

Reduced rank covariances for the analysis of environmental data ¹

Orietta Nicolis ²

Dept. of Information Technology and Mathematical Methods, e-mail:
orietta.nicolis@unibg.it

Doug Nychka ³

Institute for Mathematics Applied to Geosciences, e-mail: nychka@ucar.edu

Abstract: In this work we propose an a Monte Carlo estimator for non stationary covariances of incomplete large spatial data and irregularly distributed observations. In particular, following the multiresolution approach introduced by Nychka *et al.* (2003) and Matsuo *et al.* (2008), we estimate a spatial reduced rank covariance starting from a Matérn model and using the W-transform basis in its eigenvector/eigenvalue decomposition. Since this method doesn't need to compute the wavelet coefficients, it is very efficient for the analysis of large dataset. Finally, the method is applied to aerosol optical thickness (AOT) satellite data, observed in Northern Italy, for the estimation of missing values.

Keywords: Monte-Carlo estimator, multi-resolution basis, nonstationary covariance, large data sets.

1. Introduction

The analysis of many geophysical and environmental problems requires the applications of interpolation techniques based on the estimation of covariance matrices. Due to non stationarity of data and to the large size of the dataset it is often difficult to use the classical covariance models. In this work we proposed a non parametric estimator for computing the covariance matrices in large dataset. By the multiresolution approach (Nychka *et al.* (2003)), a covariance function can be decomposed as Eq. (1). The basic idea of the work is to use a low resolution grid for estimating the matrix H with the advantage of reducing the rank of the corresponding covariance. The method can be used for the estimation of covariances of large irregular observational data and lattice data with many missing values. In this work we apply the multiresolution method based on the reduced rank covariance to the AOT satellite data (see Nicolis *et al.* (2008) for a description of the data) for estimating the missing values of aerosol optical thickness.

Next Section discuss the basics of multiresolution approach. Section 3 describes the algorithm for the estimation of the reduced rank covariance. Some results from the application of AOT satellite data are shown in Section 4.

¹The present paper is partially supported by PRIN project 2006.

²Address of correspondence: University of Bergamo, viale Marconi, 5, 24055 Dalmine, Italy.

³Address of correspondence: National Center for Atmospheric Research, Boulder, CO (USA).

2. The multiresolution approach

Let $\mathbf{y}$ be the values of the field on a fine grid. Denote the total number of grid points as m and let $\Sigma = COV(z)$ be the $m \times m$ covariance matrix among grid points. By the multiresolution approach (Nychka et al. 2003) the covariance matrix, Σ , can be decomposed as

$$\Sigma = WHH^T = WD^TW^T \quad (1)$$

where: W is the matrix of wavelet basis evaluated at m grid points; $D = W^{-1}\Sigma W^{-T}$ and $H = D^{1/2}$. Equivalently, we have the representation

$$\mathbf{y} = WHa$$

where a is a vector of independent $N(0, 1)$ random variables. The eigenvector/eigenvalue decomposition (1) suggests an alternative way of building the covariance by specifying the basis functions and a matrix H .

In this work we will use particular type of non-orthogonal scaling and wavelet functions known as the W-transform (Kwong and Tang (1994)). The W-transform is implemented in the R statistical language (<http://www.r-project.org>) in the fields package (Nychka (2005)).

The use of W-transform class in the spatial covariance model is motivated from the fact that they appear to approximate the shape of common covariance models and handle boundaries easily (Matsuo *et al.* (2008)).

3. The estimate of spatial covariance

In many geophysical applications the spatial fields are observed over time and one can exploit temporal replication to estimate sample covariances. In this work we focus on this case and also for gridded data with the goal of deriving a estimator that scale to large problems.

Suppose that the observations $\mathbf{y}$ are samples of a centered Gaussian random field and are composed of two components: the observations at irregularly distributed locations, $\mathbf{y}_o$, and the missing observations, $\mathbf{y}_m$,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_o \\ \mathbf{y}_m \end{pmatrix} \sim MN(0, \Sigma) \quad (2)$$

where WHH^TW^T . The observational model can be written as $\mathbf{y}_o = K\mathbf{y} + \varepsilon$ where ε is a multivariate normal $MN(0, \sigma^2 I)$ and K is a incidence matrix of ones and zeroes with a single one in each row indicating the position of each observation with respect to the grid. The conditional distribution of $\mathbf{y}_m$ given $\mathbf{y}_o$ is a multivariate normal with mean

$$\Sigma_{o,m}(\Sigma_{o,o})^{-1}\mathbf{y}_o \quad (3)$$

and variance

$$\Sigma_{m,m} - \Sigma_{m,o}(\Sigma_{o,o})^{-1}\Sigma_{o,m} \quad (4)$$

where $\Sigma_{o,m} = W_oHH^TW_m^T$ is the cross-covariance between observed and missing data, $\Sigma_{o,o} = W_oHH^TW_o^T + \sigma^2 I$ is covariance of observed data and $\Sigma_{m,m} = W_mHH^TW_m^T$ is

the covariance of missing data. The matrices W_o and W_m are wavelet basis evaluated at the observed and missing data, respectively.

For a chosen multiresolution basis and a sparse matrix H there are fast recursive algorithms for computing the covariance Σ . Matsuura et al. (2008) proposed a method that allows for sparse covariance matrices for the basis coefficients. However the evaluation of wavelet coefficients can be slow for large data sets.

In this work we propose an estimator for the H matrix, based on the reduced rank of the covariance Σ . The idea is to estimate Σ on a small sub-grid $\mathcal{G}$ of size $(g \times g)$ starting from a Matérn model. Monte Carlo simulation provide and an efficient estimator for the conditional mean.

