Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case

François Bachoc

Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case

François Bachoc
Department of Statistics, University of Vienna

December 4, 2014

Abstract

In parametric estimation of covariance function of Gaussian processes, it is often the case that the true covariance function does not belong to the parametric set used for estimation. This situation is called the misspecified case. In this case, it has been observed that, for irregular spatial sampling of observation points, Cross Validation can yield smaller prediction errors than Maximum Likelihood. Motivated by this comparison, we provide a general asymptotic analysis of the misspecified case, for independent observation points with uniform distribution. We prove that the Maximum Likelihood estimator asymptotically minimizes a Kullback-Leibler divergence, within the misspecified parametric set, while Cross Validation asymptotically minimizes the integrated square prediction error. In a Monte Carlo simulation, we show that the covariance parameters estimated by Maximum Likelihood and Cross Validation, and the corresponding Kullback-Leibler divergences and integrated square prediction errors, can be strongly contrasting. On a more technical level, we provide new increasing-domain asymptotic results for the situation where the eigenvalues of the covariance matrices involved are not upper bounded.

1 Introduction

Kriging models (Stein, 1999; Rasmussen and Williams, 2006) consist in inferring the values of a Gaussian random field given observations at a finite set of observation points. They have become a popular method for a large range of applications, such as numerical code approximation (Sacks et al., 1989; Santner et al., 2003) and calibration (Paulo et al., 2012) or global optimization (Jones et al., 1998).

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 (Stein, 1999, Ch.6.8) the estimate in the Kriging equations. Usually, it is assumed that the covariance function belongs to a given parametric family (see Abrahamsen (1997) for a review of classical families). In this case, the estimation boils down to estimating the corresponding covariance parameters.

For covariance parameter estimation, Maximum Likelihood (ML) is the most studied and used method, while Cross Validation (CV) (Sundararajan and Keerthi, 2001; Zhang and Wang, 2010) is an alternative technique. The two estimators have been compared by several references. Consider first the case where the true covariance function of the Gaussian process belongs to the parametric family of covariance functions used for estimation, that we call the well-specified case. Then Stein (1990b) shows that for the estimation of a signal-to-noise ratio parameter of a Brownian motion, CV has twice the asymptotic variance of ML. In the situations treated by Bachoc (2014), the asymptotic variance is also larger for CV than for ML. Several numerical results, showing an advantage for ML over CV as well, are available, coming either from Monte Carlo studies as in (Santner et al., 2003, Ch.3) or deterministic studies as in Martin and Simpson (2004). The settings of both the above studies can arguably be classified in the well-specified case, since the interpolated functions are smooth, and the covariance structures are adapted, being Gaussian in Martin and Simpson (2004) and having a free smoothness parameter in Santner et al. (2003). Finally, in situations similar to the well-specified case, ML-type methods have been shown to be preferable over CV-type methods in Stein (1993) for estimation and prediction.

Consider now the case where the true covariance function of the Gaussian process does not belong to the parametric family of covariance functions used for estimation, that we call the misspecified case. This can occur in many situations, given for example that it is frequent to enforce the smoothness parameter in the Matérn model to an arbitrary value (e.g. 3/2 in Chevalier et al. (2014)), which de facto makes the covariance model misspecified if the Gaussian process has a different order of smoothness. In the misspecified case, Bachoc (2013) shows that, provided the spatial sampling of observation points is not too regular, CV can yield smaller prediction errors than ML. In a context of spline approximation methods, Stein (1993) and Kou (2003) also suggest that CV-type methods can provide better predictions than ML-type methods under misspecification.

In this paper, we aim at providing a general asymptotic analysis of the misspecified case, which would confirm the findings of the aforementioned references on the comparison between ML and CV. The two most studied asymptotic frameworks are the increasing-domain and fixed-domain asymptotics (Stein, 1999, p.62). In increasing-domain asymptotics, the average density of observation points is bounded, so that the infinite sequence of observation points is unbounded. In fixed-domain asymptotics, this 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 Stein (1999), a covariance parameter is microergodic if two covariance functions are orthogonal whenever they differ for it (as in Stein (1999), 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 (Stein, 1988, 1990a,c; Zhang, 2004). On the contrary, there is a fair amount of literature on consistent estimation of microergodic covariance parameters (Ying, 1991, 1993; Zhang, 2004; Loh, 2005; Anderes, 2010). Microergodic covariance parameters have an asymptotic influence on predictions, as shown in (Vazquez, 2005, Ch.5).

