the value necessary to explain the remaining variance. Further more, as the regularization path is computed for a vector of penalty values, we do not need to recompute the LASSO when the value of σ^2 is updated. This will be further explained in section 4.

It is worth noting that we are not interested in the estimated values of β, rather by the relevance of the associated basis function in the explanation of the fluctuations of $z - \mu$. Furthermore, as the remaining terms S and ε deal with smaller scale structure, they do not interfere in this selection procedure. For huge datasets, the computational effort of this step can be alleviated using stochastic gradient algorithms where the optimization is conducted iteratively over subsets of data, see e.g. Bottou (2010).

3.3 Inference of the medium scale structure

Once the p basis functions have been selected, the covariance matrix of their components has to be estimated and the weight of the small scale structure updated. As proposed in (Katzfuss and Cressie, 2011), we choose to perform maximum likelihood estimation through an expectation-maximization (EM) algorithm. The EM objective function writes:

$$\begin{aligned} Q(\Theta, \Theta_l) = &- \frac{n+p}{2}\ln(2\pi) - \frac{1}{2\sigma^2} z^t A_\theta^{-1} z + \frac{1}{\sigma^2} z^t A_\theta^{-1} P\mu - \frac{1}{2}\ln\det(\sigma^2 A_\theta) \\ &- \frac{1}{2}\mathrm{Tr}\left((\frac{1}{\sigma^2} P^t A_\theta^{-1} P + B^{-1})(C + \mu\mu^t) \right) - \frac{1}{2}\ln\det(B), \end{aligned} \tag{18}$$

where $\mathrm{Tr}(X)$ is the trace of matrix X, $C = \mathbb{V}(\eta|Z = z, \Theta_l) = \left(\frac{1}{\sigma_l^2} P^t A_\theta^{-1} P + B_l^{-1}\right)^{-1}$ and $\mu = \mathbb{E}(\eta|Z = z, \Theta_l) = \frac{1}{\sigma_l^2} C P^t A_\theta^{-1} z$.

The update formulas for the parameters σ^2 and B are:

$$\begin{aligned}\sigma^2_{\text{new}} &= \frac{1}{n}\mathbb{E}\left((Z-P\xi)^t A_\theta^{-1}(Z-P\xi)|Z,\Theta_l\right)\\ &= \frac{1}{n}\left(Z^t A_\theta^{-1} Z - 2Z^t A_\theta^{-1} P\mu + \text{Tr}((P^t A_\theta^{-1} P + B^{-1})(C+\mu\mu^t))\right), \qquad (19)\\ B_{\text{new}} &= C + \mu\mu^t. \qquad (20)\end{aligned}$$

The algorithm is initialized with the value of σ^2 obtained at the first step and B as the identity and run until a stopping criterion is met. The stopping criterion we use is defined as:

$$\min\left(|\sigma^2_{l+1}-\sigma^2_l|, \frac{|\sigma^2_{l+1}-\sigma^2_l|}{\sigma^2_{l+1}}\right) + \min\left(\left|\log\left(\frac{\det(B_{l+1})}{\det(B_l)}\right)\right|, \left|\frac{\log\left(\frac{\det(B_{l+1})}{\det(B_l)}\right)}{\log\det(B_{l+1})}\right|\right). \qquad (21)$$

We decide to stop the algorithm when the criterion is less than a given tolerance value for 20 iterations in a row.

3.4 Update of the selection and final model

In the previous step, the amount of variability explained by the small scale structure modeled by σ^2 is updated. It can be lowered or increased depending on the data. It is thus possible, and sometimes desirable, to update the selection of the basis functions accordingly.

As said earlier, the whole regularization path is computed in the selection step. Consequently, when the value σ^2 is updated in the EM algorithm, there is no need for any further computation to update the selection. The penalty value has only to be set so that the basis functions explains the updated remaining variance.

Then, the EM algorithm is run once again to get the final estimate of the matrix B, while the value of σ^2 is left unchanged. It is possible to iterate the process, updating σ^2 and redoing a selection, but this does not prove useful in practice.

3.5 Prediction

It is worth noting that most of the quantities needed for the prediction have already been computed in the EM step. Indeed, under the Gaussian hypothesis, the kriging predictor is identical to the conditional expectation and the kriging variance to the conditional variance. These quantities can be straightforwardly computed knowing $\mathbb{P}(\eta|z)$. Consequently the kriging predictor at a new location x_0 writes:

$$\mathbb{E}(Z(x_0)|z) = P_0^t\mu + A_0 A^{-1}(z - P^t\mu), \qquad (22)$$

where S_0 is the small scale component value at x_0, $\mu = \mathbb{E}(\eta|z)$, P_0 is the vector of the basis functions at x_0 and A_0 is the covariance of the small scale structure between the target and the observation. Similarly, the kriging variance writes:

$$\mathbb{V}(Z(x_0)|z) = \sigma^2 - A_0 A^{-1} A_0 - 2A_0 A^{-1} P^t C P_0 + P_0^t C P_0, \qquad (23)$$

where $C = \mathbb{V}(\eta|z)$. These formulas provides the most efficient way to compute these quantities. They only require to compute the cholesky decompositions of the sparse matrix A and the small matrix C. It is interesting to note that although the covariance model is built so that we can use the Woodbury-Morrison formula to solve the kriging system, this formula is not used explicitly to compute the predictor and the prediction variance.

3.6 Filtering a measurement error

When an observation error exists, it is important to filter it out from the prediction, including at observation locations. In that case, the observation error is modeled by a nugget effect. A first idea could be to fit the variance of the nugget at the variogram fitting step and to keep it fixed in the update of the variance term in the EM step. In other words, we would have $\sigma^2 \mathrm{A}_\theta(x,y) = \sigma^2_\phi \mathrm{R}(x,y) + \sigma^2_\varepsilon \delta_{xy}$, where σ_ε is fixed. When doing this unfortunately, there is no closed form expression for the update formula of σ^2. Therefore a numerical optimization is needed at each iteration of the EM, which slows down the calculations.

Here we propose a alternative scheme. We first rewrite the small scale covariance as $\sigma^2 \mathrm{A}_\theta(x,y) = \sigma^2_\phi \left(\mathrm{R}(x,y) + \frac{\sigma^2_\varepsilon}{\sigma^2_\phi} \delta_{xy} \right)$. Then, we fix the value of $\frac{\sigma^2_\varepsilon}{\sigma^2_\phi}$ as its first estimates $\frac{\widehat{\sigma^2_\varepsilon}}{\widehat{\sigma^2_\phi}^{\text{old}}}$ prior to go through the inference steps described above, where we get a new estimate $\widehat{\sigma^2_\phi}^{\text{new}}$. Finally, as the nugget variance is generally correctly fitted at the variogram fitting step and we do not want to change it, we reinject it in the estimated small scale covariance so that it takes the following form:

$$\widehat{\sigma^2} \mathrm{A}_{\widehat{\theta}}(x,y) = \widehat{\sigma^2_\phi}^{\text{new}} \left(1 + \frac{1}{\widehat{\sigma^2_\phi}^{\text{old}}} - \frac{1}{\widehat{\sigma^2_\phi}^{\text{new}}} \right) \mathrm{R}(x,y) + \widehat{\sigma^2_\varepsilon} \delta_{xy}. \tag{24}$$

By doing this, we keep constant the nugget effect and update the weight of the small scale component. We can reasonably assume that this simplification has a negligible effect on the computation of the likelihood in the EM algorithm.