Since the W is fixed for a chosen basis, the estimation procedure for the conditional covariance is given by the estimation of the matrix H after a sequence of approximations. The algorithm can be described in the following steps.

1. Fit an initial stationary covariance model (for example, the Matérn class) to the observed data and compute the covariance matrix, $\Sigma_{g,g}$, on a small resolution grid $\mathcal{G}$. From the relation (1), derive the $(g \times g)$ matrix $\tilde{H}_g = (W_g^{-1} \Sigma_{g,g} W_g^T)^{1/2}$ where W_g are basis functions evaluated at the grid points.
2. Use $\tilde{H}_g$ for the kriging prediction at $\mathcal{G}$, $\hat{y}_g = \Sigma_{o,g}(\Sigma_{o,o})^{-1} \mathbf{y}_o$ where $\Sigma_{o,g} = W_o \tilde{H}_g \tilde{H}_g^T W_g^T$ is cross-covariance between observed data and the grid points.
3. Generate T draws from the conditional distribution of the complete-data given the observations and $\tilde{H}_g$. In particular, for each sample,
 - generate a $m \times 1$ vector of data $\mathbf{y}^s = W \tilde{H}_g a$ with $a \sim N(0, 1)$ and select a $n \times 1$ pseudo-data vector $\mathbf{y}_o^s$. (note that W is the matrix of basis functions evaluated on the full data m);
 - compute the kriging prediction on the grid $\mathcal{G}$, $\mathbf{y}_g^s = \Sigma_{o,g}(\Sigma_{o,o})^{-1} \mathbf{y}_o^s$, where $\Sigma_{o,g} = W_o \tilde{H}_g \tilde{H}_g^T W_g^T$;
 - compute the kriging error $\mathbf{u}^* = \mathbf{y}_g^s - \hat{y}_g$. Simple matrix algebra will show that $\hat{y}_u = \hat{y}_g + \mathbf{u}^*$ will be a sample from the correct conditional distribution.
4. Compute the conditional covariance on T replications, $\Sigma_u = COV(\hat{y}_u)$ and use the new $\tilde{H}_g$ in the step 2.

Performing this several times will give an ensemble of fields and, of course, finding the sample covariance across the ensemble provides a Monte Carlo based estimate of the conditional covariance.

4. Application

We consider the column aerosol optical thickness (AOT) derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) on the Terra/Aqua satellites in Northern Italy. The meteorological conditions determine the AOT retrieval since the availability of the data is restricted to cloud-free conditions. This cause a large number of missing data. In this work we consider daily AOT for the period April 1st - June 30th, 2006 on a spatial grid of 10 km^2 resolution. Figure 1 shows the nonstationary structure of the estimated covariance $W \tilde{H}_g \tilde{H}_g^T W^T$ obtained after five iterations of MC simulations. The kriging prediction for the day 207 is shown in Figure 2. We can note that higher values of AOT are concentrated in the Po Valley.

Figure 1: This figure shows non-stationary covariance $W \tilde{H}_g \tilde{H}_g^T W^T$ obtained after five iterations of MC simulations. Each panel shows the covariance between the point indicated by black circle and the rest of grid points.

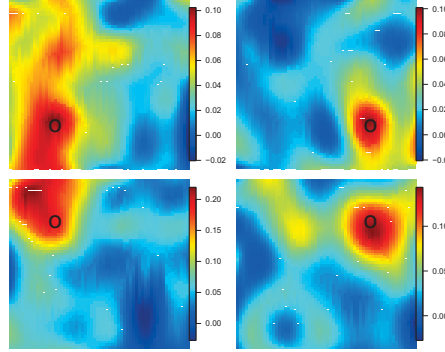
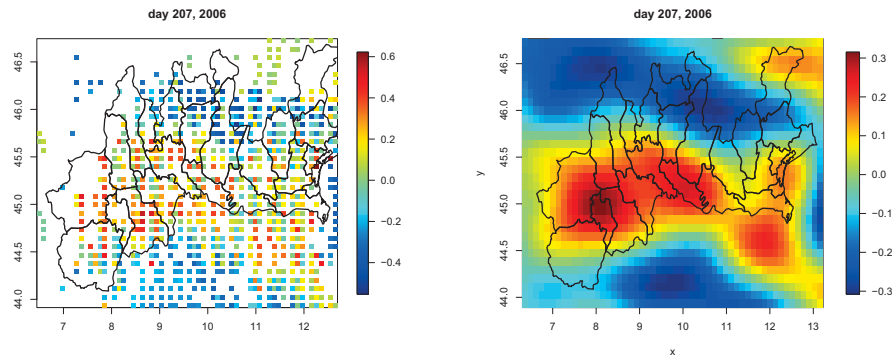


Figure 2: This figure shows the observations and the kriging predictions using wavelet-based non stationary covariance.



References

- Kwong M.K. and Tang P.T.P. (1994) W-matrices, nonorthogonal multiresolution analysis, and finite signals of arbitrary length, *Argonne National Laboratory Technical Report*.
- Matsuo T., Nychka D. and Paul D. (2008) Nonstationary covariance modeling for incomplete data: Smoothed monte carlo em approach, *In review*.
- Nicolis O., Fassò A. and Mannarini G. (2008) Aot calibration by spatio-temporal model in northern italy, *Spatio-Temporal Modelling (METMA 5)*, 24-26 September, Alghero, Sardinia.
- Nychka D. (2005) Fields: Tools for spatial data, *National Center for Atmospheric Research, Boulder, CO*.
- Nychka D., Wikle C. and Royle J.A. (2003) Multiresolution models for nonstationary spatial covariance functions, *Statistical Modeling*, 2, 315–332.