Nevertheless, similarly to Bachoc (2014), we do not address fixed-domain asymptotics in this paper. In Bachoc (2014), the reason is that fixed-domain asymptotics yields no impact of the spatial sampling of observation points on covariance function estimation. We follow this line here, since Bachoc (2013) makes the important observation that the comparison between ML and CV strongly depends on the spatial sampling. In this paper, we indeed derive asymptotic results for CV for which the type of spatial sampling used is crucial. Furthermore, another downside of fixed-domain asymptotics is that the results currently under reach, despite their significant insights, are restricted in terms of covariance model. For example, Ying (1993) addresses ML for the tensorized exponential model only and Loh (2005) addresses ML for the Matérn 3/2 covariance model only.

Hence, in this paper, we work under increasing-domain asymptotics, which solves the two aforementioned issues. Indeed, first Bachoc (2014) shows that the asymptotic distributions of the ML and CV estimators (in the well-specified case) strongly depends on the spatial sampling. Second, increasing-domain asymptotic results (Mardia and Marshall, 1984; Cressie and Lahiri, 1993, 1996; Bachoc, 2014) are available for fairly general covariance models. In fact, generally speaking, under increasing-domain asymptotics, all (identifiable) covariance parameters have a strong asymptotic influence on predictions (Bachoc, 2014) and can be consistently estimated with asymptotic normality (Mardia and Marshall, 1984; Bachoc, 2014). This is because increasing-domain asymptotics is characterized by a vanishing dependence between observations from distant observation points, so that a large sample size gives more and more information about the covariance structure.

The increasing-domain asymptotic spatial sampling we consider consists of n independent observation points with uniform distribution on $[0, n^{1/d}]^d$, for $d \in \mathbb{N}^*$. We prove that ML asymptotically minimizes, within the misspecified model, the Kullback-Leibler divergence from the true covariance function, defined at the observation vector. As we discuss after Theorem 3.3, this can actually be shown under more general spatial samplings than that considered here, but still necessitates an original proof. In the misspecified case, the important point is that no information is provided on the quality of the ML estimator for subsequent predictions of the Gaussian process at new points.

In Theorem 3.4, we prove that CV asymptotically minimizes the integrated square prediction error, within the misspecified set of covariance functions used for estimation. As discussed after Theorem 3.4, considering independent observation points is crucial. Thus, an asymptotic confirmation is given to the empirical finding of Bachoc (2013), that when the spatial sampling is not too regular, CV generally performs better than ML for prediction in the misspecified case, while it can instead perform poorly under regular sampling. On a more

technical level, the proof of Theorem 3.4 tackles a case where the eigenvalues of the covariance matrices involved are not upper bounded as $n \to \infty$. To the best of our knowledge, this situation has never been addressed in the increasing-domain asymptotic literature.

We conclude this paper by a Monte Carlo simulation, illustrating Theorems 3.3 and 3.4. The simulation highlights that the ML and CV estimators can estimate radically different covariance parameters, and that their subsequent performances for the Kullback-Leibler divergence and the integrated square prediction error can be strongly contrasting.

The rest of the paper is organized as follows. We present the context on parametric covariance function estimation in the misspecified case and on the spatial sampling in Section 2. We give the asymptotic optimality results for ML and CV in Section 3. We discuss the simulation results in Section 4. All the proofs are given in the appendix.

2 Context

2.1 Presentation and notation for the covariance model

We consider a stationary Gaussian process Y on $\mathbb{R}^d$ with zero mean function and covariance function K_0. Noisy observations of Y are obtained at the random points $X_1, ..., X_n \in \mathbb{R}^d$, for $n \in \mathbb{N}^*$. That is, for $i = 1, ..., n$, we observe $y_i = Y(X_i) + \epsilon_i$, where $\epsilon = (\epsilon_1, ..., \epsilon_n)^t$, Y and $(X_1, ..., X_n)$ are mutually independent and ϵ follows a $\mathcal{N}(0, \delta_0 I_n)$ distribution, with $\delta_0 \geq 0$ and I_n the identity matrix of size n. The distribution of $(X_1, ..., X_n)$ is specified and discussed in Condition 2.3 below.