Finally, to filter the error from the observation, the kriging is performed considering the following small scale covariance matrix:

$$\widehat{\sigma^2_\phi}^{\text{new}} \left(1 + \frac{1}{\widehat{\sigma^2_\phi}^{\text{old}}} - \frac{1}{\widehat{\sigma^2_\phi}^{\text{new}}} \right) \mathrm{R}(x,y). \tag{25}$$

4 Application

In this section, we provide experiments of the proposed method and comparison with existing approaches, on two simulated and one real datasets. The different approaches have been implemented in `R` with an extensive use of the packages `spam` (Furrer and Sain, 2010), `glmnet` (Friedman et al, 2010) and `RGeostats` (Renard et al, 2010). The computations have been performed a laptop computer with 3rd generation Intel Core i7-3687U processor (year 2013) and 8 GB of memory.

4.1 Nested covariance

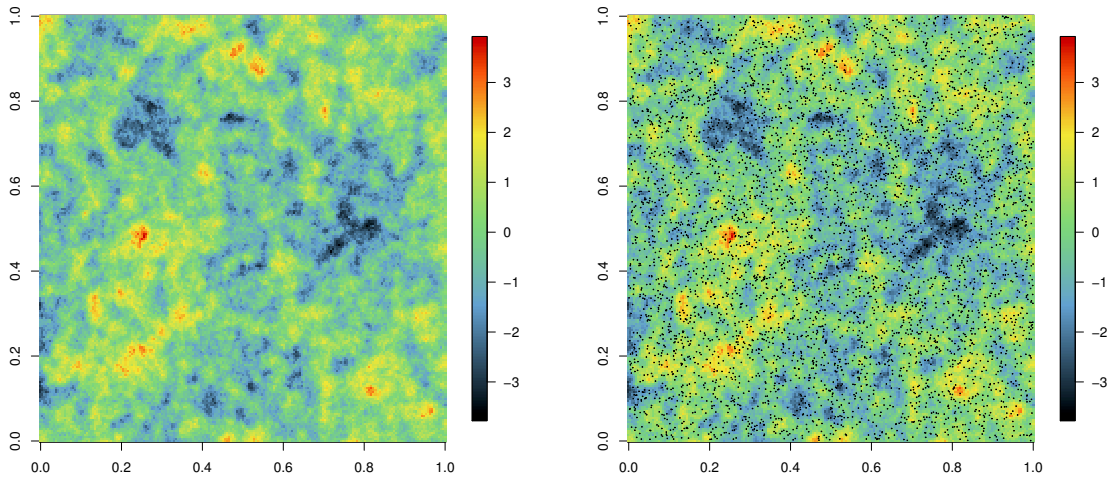
In this first example, we consider a simulated Gaussian random field (GRF) on $[0,1]^2$, centered with covariance

$$\mathrm{C}(x,y) = 0.5\mathrm{sph}(|x-y|, 0.05) + 0.5\exp(|x-y|, 0.1), \tag{26}$$

where sph stands for the spherical covariance function, exp for the exponential covariance function and the arguments are the distance and the scale parameter. This kind of nested covariance is often used in practice to model variables that exhibit variations at two different scales

The GRF is simulated on a 200x200 grid and sampled at 5000 points, as shown in figures 1a and 1b

From this dataset, we infer the combination model described in paragraph 2.4 following the steps detailed in 3. The first step consits in computing the empirical variogram at short distances and fitting



(a) Reference realization

(b) Sampling at 5000 locations

Figure 1: Reference simulation and sampling

a tapered variogram with a small range. This step is illustrated in figure 2a. The structured inferred for the small scale component is:

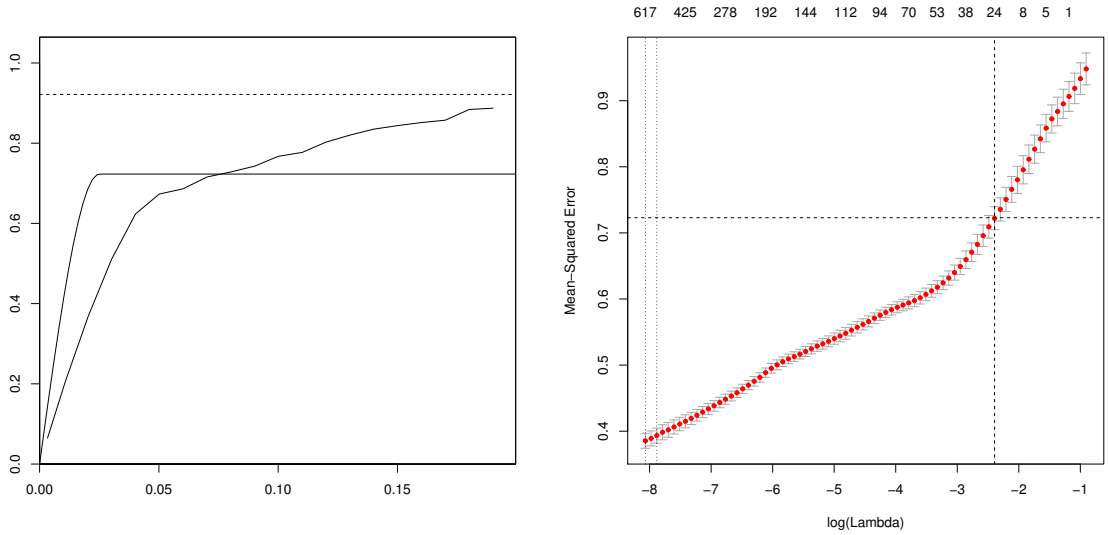
$$A(x, y) = \sigma^2 \exp_\theta(x, y) \mathrm{sph}_{0.025}(x, y) \tag{27}$$

The quantity σ^2 can be interpreted as the amount of variability explained by the short scale structure. As can be seen, from figure 2a, this value is overestimated but will be updated later. To explain the remaining variability, we will select a set of basis functions from a large set of candidates. The dictionary of candidates is built combining several ingredients: two kinds of compactly supported basis functions built from the cubic and spherical covariance models, two ranges are considered for each one, 0.5 and 0.2, and four different anisotropy angles 0, $\pi/4$, $\pi/2$, $3\pi/4$. The basis functions are computed over knots forming two different grids whose discretization step is set according to the range of the basis function.

This makes a total of 2658 candidates basis functions. The relevant functions are selected by the LASSO using the `cv.glmnet` function of the package `glmnet` (Friedman et al, 2010). Using this function, a 5-fold cross-validation is performed so as to assess the stability of the selection procedure. Then, the penalty is set so that the basis functions explain the remaining variance, as shown in figure 2b. We can see that 24 basis functions are selected as a first guess.

