Journal of Multivariate Analysis 125 (2014) 1–35

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes ☆

François Bachoc *

CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France
Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot, 75205 Paris cedex 13, France

ARTICLE INFO

Article history:
Received 17 January 2013
Available online 12 December 2013

AMS subject classifications:
primary 62M30
secondary 62F12

Keywords:
Uncertainty quantification
Metamodel
Kriging
Covariance parameter estimation
Maximum likelihood
Leave-one-out
Increasing-domain asymptotics

ABSTRACT

Covariance parameter estimation of Gaussian processes is analyzed in an asymptotic framework. The spatial sampling is a randomly perturbed regular grid and its deviation from the perfect regular grid is controlled by a single scalar regularity parameter. Consistency and asymptotic normality are proved for the Maximum Likelihood and Cross Validation estimators of the covariance parameters. The asymptotic covariance matrices of the covariance parameter estimators are deterministic functions of the regularity parameter. By means of an exhaustive study of the asymptotic covariance matrices, it is shown that the estimation is improved when the regular grid is strongly perturbed. Hence, an asymptotic confirmation is given to the commonly admitted fact that using groups of observation points with small spacing is beneficial to covariance function estimation. Finally, the prediction error, using a consistent estimator of the covariance parameters, is analyzed in detail.

1. Introduction

In many areas of science that involve measurements or data acquisition, one often has to answer the question of how the set of experiments should be designed [19]. It is known that in many situations, an irregular, or even random, spatial sampling is preferable to a regular one. Examples of these situations are found in many fields. For numerical integration, Gaussian quadrature rules generally yield irregular grids [22, Ch. 4]. The best known low-discrepancy sequences for quasi-Monte Carlo methods (van der Corput, Halton, Sobol, Faure, Hammersley, etc.) are not regular either [20]. In the compressed sensing domain, it has been shown that one can recover a signal very efficiently, and at a small cost, by using random measurements [6].

In this paper, we are focused on the role of spatial sampling for meta-modeling. Meta-modeling is particularly relevant for the analysis of complex computer models [28]. We will address the case of Kriging models, which consist in interpolating the values of a Gaussian random field given observations at a finite set of observation points. Kriging has become a popular method for a large range of applications, such as numerical code approximation [27,28] and calibration [21] or global optimization [13].

One of the main issues regarding Kriging is the choice of the covariance function for the Gaussian process. Indeed, a Kriging model yields an unbiased predictor with minimal variance and a correct predictive variance only if the correct covariance function is used. The most common practice is to statistically estimate the covariance function, from a set of observations of the Gaussian process, and to plug [29, Ch. 6.8] the estimate in the Kriging equations. Usually, it is assumed

☆ A supplementary material is attached in the electronic version of the article (see Appendix F).

* Corresponding author at: CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France.
E-mail addresses: francois.bachoc@cea.fr, francois.bachoc@gmail.com.

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.11.015

that the covariance function belongs to a given parametric family (see [1] for a review of classical families). In this case, the estimation boils down to estimating the corresponding covariance parameters.

The spatial sampling, and particularly its degree of regularity, plays an important role for the covariance function estimation. In Chapter 6.9 of [29], it is shown that adding three observation points with small spacing to a one-dimensional regular grid of twenty points dramatically improves the estimation in two ways. First, it enables to detect without ambiguities that a Gaussian covariance model is poorly adapted, when the true covariance function is Matérn $\frac{3}{2}$. Second, when the Matérn model is used for estimation, it subsequently improves the estimation of the smoothness parameter. It is shown in [40] that the optimal samplings, for maximizing the log of the determinant of the Fisher information matrix, averaged over a Bayesian prior on the true covariance parameters, contain closely spaced points. Similarly, in the geostatistical community, it is acknowledged that adding sampling crosses, that are small crosses of observation points making the different input quantities vary slightly, enables a better identification of the small scale behavior of the random field, and therefore a better overall estimation of its covariance function [12]. The common conclusion of the three examples we have given is that irregular samplings, in the sense that they contain at least pairs of observation points with small spacing, compared to the average density of observation points in the domain, work better for covariance function estimation than regular samplings, that is samplings with evenly spaced points. This conclusion has become a commonly admitted fact in the Kriging literature.