The case where Y is observed exactly is treated by this framework by letting $\delta_0 = 0$. Otherwise, letting $\delta_0 > 0$ can correspond for instance to measure errors (Bachoc et al., 2014) or to Monte Carlo computer experiments (Le Gratiet and Garnier, 2014). Note also that the case of a Gaussian process with discontinuous covariance function at 0 (nugget effect) is mathematically equivalent to this framework if the observation points $X_1, ..., X_n$ are two by two distinct. [This is the case in this paper, in an almost sure sense, see Condition 2.3.]

Let $p \in \mathbb{N}^*$ and let Θ be the compact subset $[\theta_{inf}, \theta_{sup}]^p$ with $-\infty < \theta_{inf} < \theta_{sup} < +\infty$. We consider a parametric model attempting to approximate the covariance function K_0 and the noise variance δ_0: $\{(K_\theta, \delta_\theta), \theta \in \Theta\}$, with K_θ a stationary covariance function and $\delta_\theta > 0$. We call the case where there exists $\theta_0 \in \Theta$ so that $(K_0, \delta_0) = (K_{\theta_0}, \delta_{\theta_0})$ the well-specified case. The converse case, where $(K_0, \delta_0) \neq (K_\theta, \delta_\theta)$ for all $\theta \in \Theta$ is called the misspecified case.

The well-specified case has been extensively studied in the Gaussian process literature, see the references given in Section 1. Nevertheless, the misspecified case can occur in many practical applications. Indeed, even if we assume $\delta_\theta = \delta_0$ for all θ, the standard covariance models $\{K_\theta, \theta \in \Theta\}$ are often driven by a limited number of parameters and thus restricted in some ways. For instance, one common practice (Martin and Simpson, 2004) is to use the Gaussian covariance model, where $p = d + 1$, $\Theta \subset (0, \infty)^p$, $\theta = (\sigma^2, \ell_1, ..., \ell_d)$ and $K_\theta(t) = \exp(-\sum_{i=1}^d t_i^2/\ell_i^2)$. With the Gaussian covariance model, all the covariance functions K_θ generate Gaussian process realizations that are almost surely infinitely differentiable. Thus, the Gaussian model is de facto misspecified if the realizations of Y have only a finite order of differentiability. In theory, the Matérn model considered in Section 4 brings a solution to this problem, by incorporating a tunable smoothness parameter $\nu > 0$. However, it is also common practice to enforce a priori this parameter ν to a fixed value (e.g. 3/2 in Chevalier et al. (2014)).

In this paper, we are primarily interested in analyzing the misspecified case and stressing that the conclusions for it differ from those that are usually derived from the well-specified case. Nevertheless, the asymptotic results that are given in Section 3 actually do not assume one of the two cases and are valid for both.

We let $X = (X_1, ..., X_n)$ be the random n-tuple of the n observation points. For $\theta \in \Theta$, we define the $n \times n$ random matrix R_θ by $(R_\theta)_{i,j} = K_\theta(X_i - X_j) + \delta_\theta \mathbf{1}_{i=j}$. We define the $n \times n$ random matrix R_0 by $(R_0)_{i,j} = K_0(X_i - X_j) + \delta_0 \mathbf{1}_{i=j}$. We do not write explicitly the dependence of R_θ and R_0 with respect to X and n. We define the random vector $y = (y_1, ..., y_n)^t$ of size n by $y_i = Y(X_i) + \epsilon_i$. Then, conditionally to X, y follows a $\mathcal{N}(0, R_0)$ distribution and is assumed to follow a $\mathcal{N}(0, R_\theta)$ distribution under parameter θ. We do not write explicitly the dependence of y with respect to X, n and ϵ.