The next step consists in running the EM algorithm to estimate the covariance matrix B of the random coefficients of the basis functions and update the value of σ^2. This allows to adjust the part of the variations explained by the small scale component. The EM algorithm is run for this selection until the stopping criterion is met, with a tolerance value set at 10^-3. The updated variogram is shown in figure 3a, where we can see that the fit is now more precise at small distances. A new set of basis functions is selected accordingly, that represents 287 basis functions, as can be seen in figure 3b.

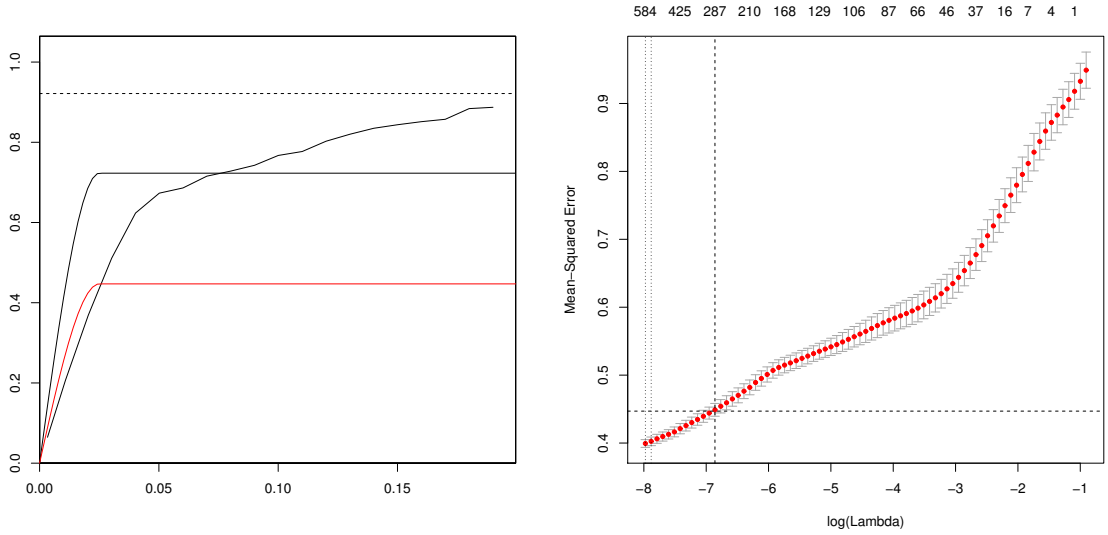
The EM algorithm is run once again with the new set of basis functions to estimate B, while the value of σ^2 is kept fixed. Then the kriging predictor can be computed on the 200x200 discretization grid of the unit square. We also computed the map obtained using the tapering term only (27), the basis functions only, where these are selected setting the penalty value at the largest value such that the error is within 1 standard error of the minimum (560), and the conditional expectation obtained using the true covariance. The kriging maps are depicted in figure 4.



(a) First fitting of the small scale structure

(b) First selection of basis functions

Figure 2: First fitting of the model



(a) Updated variogram

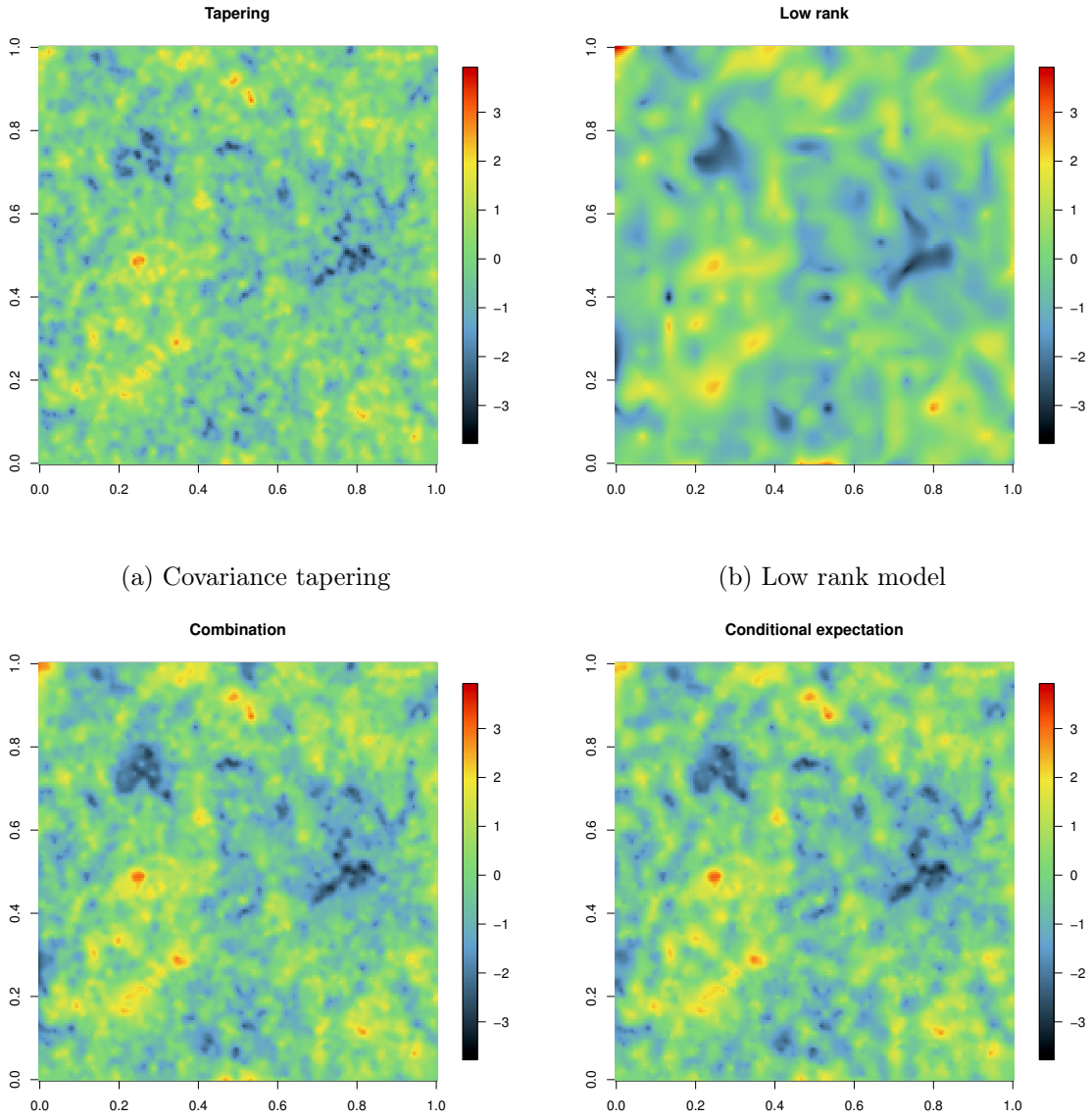
(b) Updated selection of basis functions

Figure 3: Model update

While there are few visible differences between the maps obtained with covariance tapering, the combination and the true model, it can be seen in figure 4b that the map obtained with the low rank model is much smoother. This is due to the bad representation of the small scale fluctuations using only the basis functions. In that case, the small scale fluctuations are modeled by an unstructured noise whose variance is close to the σ^2 term estimated in the combination approach.

