

Distributed computing for spatio-temporal hierarchical models¹

Michela Cameletti²

Dept. of Mathematics, Statistics, Computing and Applications, University of Bergamo
e-mail: michela.cameletti@unibg.it

Abstract: This work focuses on spatio-temporal hierarchical models and on the parallelization of the related computing intensive procedures using R statistical computing language.

Keywords: parallel computing, computing intensive procedures, R language, parametric bootstrap, spatio-temporal covariance function, environmental statistics.

1. Introduction

Over the last decade, hierarchical spatio-temporal models have become popular especially for environmental modeling. This is related to their flexible conditional point of view that makes it possible to express the joint probability distribution of a spatio-temporal process as the product of some simpler distributions (Wikle (2003)). However likelihood based inferences for spatio-temporal models (i.e. parameter estimation and spatial prediction) involve the use of computing intensive statistical methods, such as bootstrap simulations or Markov Chain Monte Carlo (MCMC) methods. Moreover, large spatio-temporal datasets are available for many environmental applications giving rise to the so-called "big n" problem in geostatistics and to the infeasibility of matrix operations whose complexity increases in cubic order with the number of data. An approach to tackle these computational burdens is given by parallel computing that can be easily implemented using R statistical computing language (R Development Core Team (2006)).

In this work we focus on the spatio-temporal hierarchical model presented in Fassò and Cameletti (2009a,b) and describe how it is possible to parallelize the related computing intensive procedures.

2. Spatio-temporal hierarchical models

Let $Z(s, t)$ be the observed scalar spatio-temporal process at time t and site s . Let $Z_t = (Z(s_1, t), \dots, Z(s_d, t))'$ denote the data vector at time t and at d geographical locations $s_1, \dots, s_d$. Moreover let $Y_t = (Y_1(t), \dots, Y_p(t))'$ be a p -dimensional vector for the unobserved temporal process at time t . The spatio-temporal hierarchical model defined by the following equations for $t = 1, \dots, T$

$$Z_t = X_t\beta + KY_t + e_t \quad (1)$$

$$Y_t = GY_{t-1} + \eta_t \quad (2)$$

¹This work is partially supported by PRIN project n. 2006131039.

²Address of correspondence: Via dei Caniana 2, 24127 Bergamo (BG).

can be considered as a classical *state-space model* where (1) is the measurement equation and (2) is the state equation. The term X_t is the $(d \times k)$ covariate matrix, K_t is a $(d \times p)$ matrix of ones, $G = \text{diag}(g_1, \dots, g_p)$ is the transition matrix and the innovation η_t is a p -dimensional Gaussian white-noise process with covariance matrix $\Sigma_\eta = \text{diag}(\sigma_{\eta_1}^2, \dots, \sigma_{\eta_p}^2)$. In Eq.(1) the error e_t has a zero-mean Gaussian distribution with covariance matrix $\Sigma_e = \sigma_\omega^2 \Gamma(\|s_i - s_j\|)_{i,j=1,\dots,d}$, with Γ given by the following scaled spatial covariance function

$$\Gamma_{\gamma,\theta}(h) = \begin{cases} 1 + \gamma & h = 0 \\ C_\theta(h) & h > 0 \end{cases} \quad (3)$$

where $\gamma = \sigma_\varepsilon^2/\sigma_\omega^2$ and $h = \|s_i - s_j\|$ is the Euclidean distance between site i and j . The measurement error variance σ_ε^2 can be interpreted in geostatistical terms as the “nugget effect” while σ_ω^2 is the so-called partial sill (Cressie (1993)). The constant in time function $C_\theta(h)$ of Eq.(3) is defined as

$$\text{Cov}(\omega(s_i, t), \omega(s_j, t')) = \sigma_\omega^2 C_\theta(h) \quad \forall i \neq j, t \neq t' \quad (4)$$

and is continuous at $h = 0$ with $\lim_{h \rightarrow 0} C_\theta(h) = 1$.