2.2 Maximum Likelihood and Cross Validation estimators

The Maximum Likelihood (ML) estimator is defined by $\hat{\theta}_{ML} \in \text{argmin}_\theta \, L_\theta$, with

$$L_\theta := \frac{1}{n} \log\left(\det\left(R_\theta\right)\right) + \frac{1}{n} y^t R_\theta^{-1} y. \tag{1}$$

Remark 2.1. *For concision, we do not write explicitly the dependence of L_θ on X, n, Y and ϵ. We make the same remark for the CV criterion in* (2) *and* (3)*.*

Remark 2.2. *In this paper, we enable the criterion* (1) *to have more than one global minimizer, in which case, the asymptotic results of Section 3 hold for any sequence of random variables $\hat{\theta}_{ML}$ minimizing it. The same remark can be made for the CV criterion* (2)*. We refer to Remark 2.1 in Bachoc (2014) for the existence of measurable minimizers of the ML and CV criteria.*

The functional L_θ is the modified opposite log-likelihood. It is globally admitted that, in the well-specified case, the ML estimator is preferable over most other potential estimators, both from asymptotic grounds and in practice. Under increasing-domain asymptotics, ML is consistent and asymptotically normal, with mean vector 0 and covariance matrix the inverse of the Fisher information matrix. This is shown in Mardia and Marshall (1984), assuming either some convergence conditions on the covariance matrices and their derivatives or gridded observation points. Similar results are provided for Restricted Maximum Likelihood in Cressie and Lahiri (1993, 1996). In Bachoc (2014) asymptotic normality is also shown for Maximum Likelihood, using only simple conditions on the covariance model and for observation points that constitute a randomly perturbed regular grid.

The Cross Validation (CV) estimator, minimizing the Leave One Out (LOO) mean square error is defined by $\hat{\theta}_{CV} \in \text{argmin}_\theta \, CV_\theta$, with

$$CV_\theta := \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_{i,\theta}\right)^2, \tag{2}$$

where $\hat{y}_{i,\theta} := \mathbb{E}_{\theta|X}(y_i | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$ is the LOO prediction of y_i with parameter θ. The conditional mean value $\mathbb{E}_{\theta|X}$ denotes the expectation with respect to the distribution of Y and ϵ with the covariance function K_θ and the variance δ_θ, given X.

Let $r_{i,\theta} = (K_\theta(X_i, X_1), ..., K_\theta(X_i, X_{i-1}), K_\theta(X_i, X_{i+1}), ..., K_\theta(X_i, X_n))^t$. Define $r_{i,0}$ similarly with K_0. Define the $(n-1) \times (n-1)$ covariance matrix $R_{i,\theta}$ as the matrix extracted from R_θ by deleting its line and column i. Define $R_{i,0}$ similarly with R_0. Then, with $y_{-i} = (y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)^t$, we have $\hat{y}_{i,\theta} = r_{i,\theta}^t R_{i,\theta}^{-1} y_{-i}$.

Note that $\hat{y}_{i,\theta}$ is invariant if K_θ and δ_θ are multiplied by a common positive constant. Thus, the CV criterion (2) is designed to select only a correlation function $K_\theta / K_\theta(0)$ and a corresponding relative noise variance $\delta_\theta / K_\theta(0)$. For the question of selecting the variance $K_\theta(0)$ by CV, two cases are possible, depending on the parametric model:

First, several parametric models can be written so that they satisfy the decomposition $\theta = \left(\sigma^2, \tilde{\theta}\right)$, with $\sigma^2 > 0$, $\tilde{\theta} \in \mathbb{R}^{p-1}$ and $(K_\theta, \delta_\theta) = (\sigma^2 \tilde{K}_{\tilde{\theta}}, \sigma^2 \tilde{\delta}_{\tilde{\theta}})$, with $\tilde{K}_{\tilde{\theta}}$ a stationary correlation function and $\tilde{\delta}_{\tilde{\theta}} \geq 0$. Hence, in this case, $\tilde{\theta}$ would be estimated by (2) in a first step. In a second step, σ^2 can be estimated as in Bachoc (2013) by the equation $\hat{\sigma}^2_{CV}(\tilde{\theta}) = (1/n) \sum_{i=1}^n \{y_i - \hat{y}_{i,\tilde{\theta}}\}^2 / c^2_{i,\tilde{\theta}}$, where $c^2_{i,\tilde{\theta}} = \text{var}_{\tilde{\theta}|X}\left(y_i | y_1, ..., y_{i-1}, y_{i+1}, ..., y_n\right)$ is the LOO predictive variance for y_i with parameters $\sigma^2 = 1$ and $\tilde{\theta}$. The conditional variance $\text{var}_{\tilde{\theta}|X}$ denotes the variance with respect to the distribution of Y and ϵ with the covariance function $\tilde{K}_{\tilde{\theta}}$ and the noise variance $\tilde{\delta}_{\tilde{\theta}}$, given X. Even if the decomposition $\theta = \left(\sigma^2, \tilde{\theta}\right)$ is possible, we address only the estimation of $\tilde{\theta}$ in this paper, that is given by (2).