We also computed the difference maps between the three fast approaches and the conditional expectation (figure 5).

Finally, we compute the mean squared prediction error (MSPE) obtained by the four methods. They are shown in table 2, together with the details of each model. We can notice here that the



(a) Covariance tapering

(b) Low rank model

(c) Combination

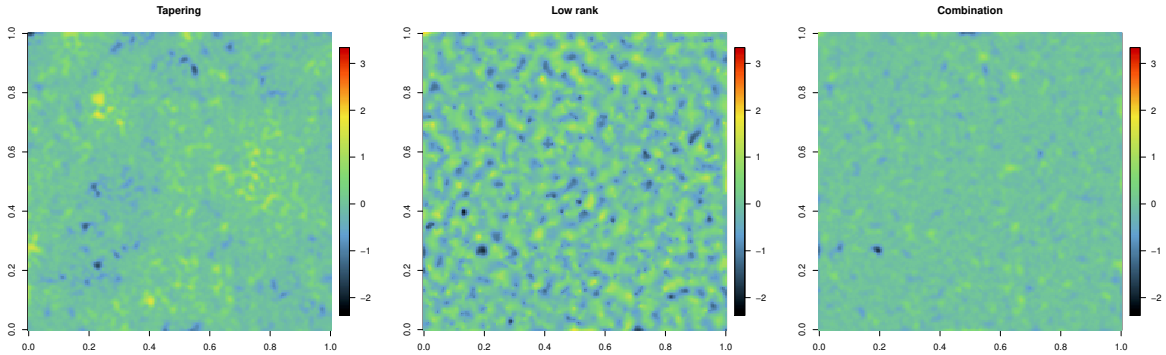
(d) Conditional expectation

Figure 4: Kriging maps

low rank approach obtains the worse result among the different methods implemented, while using a large number of basis functions. As explained earlier, it fails at capturing the small scale fluctuations, modelling them as a nugget effect with a large variance and therefore filters out this information while predicting, resulting in an oversmooth kriging map.

4.2 Non stationary Matérn covariance

Recently, new non stationary covariance constructions have been proposed, see e.g. (Fouedjio et al, 2016; Porcu et al, 2009). These models allow to make vary the anisotropy or even the regularity of the random field over the domain considered. The inference of such model relies on local methods (local variogram fitting or local likelihoods) combined with an interpolation of the locally fitted parameters. This procedure turns out to be computationally intensive. The resulting dense covariance matrices are



(a) Covariance tapering (b) Low rank model (c) Combination

Figure 5: Difference maps

Table 2: Summary

model	small scale structure	number of basis functions	MSPE
Covariance tapering	range 0.025	0	0.24
Low rank	nugget 0.40	560	0.35
Combination	range 0.025	287	0.19
Conditional expectation			0.17

also cumbersome to compute. For our second example, we consider a GRF on $[0,1]^2$, centered, with covariance

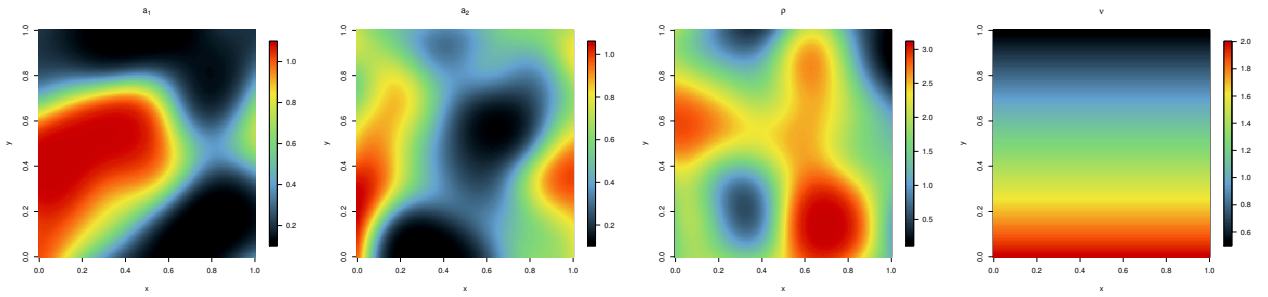
$$\mathrm{Cov}(Z(x), Z(y)) = \phi_{xy} \frac{2^{1-\nu(x,y)}}{\sqrt{\Gamma(\nu(x))\Gamma(\nu(y))}} M_{\nu(x,y)}\left(\sqrt{Q_{xy}(x-y)}\right), \tag{28}$$

where Σ_x is a positive definite matrix accounting for the local anisotropy at location x, $\phi_{xy} = |\Sigma_x|^{1/4}|\Sigma_y|^{1/4}\left|\frac{\Sigma_x+\Sigma_y}{2}\right|^{-1/2}$, $Q_{xy}(h) = h^t\left(\frac{\Sigma_x+\Sigma_y}{2}\right)^{-1}h$, $M_{\nu(x,y)}(h) = h^{\nu(x,y)}K_{\nu(x,y)}(h)$, $K_{\nu(x,y)}$ is the modified Bessel function of the second kind and $\nu(x,y) = \frac{\nu(x)+\nu(y)}{2}$.

A simulation method for this type of covariance has been recently proposed in (Emery and Arroyo, 2017). The simulation is obtained as a weighted sum of cosine waves, with random frequencies computed from an instrumental stationary spectral density and random phases. A location specific importance weight is applied so as to respect the local spectral density of (28).

To build our example, we first simulate a map of varying parameters over the unit square. The two scale parameters and the anisotropy angle are simulated as uniform transforms of a Gaussian random field with a gaussian covariance and a large scale parameter. The scale parameters are comprised between 0.1 and 1.1, while the anisotropy angle is allowed to vary between 0 and π. The smoothness parameter varies linearly between 0.5 and 2 along the vertical direction. The varying parameters are plotted in figure 6.