In this paper, we aim at confirming this fact in an asymptotic framework. Since exact finite-sample results are generally not reachable and not meaningful as they are specific to the situation, asymptotic theory is widely used to give approximations of the estimated covariance parameter distribution.

The two most studied asymptotic frameworks are the increasing-domain and fixed-domain asymptotics [29, p. 62]. In increasing-domain asymptotics, a minimal spacing exists between two different observation points, so that the infinite sequence of observation points is unbounded. In fixed-domain asymptotics, the sequence is dense in a bounded domain.

In fixed-domain asymptotics, significant results are available concerning the estimation of the covariance function, and its influence on Kriging predictions. In this asymptotic framework, two types of covariance parameters can be distinguished: microergodic and non-microergodic covariance parameters. Following the definition in [29], a covariance parameter is microergodic if two covariance functions are orthogonal whenever they differ for it (as in [29], we say that two covariance functions are orthogonal if the two underlying Gaussian measures are orthogonal). Non-microergodic covariance parameters cannot be consistently estimated, but have no asymptotic influence on Kriging predictions [30–32,38]. On the contrary, there is a fair amount of literature on consistently estimating microergodic covariance parameters, notably using the Maximum Likelihood (ML) method. Consistency has been proved for several models [36,37,17,38,16,4]. Microergodic covariance parameters have an asymptotic influence on predictions, as shown in [35, Ch. 5].

Nevertheless, the fixed-domain asymptotic framework is not well adapted to study the influence of the irregularity of the spatial sampling on covariance parameter estimation. Indeed, we would like to compare sampling techniques by inspection of the asymptotic distributions of the covariance parameter estimators. In fixed-domain asymptotics, when an asymptotic distribution is proved for ML [36,37,10], it turns out that it is independent of the dense sequence of observation points. This makes it impossible to compare the effect of spatial sampling on covariance parameter estimation using fixed-domain asymptotics techniques.

The first characteristic of increasing-domain asymptotics is that, as shown in Section 5.1, all the covariance parameters have strong asymptotic influences on predictions. The second characteristic is that all the covariance parameters (satisfying a very general identifiability assumption) can be consistently estimated, and that asymptotic normality generally holds [34,18,7]. Roughly speaking, increasing-domain asymptotics is characterized by a vanishing dependence between observations from distant observation points. As a result, a large sample size gives more and more information about the covariance structure. Finally, we show that the asymptotic variances of the covariance parameter estimators strongly depend on the spatial sampling. This is why we address the increasing-domain asymptotic framework to study the influence of the spatial sampling on the covariance parameter estimation.

We propose a sequence of random spatial samplings of size $n \in \mathbb{N}$. The regularity of the spatial sampling sequence is characterized by a regularity parameter $\epsilon \in [0, \frac{1}{2})$. $\epsilon = 0$ corresponds to a regular grid, and the irregularity is increasing with ϵ. We study the ML estimator, and also a Cross Validation (CV) estimator [33,39], for which, to the best of our knowledge, no asymptotic results are yet available in the literature. For both estimators, we prove an asymptotic normality result for the estimation, with a $\sqrt{n}$ convergence, and an asymptotic covariance matrix which is a deterministic function of ϵ. The asymptotic normality yields, classically, approximate confidence intervals for finite-sample estimation. Then, carrying out an exhaustive analysis of the asymptotic covariance matrix, for the one-dimensional Matérn model, we show that large values of the regularity parameter ϵ always yield an improvement of the ML estimation. We also show that ML has a smaller asymptotic variance than CV, which is expected since we address the well-specified case here, in which the true covariance function does belong to the parametric set used for estimation. Thus, our general conclusion is a confirmation of the aforementioned results in the literature: using a large regularity parameter ϵ yields groups of observation points with small spacing, which improve the ML estimation, which is the preferable method to use.

The rest of the article is organized as follows. In Section 2, we introduce the random sequence of observation points, that is parameterized by the regularity parameter ϵ. We also present the ML and CV estimators. In Section 3, we give the asymptotic normality results. In Section 4, we carry out an exhaustive study of the asymptotic covariance matrices for the Matérn model in dimension one. In Section 5, we analyze the Kriging prediction for the asymptotic framework we consider.