Second, assume that the the aforementioned decomposition is not possible. One important example is the one addressed in Section 4, where a fixed value δ_1 of the noise variance is assumed, so that $\delta_\theta = \delta_1$ for all $\theta \in \Theta$. Then, usually, there is a unique parameter $\hat{\theta}_{CV}$ minimizing (2). In this case, the estimated $(K_{\hat{\theta}_{CV}}, \delta_{\hat{\theta}_{CV}})$ is relevant in the misspecified case only for providing the predictor $\mathbb{E}_{\hat{\theta}_{CV}|X}(Y(t)|y)$ of the value of Y at a new point t. Other quantities like $\text{var}_{\hat{\theta}_{CV}|X}(Y(t)|y)$, where $\text{var}_{\theta|X}$ denotes the variance under parameter θ given X, can be unreliable. Nevertheless, the predictor $\mathbb{E}_{\hat{\theta}_{CV}|X}(Y(t)|y)$ alone has the same applicability as many regression techniques like

kernel regression or neural network methods and can be used in a wide range of practical applications. Though one can consider other Cross Validation criteria to optimize, like the log predictive probability (Rasmussen and Williams (2006), chapter 5, Zhang and Wang (2010), Sundararajan and Keerthi (2001)), this paper is dedicated to the criterion (2), and toward analyzing it according to the accuracy of the predictor $\mathbb{E}_{\hat{\theta}_{CV}|X}(Y(t)|y)$.

The two aforementioned cases are treated in the same fashion in the rest of the paper. Indeed, we only work under the assumption that $\hat{\theta}_{CV}$ is a random variable minimizing (2) which is allowed to have more than one global minimizer (Remark 2.2).

The criterion (2) can be computed with a single matrix inversion, by means of virtual LOO formulas (see e.g Ripley (1981); Dubrule (1983)). These virtual LOO formulas yield, when writing $\mathrm{diag}\,(A)$ for the matrix obtained by setting to 0 all the off diagonal elements of a square matrix A,

$$CV_\theta := \frac{1}{n} y^t R_\theta^{-1} \,\mathrm{diag}\left(R_\theta^{-1}\right)^{-2} R_\theta^{-1} y, \tag{3}$$

which is useful both in practice (to compute the CV criterion quickly) and in the proofs for CV.

Finally, in Bachoc (2014) it is shown that, in the well-specified case, the CV estimator is consistent and asymptotically normal for estimating correlation parameters, under increasing-domain asymptotics. Nevertheless, in the well-specified case, ML is preferable as discussed above. In the misspecified case, this hierarchy can change, as shown numerically in Bachoc (2013). In Section 3, we give asymptotic grounds to this finding.

2.3 Random spatial sampling

We consider an increasing-domain asymptotic framework where the observation points are independent and uniformly distributed.

Condition 2.3. *For all $n \in \mathbb{N}^*$, the observation points $X_1, ..., X_n$ are random and follow independently the uniform distribution on $[0, n^{1/d}]^d$. The three variables Y, $(X_1, ..., X_n)$ and ϵ are mutually independent.*

Condition 2.3 constitutes an increasing-domain asymptotic framework in the sense that the volume of the observation domain is n and the average density of observation points is constant. Some authors define increasing-domain asymptotics by the condition that the minimum distance between two different observation points is bounded away from zero (e.g. Zhang and Zimmerman (2005)), which is not the case here. In Lahiri (2003) and Lahiri and Mukherjee (2004), the term increasing-domain is also used, when points are sampled randomly on a domain with volume proportional to n. Observation points that are sampled uniformly and independently constitute the archetype of an irregular spatial sampling. Thus, as discussed in Section 1, CV is expected to perform well under Condition 2.3. This is indeed the case, as is shown in Section 3.