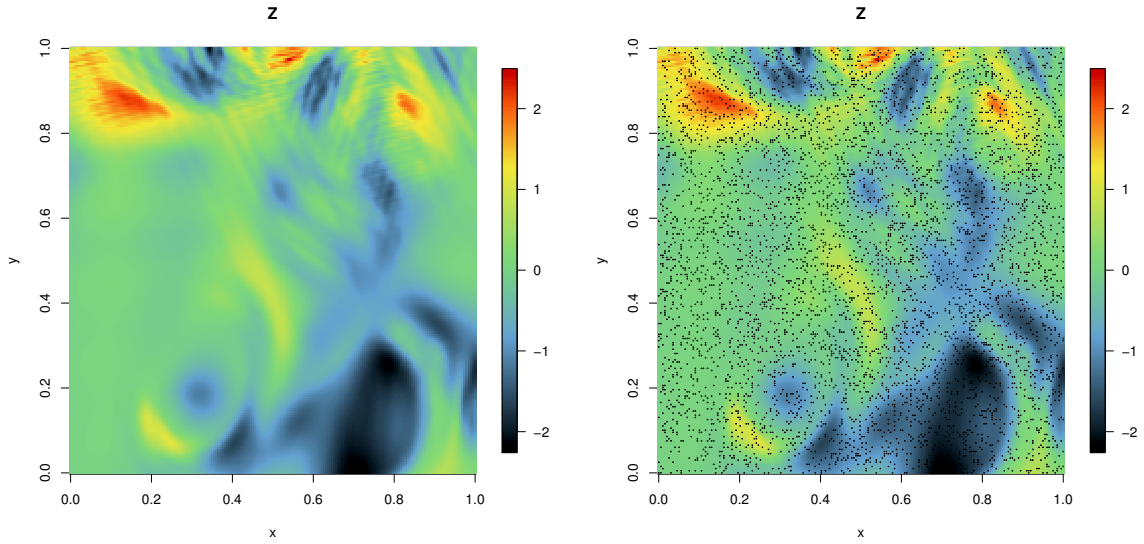
10000 cosine waves are used to generate the realisation plotted in figure 7. We can clearly see the effect of the varying anisotropy and smoothness parameters. Besides, some artifacts appear at the top of the domain, where the smoothness parameter is the lowest. They are due to a too small number of cosines in the simulation, which implies a incomplete sampling of the highest frequencies. This is not an important problem in this example however as these artifacts will be treated as patterns in the original image. This realisation is sampled at 5000 locations uniformly distributed over the domain,



(a) First scale parameter (b) Second scale parameter (c) Anisotropy angle (d) Smoothness parameter

Figure 6: Covariance parameters

as in the previous example.



(a) Reference realization (b) Sampling at 5000 locations

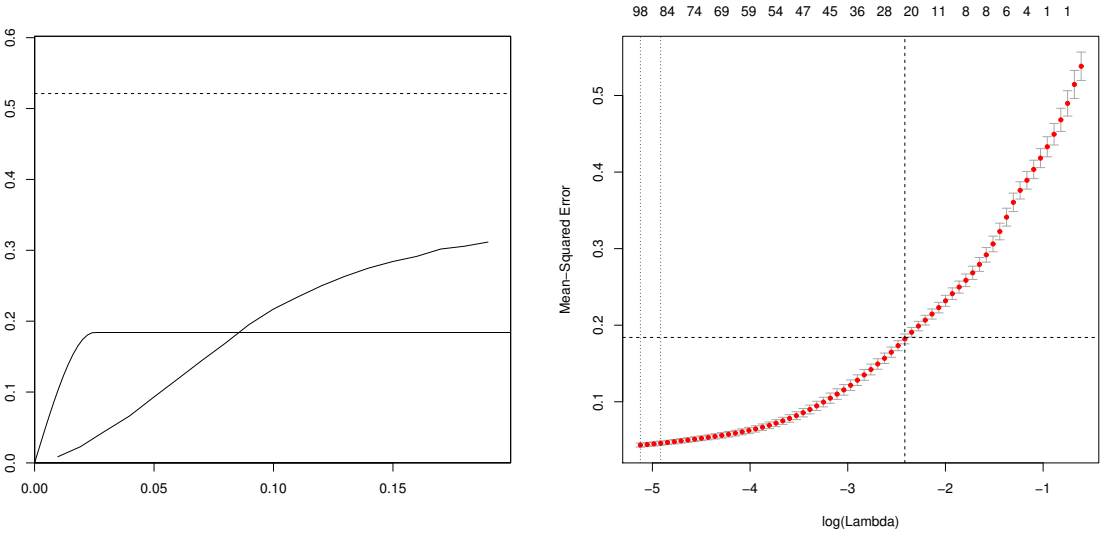
Figure 7: Reference simulation and sampling

The inference is conducted following the same steps as in the previous example. Moreover, the same set of 2658 basis functions is used to constitute the dictionary of candidates.

We can see in figure 8a that the variogram fit is not good, which is due to the low value for the range of the taper compared to what should be the range. In particular, a too large weight is associated to the small scale component, which results in the selection of only 23 basis functions.

After the EM algorithm, the updated value of σ^2 provides a better fit of the variogram and 69 basis functions are now selected.

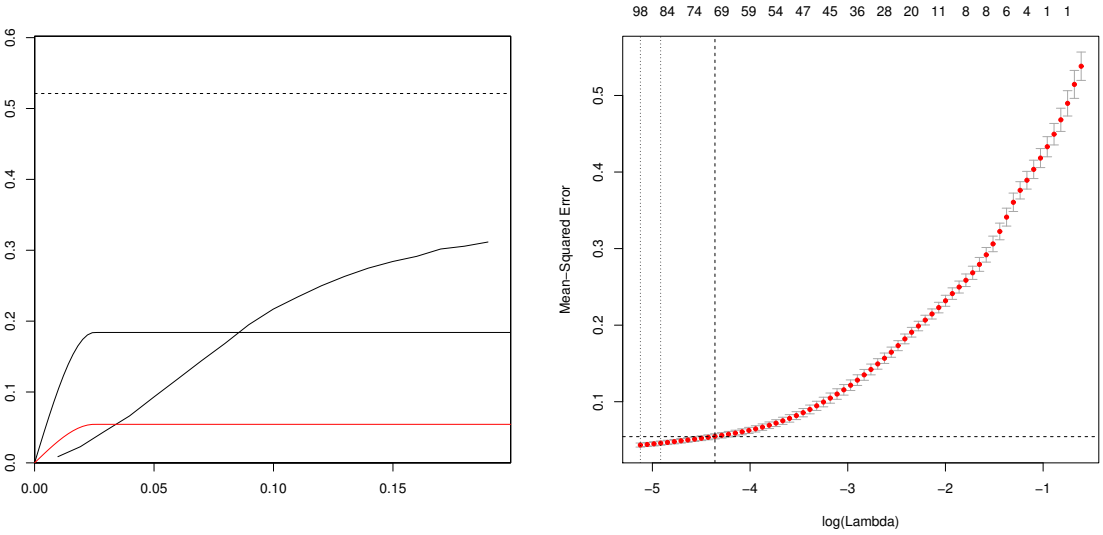
The difference in the kriging maps shown in figure 10 are clearly visible. The small value chosen for the tapering range causes the prediction to be equal to the mean when the distance from observation locations increase. The low rank model provides on the contrary an oversmooth map. The combination allows capturing both scales of variation and offers a result closer to the conditional expectation. In terms of mean square prediction error (MSPE), an important improvement is obtained with the combination as can be read in table 3.