Hence, the parameter vector to be estimated is given by $\Psi = \{\beta, \sigma_\omega^2, G, \Sigma_\eta, \mu_0, \Sigma_0, \gamma, \theta\}$ where μ_0 and Σ_0 are the parameters of the initial Gaussian state Y_0 . The maximum likelihood (ML) estimation of the unknown parameter vector Ψ is performed using the iterative procedure given by the EM algorithm (McLachlan and Krishnan (1997), Xu and Wikle (2007)). All the details regarding the Expectation and the Maximization steps of the algorithm are reported in Fassò and Cameletti (2009a). In particular, it is shown that the maximization step is based on closed form formulas for all the parameters except for the spatial covariance ones, which are obtained by the Newton-Raphson algorithm. This means that the stability of the algorithm is enhanced with respect to the classical Newton-Raphson likelihood optimization methods. On the other hand, a disadvantage of the EM algorithm is that it does not provide ready-to-use standard errors of the parameter estimates. Consequently a parametric bootstrap is required for approximating the above standard errors. Since this is a computing intensive procedure that requires to compute the EM algorithm for each bootstrap iteration, it is performed in parallel as described in the following section.

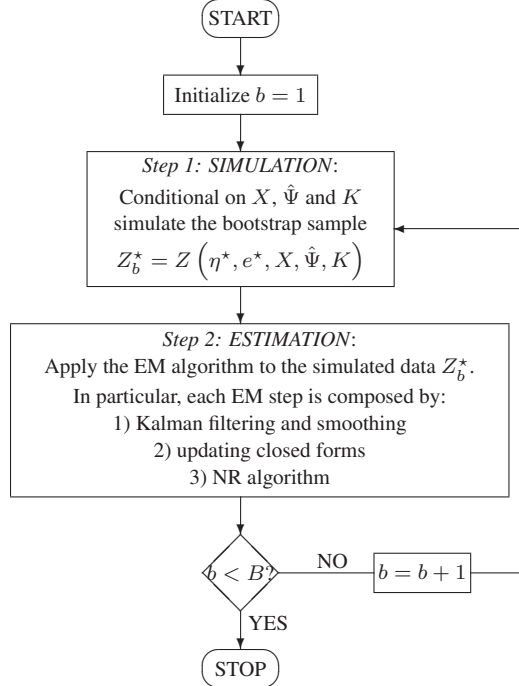
3. Parallel computing for spatio-temporal hierarchical models with R

Parallel computing consists in dividing a job into small tasks for parallel execution. In particular, the granularity of the parallelization is defined as the ratio between computation and communication, that is the amount of independent parallel processing carried out before requiring some sort of communication or synchronization (Kontoghiorghe (2006)). Statistical procedures that require independent repeated computations on different data sets, such as *bootstrap* and *cross-validation*, are examples of coarse grain parallelization, also referred as embarrassingly parallel problems.

The spatio-temporal bootstrap used for the hierarchical model of Section 2. is a sequence of B simulations based on the assumption that the estimated parametric model is the correct one. In particular, we simulate directly from the involved Gaussian distributions

using equations (1) and (2), with Ψ replaced by its ML estimate $\hat{\Psi}$ and the covariates $X = (X_1, \dots, X_T)$ kept fixed for all the B simulations. The details of the b -th single bootstrap simulation, for fixed $b = 1, \dots, B$, are schematically described in Fig.1. The computing intensive bootstrap procedure is performed using a computer cluster with 4 cpu's of the quad core Intel Xeon processor running at 2.66 GHz and a Linux environment. In particular, the code for estimation and bootstrap has been organized in a R library called `stem` from the acronym of Spatio-Temporal EM (Cameletti (2009)). The parallelization of the procedure is then implemented using the R packages called `RMPI` and `SNOW`. The first package is an interface to MPI system library (Message-Passing Interface) which is a standardized and portable message-passing system for defining the cluster and the coordination of the node work. The second package, `SNOW`, provides a high-level interface for delivering the job through the nodes of the cluster and contains the parallel version of the classical R functions, such as `apply` and `lapply`. For a complete and recent overview of parallel computing techniques with R see Schmidberger *et al.* (2009).