3 Asymptotic optimality results

3.1 Technical assumptions

We shall assume the following condition for the covariance function K_0, which is satisfied in all the most classical cases, and especially for the Matérn covariance function. Let $|t| = \max_{i=1,...,d} |t_i|$.

Condition 3.1. *The covariance function K_0 is stationary and continuous on $\mathbb{R}^d$. There exists $C_0 < +\infty$ so that for $t \in \mathbb{R}^d$,*

$$|K_0(t)| \leq \frac{C_0}{1 + |t|^{d+1}}.$$

Next, the following condition for the parametric set of covariance functions and noise variances is slightly non-standard but not restrictive. We discuss it below.

Condition 3.2. *For all $\theta \in \Theta$, the covariance function K_θ is stationary. For all fixed $t \in \mathbb{R}^d$, $K_\theta(t)$ is $p+1$ times continuously differentiable with respect to θ. For all $i_1, ..., i_p \in \mathbb{N}$ so that $i_1 + ... + i_p \leq p+1$, there exists $A_{i_1,...,i_p} < +\infty$ so that for all $t \in \mathbb{R}^d$, $\theta \in \Theta$,*

$$\left| \frac{\partial^{i_1}}{\partial \theta_1^{i_1}} ... \frac{\partial^{i_p}}{\partial \theta_p^{i_p}} K_\theta(t) \right| \leq \frac{A_{i_1,...,i_p}}{1 + |t|^{d+1}}.$$

There exists a constant $C_{inf} > 0$ so that, for any $\theta \in \Theta$, $\delta_\theta \geq C_{inf}$. Furthermore, δ_θ is $p+1$ times continuously differentiable with respect to θ. For all $i_1, ..., i_p \in \mathbb{N}$ so that $i_1 + ... + i_p \leq p+1$, there exists $B_{i_1,...,i_p} < +\infty$ so that for all $\theta \in \Theta$,

$$\left| \frac{\partial^{i_1}}{\partial \theta_1^{i_1}} ... \frac{\partial^{i_p}}{\partial \theta_p^{i_p}} \delta_\theta \right| \leq B_{i_1,...,i_p}.$$

In Condition 3.2, we require a differentiability order of $p+1$ for K_θ and δ_θ with respect to θ. In the related context of Bachoc (2014), where a well-specified covariance model is studied, consistency of ML and CV can be proved with a differentiability order of 1 only. [One can check that the proofs of Propositions 3.1 and 3.4 in Bachoc (2014) require only the first order partial derivatives of the Likelihood function.] The reason for this difference is that, as discussed after Theorem 3.4, an additional technical difficulty is present here, compared to Bachoc (2014). The specific approach we use requires the condition of differentiability order of $p+1$ and we leave open the question of relaxing it. Note, anyway, that many parametric covariance models are infinitely differentiable with respect to the covariance parameters, especially the Matérn model. In Condition 3.2, assuming that the covariance function and its derivatives vanish with distance with a polynomial rate of order $d+1$ is not restrictive. Indeed, many covariance functions vanish at least exponentially fast with distance.

Finally, the condition that the noise variance δ_θ is lower bounded uniformly in θ is crucial for our proof methods. Since we address here noisy observations of Gaussian processes, this condition is reasonable so that the results of Section 4 can cover a large variety of practical situations, some of which are listed in Section 2. Note that even when the Gaussian process under consideration is observed exactly, it can be desirable to incorporate an instrumental positive term δ_θ in the parametric model, for numerical reasons or for not interpolating exactly the observed values (Andrianakis and Challenor, 2012). Thus, Condition 3.2 could also be considered for Gaussian processes that are observed without noise.