(a) First fitting of the small scale structure

(b) First selection of basis functions

Figure 8: First fitting of the model



(a) Updated variogram

(b) Updated selection of basis functions

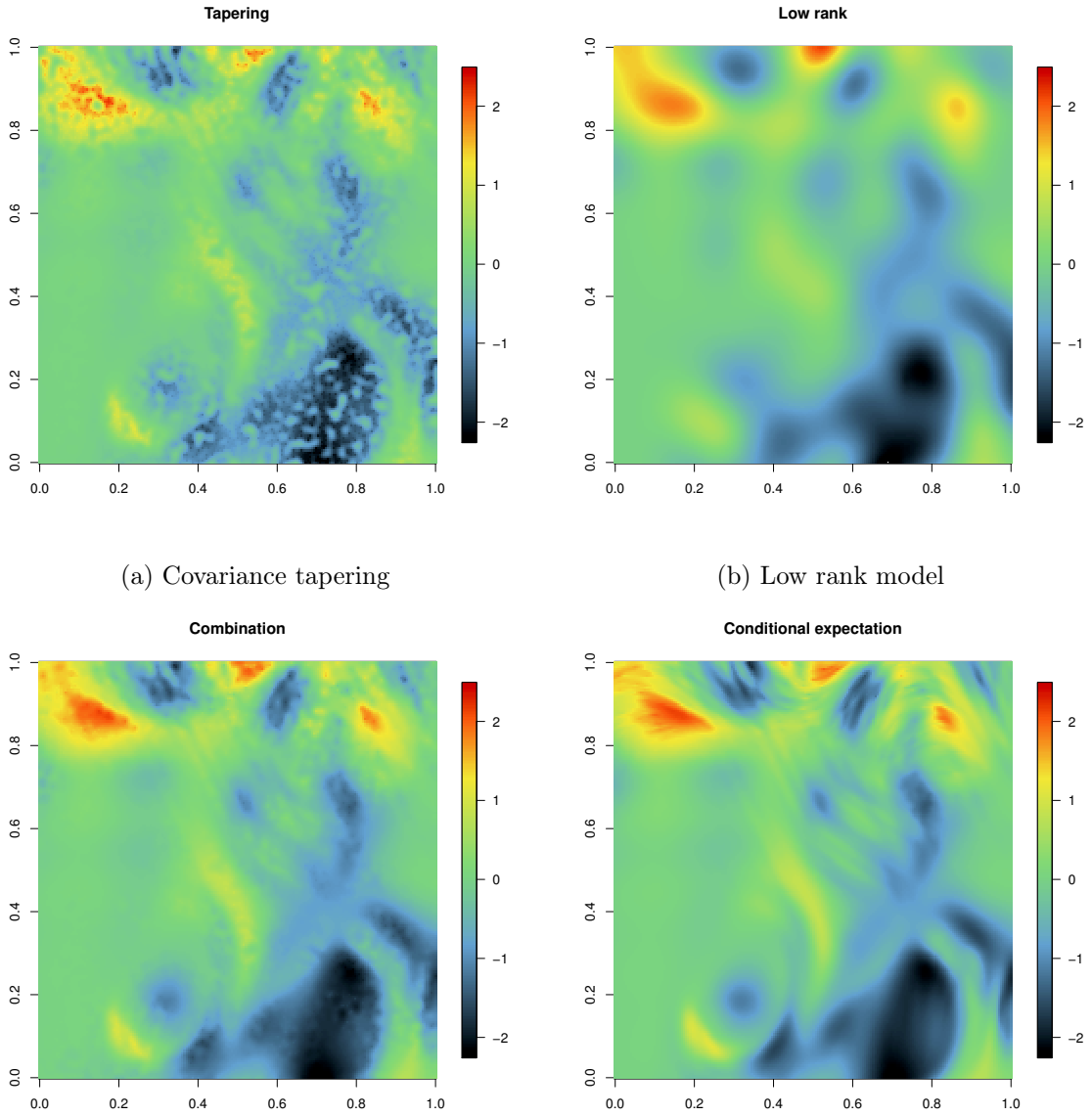
Figure 9: Model update

4.3 Total column ozone dataset

We now apply our approach to total column ozone (TCO) data acquired from an orbiting satellite mounted with a passive sensor registering backscattered solar ultraviolet radiation. The dataset we consider here has been previously analyzed in Cressie and Johannesson (2008), Bolin and Lindgren (2011) and Eidsvik et al (2014) among others. It consists of $n = 173\ 405$ measurements.

We compare covariance tapering, fixed rank kriging (FRK) and the proposed combination approach.

The predicition is performed on a 180 $\times$ 288 grid. The latitude ranges from -89.5 to 89.5 in 1° steps, while longitude ranges from -179.375 to 179.375 in 1.25° steps. This represents 51 840 prediction



(a) Covariance tapering

(b) Low rank model

(c) Combination

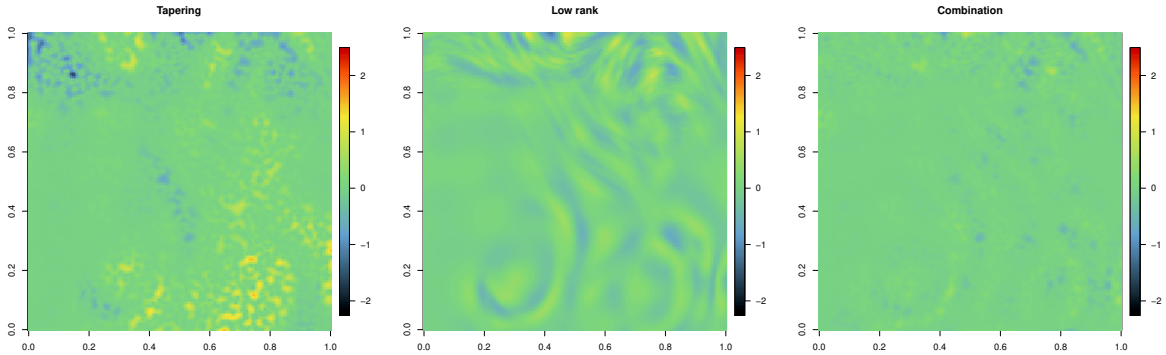
(d) Conditional expectation

Figure 10: Kriging maps

sites in total.

Covariance tapering is implemented using a Wendland2 (Gneiting, 2002) covariance with a taper range set at 275 km, which represents a 2.5° distance. Indeed, the data exhibit a rather smooth behavior that can be well represented by a regular covarariance function. The parameter estimation is performed through variogram fitting. FRK is performed with 4 resolution bi-square basis functions, representing a total of 3608 basis functions. The combination is built using the same fine scale structure as in covariance tapering and by selecting the basis functions among those used in FRK. 331 basis functions are selected. The EM step is run only once since the first fit of the variogram is deemed satisfying.