Figure 1: Schematic flowchart of the spatio-temporal bootstrap.



In Fassò and Cameletti (2009b) a similar spatio-temporal bootstrap is parallelized but with reference to a Kriging spatial interpolator with the aim to assess the map uncertainty through bootstrap standard errors. The procedure is like the one reproduced in Fig.1 but with an additional step where the spatial predictions are computed for a set of new locations conditionally on the parameter estimated at Step 2. Moreover, the authors perform in parallel a leave-one-out cross-validation analysis in order to evaluate the spatial capability of the spatio-temporal hierarchical model with different covariates. This is another example of coarse grain parallelization: in fact the same inferential procedures (parameter estimation with the EM algorithm and Kriging spatial prediction) are independently repeated on d different data sets obtained removing one observation at a time.

Extensions in parallel computing

If a Bayesian approach is adopted to infer spatio-temporal hierarchical models (i.e. parameter estimation and spatial prediction) MCMC methods are required. They are notoriously computing intensive procedures especially when the sample size increases and matrix operations can become infeasible. The parallelization of MCMC procedures is a fine grain problem because they are inherently sequential and the simulation of the next values of the chain can not begin until the current value has been simulated. Nevertheless there are opportunities for parallelization. First of all, it is possible to run trivially several parallel chains, especially for diagnostic purposes. Moreover, for high-dimensional problems, such as the case of spatio-temporal processes, it happens that every iteration is computationally expensive and so it is convenient to parallelize the computations required for each iteration (Yan *et al.* (2007)). In particular, it is possible to use the conditional independence property to update independently groups of parameters or to perform matrix calculations in parallel. This last case is particularly challenging because it would be a solution also for the "big n" problem computational burden that occurs in applications with large datasets, especially when the constance in time or the separability assumptions can not be used for the spatio-temporal covariance function of Eq.(4). With this regard the `RScalAPACK` is a R package based on the high-performance ScaLAPACK library for linear algebra routines and can be used for carrying out parallel matrix computations through a single function call from the R environment.

References

- Cameletti M. (2009) Stem: Spatio-temporal models in R, <http://cran.r-project.org/web/packages/Stem/>.
- Cressie N. (1993) *Statistics for Spatial Data*, Wiley, New York.
- Fassò A. and Cameletti M. (2009a) The EM algorithm in a distributed computing environment for modelling environmental space-time data, *Environmental Modeling and Software*, 24, 1027–1035.
- Fassò A. and Cameletti M. (2009b) A unified statistical approach for simulation, modelling, analysis and mapping of environmental data, accepted for publication in *Simulation: Transactions of the Society for Modeling and Simulation International*.
- Kontoghiorghe E.J. (Ed.) (2006) *Handbook of parallel computing and statistics*, Chapman and Hall.
- McLachlan G.J. and Krishnan T. (1997) *The EM Algorithm and Extensions*, Wiley, New York.
- R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Schmidberger M., Morgan M., Eddelbuettel D., Yu H., Tierney L. and Mansmann U. (2009) State-of-the-art in parallel computing with R, Technical Report 47, Dept. of Statistics, Univ. of Munich.
- Wikle C.K. (2003) Hierarchical models in environmental science, *International Statistical Review*, 71, 181–199.
- Xu K. and Wikle C.K. (2007) Estimation of parameterized spatio-temporal dynamic models, *Journal of Statistical Inference and Planning*, 137, 567–588.
- Yan J., Cowles M., Wang S. and Armstrong M. (2007) Parallelizing MCMC for Bayesian spatiotemporal geostatistical models, *Statistics and Computing*, 17, 4, 323–335.