3.2 Maximum Likelihood

In this paper, the analysis of the ML estimator in the misspecified case is based on the Kullback-Leibler divergence of the distribution of y assumed under $(K_\theta, \delta_\theta)$, for $\theta \in \Theta$, from the true distribution of y. More precisely, conditionally to X, y has a $\mathcal{N}(0, R_0)$ distribution and is assumed to have a $\mathcal{N}(0, R_\theta)$ distribution. The Kullback-Leibler divergence of the latter distribution from the former is, after multiplication by $2/n$,

$$D_{n,\theta} := \frac{1}{n} \left\{ \log \left(\det \left(R_\theta R_0^{-1} \right) \right) + \mathrm{Tr} \left(R_0 R_\theta^{-1} \right) \right\} - 1. \tag{4}$$

The normalized Kullback-Leibler divergence in (4) is equal to 0 if and only if $R_\theta = R_0$ and is strictly positive otherwise. It is interpreted as an error criterion for using $(K_\theta, \delta_\theta)$ instead of (K_0, δ_0), when making inference on the Gaussian process Y.

Note that $D_{n,\theta}$ is here appropriately scaled so that, if for a fixed θ $(K_\theta, \delta_\theta) \neq (K_0, \delta_0)$, $D_{n,\theta}$ should generally not vanish, nor diverge to infinity under increasing-domain asymptotics. This can be shown for instance in the framework of Bachoc (2014), by using the methods employed there. It is also well-known that, in the case of a regular grid of observation points for $d = 1$, $D_{n,\theta}$ converges to a finite limit as $n \to +\infty$ (Azencott and Dacunha-Castelle, 1986). This limit is twice the asymptotic Kullback information in Azencott and Dacunha-Castelle (1986) and is positive if $(K_\theta(t), \delta_\theta)$ differs from $(K_0(t), \delta_0)$ for at least one point t in the regular grid of observation points. Similarly, in the spatial sampling framework of Condition 2.3, we observe in the Monte Carlo simulations of Section 4 that the order of magnitude of (4) does not change when n increases, for $(K_\theta, \delta_\theta) \neq (K_0, \delta_0)$.

The following theorem shows that the ML estimator asymptotically minimizes the normalized Kullback-Leibler divergence.

Theorem 3.3. *Under Conditions 2.3, 3.1 and 3.2, we have, as $n \to \infty$,*

$$D_{n,\hat{\theta}_{ML}} = \inf_{\theta \in \Theta} D_{n,\theta} + o_p(1),$$

where the $o_p(1)$ in the above display is a function of X and y only that goes to 0 in probability as $n \to \infty$.

Theorem 3.3 is in line with the well-known fact that, in the *iid* setting, ML asymptotically minimizes the Kullback-Leibler divergence (than does not depend on sample size) from the true distribution, within a misspecified parametric model (White, 1982). Theorem 3.3 conveys a similar message, with the normalized Kullback-Leibler divergence that depends on the spatial sampling. As discussed above, the infimum in Theorem 3.3 is typically lower bounded as $n \to \infty$ in the misspecified case.

Note that Theorem 3.3 can be shown, in increasing-domain asymptotics, under other spatial samplings than that of Condition 2.3 (e.g. for the randomly perturbed regular grid of Bachoc (2014)). Nevertheless, to the best of our knowledge, in the context of Condition 2.3, Theorem 3.3 is not a simple consequence of the existing literature, and an original proof is provided in Section A.2.

An important message of this paper is that the normalized Kullback-Leibler divergence is one error criterion among others and may not be given a particular priority. Indeed, most importantly, this criterion addresses the distribution of the Gaussian process only at the observation points. Hence, the minimality of $D_{n,\hat{\theta}_{ML}}$ gives no information on the optimality of the inference of the values of Y at new points, obtained from $(K_{\hat{\theta}_{ML}}, \delta_{\hat{\theta}_{ML}})$. Especially, $\hat{\theta}_{ML}$ is typically not optimal for the prediction error at new points, as is illustrated in Section 4.