Figure 12 shows the prediction maps of TCO and the prediction standard deviations obtained by each three methods. We can observe that FRK predictions (fig. 12c) are much smoother than



(a) Covariance tapering (b) Low rank model (c) Combination

Figure 11: Difference maps

Table 3: Summary

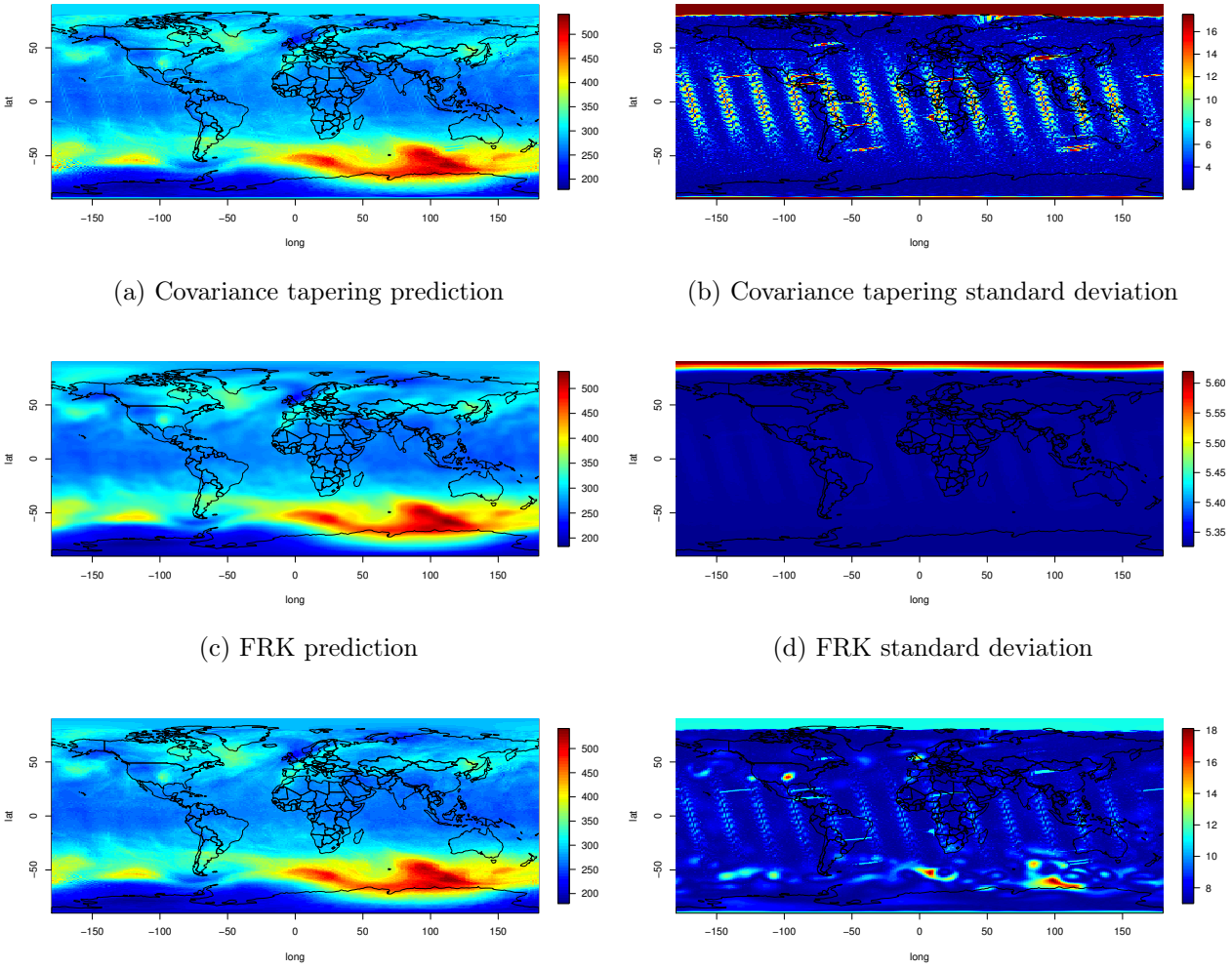
model	small scale structure	number of basis functions	MSPE
Covariance tapering	range 0.025	0	0.048
Low rank	nugget 0.03	560	0.036
Combination	range 0.025	287	0.008
Conditional expectation			0.002

combination (fig. 12e) and tapering (fig. 12a) predictions, where more details are visible. The latter presents however artifacts where the satellite coverage is less dense, especially visible along lines going south-southwest all around the globe. This effect is due to the use of a too small tapering range with respect to the range of the data. This causes the prediction to equals the mean when the data locations are too far apart. This effect is erased in the prediction map of the combination while preserving the accuracy in less dense areas.

The prediction standard deviations are much higher near the poles because of the lack of data. An increased estimated uncertainty in regions of missing data is especially visible for covariance tapering and the combination (figs. 12b and 12f) whereas this effect is almost imperceptible for FRK (fig. 12d). In the latter case, the predicted standard deviation is almost constant over the globe with a very low value compared to the two other methods. Finally, we can notice that the predicted standard deviation obtained with the combination show some areas with high value, mostly in the southern hemisphere. They correspond to the locations of the selected basis functions explaining local important departure from the mean of the TCO. They may seem a desirable effect as they correspond to additional parameters with respect to the covariance tapering approach.

Table 4 summarizes the parameter estimation results of all three methods and shows the MSPE values obtained on the validation dataset. Both covariance tapering and combination uses a nugget effect with variance one while their small scale structures differ from the weight assigned to the Wendland2 term, which is smaller for the combination. Less than 10% of the proposed basis functions are selected by the LASSO. Covariance tapering and FRK show similar results in terms of accuracy while the combination slightly outperforms both with the lowest MSPE.

Concerning the computation times, covariance tapering is the fastest with a running time of about 10, comprising about 9 minutes for parameter estimation and 1 minute for prediction. Better results could have been obtained by increasing the tapering range but the increased memory workload caused



(a) Covariance tapering prediction

(b) Covariance tapering standard deviation

(c) FRK prediction

(d) FRK standard deviation

(e) Combination prediction

(f) Combination standard deviation

Figure 12: Total Column Ozone data: Prediction and prediction standard deviation maps.

the computing time to increase considerably (more than one hour) due to swapping when solving the kriging system. The combination and FRK run in approximately the same time, around 20 minutes. Regarding the combination, the EM algorithm reaches the stopping criterion in about two minutes with the 331 selected basis functions within some hundredth of iterations, while 5 minutes are necessary to build the dictionary of basis functions, without parallelization, and 20 seconds for the lasso. Building the prediction map takes about 2 minutes. Concerning FRK, we run the EM algorithm for 10 iterations only to reduce the computing time and obviously do not reach the convergence criterion. This emphasizes the advantage to work with a low number of basis functions.

5 Conclusion

We have exposed here an original approach to interpolate large spatial datasets that somehow reconciliates existing approaches. It benefits from the advantages of both covariance tapering and low rank modeling, while alleviating the computational burden associated with the latter by using a limited number of basis functions. The model is thus able to represent each scale of variation of the phenomenon under study and achieves a good prediction accuracy. We have developed an inference

Table 4: Summary

model	small scale structure	number of basis functions	MSPE
Covariance tapering	nugget 4 + 125Wendland2(275)	0	24.63
Fixed rank	nugget 28	3608	24.09
Combination	nugget 4 + 80Wendland2(275)	331	21.64

approach adapted to this model. The two components are inferred almost separately through a step by step procedure. Both inference and prediction steps are fast to compute.

