Incomplete Cholesky decomposition for the kriging of large datasets

CrossMark

Thomas Romary *

Mines ParisTech, Equipe Géostatistique, Centre de Géosciences, 35 rue Saint Honoré, 77305 Fontainebleau, France

ARTICLE INFO

ABSTRACT

Kriging of very large spatial datasets is a challenging problem. The size n of the dataset causes problems in computing the kriging estimate: solving the kriging equations directly involves inverting an $n \times n$ covariance matrix. This operation requires $O(n^3)$ computations and a storage of $O(n^2)$. Under these circumstances, straightforward kriging of massive datasets is not possible. Several approaches have been proposed in the literature among which two main families exist: sparse approximation of the covariance function and low rank approaches. We propose here an approach that is built upon a low rank approximation of the covariance matrix obtained by incomplete Cholesky decomposition. This algorithm requires $O(nk)$ storage and takes $O(nk^2)$ arithmetic operations, where k is the rank of the approximation, whose accuracy is controlled by a parameter. We detail the main properties of this method and explore its links with existing methods. Its benefits are illustrated on simple examples and compared to those of existing approaches. Finally, we show that this low rank representation is also suited for inverse conditioning of Gaussian random fields.

© 2013 Elsevier B.V. All rights reserved.

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 development of remote sensing platforms on satellites or planes, spatial database paradigms have moved from small to massive. Therefore,

---

* Tel.: +33 164694704.
E-mail addresses: thomas.romary@mines-paristech.fr, thomas.romary@gmail.com.

new methods for the geostatistical analysis of large datasets have been developed. Indeed, richer datasets allow for more complex modeling but may also prevent the straightforward use of classical techniques. The challenge of handling such datasets is to extract the maximum of information that they contain while ensuring the numerical tractability of the interpolation algorithms.

Conventional geostatistical interpolation methods such as kriging are made numerically intractable by the size of the dataset. Indeed, solving the kriging equations directly involves the inversion of an $n \times n$ variance–covariance matrix C, which requires $O(n^3)$ computations. Under these circumstances, straightforward kriging of massive datasets is not possible. In practice, kriging is achieved approximately by local approaches that are based on considering only a relatively small number of points that lie close to the query point. Determining the proper neighborhood size is usually solved by ad hoc methods such as selecting a fixed number of nearest neighbors or all the points lying within a fixed radius. Such fixed neighborhood sizes may not work well for all query points, depending on the local density of the point distribution. Local methods also suffer from the problem that the resulting interpolant is not continuous, even though approaches have been proposed to tackle the latter; see Gribov and Krivoruchko (2004) and in a more general context Rivoirard and Romary (2011). Furthermore, the larger the dataset, the lower the prediction variance: the most accurate prediction cannot be obtained with a subsample. Finally, if a non-stationary covariance model is used (see e.g. Gelfand et al., 2011) a small neighborhood will not be able to capture the complex correlation structure. For these different reasons, it is better to consider and solve the global system for each interpolant. Solving such large dense systems for each query point is impractical however. To tackle this problem, several approaches have been proposed in the literature for both prediction and estimation, and are described below; see also Sun et al. (2012) for a review.

In this paper, we consider a zero-mean spatial random field $\{Z(x), x \in \mathcal{X} \subset \mathbb{R}^d\}, d \in \mathbb{N}^*$. We denote by $C(x, y)$ the covariance of Z, $(x, y) \in \mathcal{X}^2$, the covariance function. We assume throughout the paper that the parameters of C are known. The kriging problem is to predict $Z(x_0)$, $x_0 \in \mathcal{X}$, given the observation of Z at n locations $x_1, \ldots, x_n$. The BLUP at an unobserved location x_0 is

$$Z^*(x_0) = C_0{}' C^{-1} Z, \tag{1}$$

where $Z = (Z(x_1), \ldots, Z(x_n))'$, $C_{ij} = C(x_i, x_j)$ and $C_{0i} = C(x_i, x_0)$. The associated mean squared prediction error is

$$\sigma(x_0) = C(x_0, x_0) - C_0{}' C^{-1} C_0. \tag{2}$$

Note that $Z^*(x_0)$ and $\sigma(x_0)$ are the conditional expectation and variance of $Z(x_0)$, given the observations, under the Gaussian hypothesis.

This situation corresponds to the simple kriging with a known zero mean. Nonetheless, the different approaches described in this paper can be used with any kriging technique (ordinary, universal, etc.) by considering a blockwise matrix inversion.

The paper is organized as follows: In Section 2, we review the main existing approaches for the kriging of large datasets. Then in Section 3, we detail the main properties of the incomplete Cholesky decomposition when used for kriging and explore its links with existing methods. We then test the sensitivity of the different approaches to various settings on toy examples in Section 4. Finally, in Section 5, we discuss the use of incomplete Cholesky decomposition for dimension reduction in inverse problems as well as its link with the predictive process approach.

2. Existing approaches

Most existing approaches have considered approximating the covariance function C in an attempt to make the computation of C^{-1} computationally tractable. These have been classified into two main categories: sparse approximation and low rank approximation. The combination of these is also considered.