In each of the three application examples, combining covariance tapering and low rank decomposition outperformed covariance tapering and the low rank approach alone. It is true that allowing a larger tapering range would have improved the results of the former. However using a too large tapering range can reduce drastically the speed of the prediction by increasing the memory load. Therefore, our approach can be seen as a way to improve the results given by the covariance tapering by giving it a light low rank flavor: the examples indeed show that adding some basis functions to the model allow to erase the flaws obtained when using a too small taper in less sampled areas. The examples also show the versatility of the model: the same initial set of basis functions is used for the nested covariance model and the non stationary Matérn covariance model while obtaining good prediction results for each one.

Comparing computing times between different approaches for the kriging of large datasets can be troublesome. They depend on the implementation of each method and on the choice of the parameters value, namely the range in covariance tapering and the number of basis functions in fixed rank kriging. However, the application to the total column ozone dataset show that our method provides comparable computing times with exisiting appproaches, all implemented in `R` and run on a limited capacity computer.

References

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society Series B 70(4):825–848

Bolin D, Lindgren F (2011) Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. The Annals of Applied Statistics pp 523–550

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, Springer, pp 177–186

Braham H, Jemaa SB, Fort G, Moulines E, Sayrac B (2017) Fixed rank kriging for cellular coverage analysis. IEEE Transactions on Vehicular Technology 66(5):4212–4222

Chilès JP, Delfiner P (2012) Geostatistics, Modeling Spatial Uncertainty, 2nd edn. John Wiley & Sons, New-York

Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society Series B 70:209–226

Desassis N, Renard D (2013) Automatic variogram modeling by iterative least squares: Univariate and multivariate cases. Mathematical Geosciences 45(4):453–470

Eidsvik J, Shaby BA, Reich BJ, Wheeler M, Niemi J (2014) Estimation and prediction in spatial models with block composite likelihoods. Journal of Computational and Graphical Statistics 23(2):295–315

Emery X, Arroyo D (2017) On a continuous spectral algorithm for simulating non-stationary gaussian random fields. Stochastic Environmental Research and Risk Assessment

Fouedjio F, Desassis N, Romary T (2015) Estimation of space deformation model for non-stationary random functions. Spatial Statistics 13:45 – 61, DOI http://dx.doi.org/10.1016/j.spasta.2015.05.001, URL http://www.sciencedirect.com/science/article/pii/S221167531500041X

Fouedjio F, Desassis N, Rivoirard J (2016) A generalized convolution model and estimation for non-stationary random functions. Spatial Statistics 16:35–52

Friedman JH, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1–22, URL http://www.jstatsoft.org/v33/i01

Furrer R, Sain SR (2010) spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. Journal of Statistical Software 36(10):1–25, URL http://www.jstatsoft.org/v36/i10/

Furrer R, Genton MG, Nychka DW (2006) Covariance tapering for interpolation of large spatial datasets. Journal of Computational and Graphical Statistics 15(3):502–523

Gneiting T (2002) Compactly supported correlation functions. Journal of Multivariate Analysis 83(2):493–508

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer

Hsu NJ, Chang YM, Huang HC (2012) A group lasso approach for non-stationary spatial–temporal covariance estimation. Environmetrics 23(1):12–23

Katzfuss M (2017) A multi-resolution approximation for massive spatial datasets. Journal of the American Statistical Association 112(517):201–214

Katzfuss M, Cressie N (2011) Spatio-temporal smoothing and em estimation for massive remote-sensing data sets. Journal of Time Series Analysis 32(4):430–446

Kaufman CG, Schervish MJ, Nychka DW (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. Journal of the American Statistical Association 103(484):1545–1555, doi:10.1198/016214508000000959

Lindgren F, Rue H, Lindström J (2011) An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B 73(4):423–498

Loève M (1955) Probability Theory. Van Nostrand, New-York

Ma P, Kang EL (2017) Fused gaussian process for very large spatial data. arXiv preprint arXiv:170208797

Mardia KV, Marshall RJ (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. biometrika 71(1):135–46

Matheron G (1970) La théorie des variables régionalisées et ses applications. Les cahiers du centre de morphologie mathématiques de Fontainebleau, Ecole des Mines de Paris, fascicule 5

Nychka D, Wikle C, Royle JA (2002) Multiresolution models for nonstationary spatial covariance functions. Statistical Modelling 2:315–331

Nychka D, Bandyopadhyay S, Hammerling D, Lindgren F, Sain S (2015) A multiresolution gaussian process model for the analysis of large spatial datasets. Journal of Computational and Graphical Statistics 24(2):579–599

Paciorek CJ, Schervish MJ (2006) Spatial modelling using a new class of nonstationary covariance functions. Environmetrics 17(5):483–506

Parker RJ, Reich BJ, Eidsvik J (2016) A fused lasso approach to nonstationary spatial covariance estimation. Journal of agricultural, biological, and environmental statistics 21(3):569–587

Porcu E, Mateu J, Christakos G (2009) Quasi-arithmetic means of covariance functions with potential applications to space–time data. Journal of Multivariate Analysis 100(8):1830–1844

Renard D, Bez N, Desassis N, Beucher H, Ors F, Laporte F (2010) RGeostats: The Geostatistical package. Mines ParisTech, URL `http://cg.ensmp.fr/rgeos/`, version 9.0.0, Free download from: http://cg.ensmp.fr/rgeostats

Romary T (2009) Integrating production data under uncertainty by parallel interacting Markov chains on a reduced dimensional space . Computational Geosciences 13(1):103–122

Romary T (2013) Incomplete cholesky decomposition for the kriging of large datasets. Spatial Statistics 5:85–99

Romary T, Hu LY (2007) Assessing the Dimensionality of Random Fields with Karhunen-Loève Expansion. In: Petroleum Geostatistics 2007, EAGE

Sang H, Huang J (2012) A full scale approximation of covariance functions for large spatial data sets. Journal of the Royal Statistical Society Series B 74(1):111–132

Saporta G (2006) Probabilités, analyses des données et statistiques, 2nd edn. Technip

Stein ML (1993) A simple condition for asymptotic optimality of linear predictions of random fields. Statistics and Probability Letters 17:399–404

Stein ML (1999) Interpolation of spatial data: some theory for kriging. Springer

Stein ML (2008) A modeling approach for large spatial datasets. Journal of the Korean Statistical Society 37:3–10

Stein ML (2013) Statistical properties of covariance tapers. Journal of Computational and Graphical Statistics 22(4):866–885

Stein ML (2014) Limitations on low rank approximations for covariance matrices of spatial data. Spatial Statistics 8:1–19

Stein ML, Chi Z, Welty LJ (2004) Approximating likelihoods for large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66(2):275–296

Sun Y, Stein ML (2016) Statistically and computationally efficient estimating equations for large spatial datasets. Journal of Computational and Graphical Statistics 25(1):187–208

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B 58(1):267–288

Varin C, Reid NM, Firth D (2011) An overview of composite likelihood methods. Statistica Sinica 21(1):5–42