3.3 Cross Validation

Let $\hat{y}_\theta(t) = \mathbb{E}_{\theta|X}(Y(t)|y)$. With the $n \times 1$ vector $r_\theta(t)$ so that $(r_\theta(t))_j = K_\theta(t - X_j)$, we have $\hat{y}_\theta(t) = r_\theta^t(t) R_\theta^{-1} y$. Then, define the family of random variables

$$E_{n,\theta} = \frac{1}{n} \int_{[0,n^{1/d}]^d} \left(\hat{y}_\theta(t) - Y(t)\right)^2 dt, \tag{5}$$

where the integral is defined in the L^2 sense since K_0 is continuous. We call the criterion (5) the integrated square prediction error. It is natural to consider that the first objective of the CV estimator $\hat{\theta}_{CV}$ is to yield a small $E_{n,\hat{\theta}_{CV}}$. If the observation points $X_1, ..., X_n$ are regularly spaced, then this objective might however not be fulfilled. Indeed, the principle of CV does not really have grounds in this case, since the LOO prediction errors are not representative of actual prediction errors for new points. This fact is only natural and has been noted in e.g. Iooss et al. (2010) and Bachoc (2013). If however the observation points $X_1, ..., X_n$ are not regularly spaced, then it is shown numerically in Bachoc (2013) that the CV estimator $\hat{\theta}_{CV}$ can yield a small $E_{n,\hat{\theta}_{CV}}$ and, especially, smaller than $E_{n,\hat{\theta}_{ML}}$. The following theorem, which is the main contribution of this paper, supports this conclusion under increasing-domain asymptotics.

Theorem 3.4. *Under Conditions 2.3, 3.1 and 3.2, we have, as* $n \to \infty$*,*

$$E_{n,\hat{\theta}_{CV}} = \inf_{\theta \in \Theta} E_{n,\theta} + o_p(1),$$

where the $o_p(1)$ *in the above display is a function of* X*,* y *and* Y *only that goes to* 0 *in probability as* $n \to \infty$*.*

Before discussing further the content of Theorem 3.4, we say a few words about its proof (Section A.3). The main difference compared to Bachoc (2014) is that, because the observation points are independent, there is no minimum distance between two different observation points and clusters of closely spaced observation points may appear. As a consequence, the maximum eigenvalues of, say, the matrices R_θ, are not upper bounded, even in probability. To the best of our knowledge, no increasing-domain asymptotic results for Gaussian processes are yet available in the literature that both address spatial sampling without minimal distance between two different observation points and require only assumptions on the covariance functions, as Conditions 3.1 and 3.2 do. Especially, the references Mardia and Marshall (1984); Cressie and Lahiri (1993, 1996) work under non-trivial assumptions on the covariance matrices involved, and show that these assumptions are fulfilled for examples of spatial samplings for which the minimal distance between two different observation points is bounded away from zero. More recently, also Zheng and Zhu (2012) assumes that the largest eigenvalues of the covariance matrices in the parametric model are bounded as $n \to \infty$. In Bachoc (2014), the working assumptions concern only the covariance model, but there is also a minimum distance between two different observation points. Note finally that in Lahiri (2003); Lahiri and Mukherjee (2004); Lahiri and Zhu (2006), random spatial sampling with independent observation points is addressed, under increasing and mixed increasing-domain asymptotics.

Nevertheless, these references do not address covariance parameter estimation and do not deal with M-estimation where the criterion function involves $n \times n$ random matrices.

As a consequence, the proof we propose for Theorem 3.4 is original and we do not address the asymptotic distribution of the ML and CV estimators. We leave this problem open to further research. Note nevertheless that, in the misspecified case addressed here, the fact that the ML and CV estimators minimize two different criteria and are thus typically asymptotically different is, in our opinion, at least as important as their asymptotic distributions.

An important element in the proof of Theorem 3.4 is that the variable t in the expression of the integrated square prediction error $E_{n,\theta}$ in (5) plays the same role as a new point X_{n+1}, uniformly distributed on $[0, n^{1/d}]^d$ and independent of $(X_1, ..., X_n)$. Hence, using the symmetry of $X_1, ..., X_{n+1}$, for fixed θ, the mean value of $E_{n,\theta}$ is equal to the mean value of a modification of the CV criterion CV_θ in (2), where there are $n+1$ observation points instead of n. Thus, one can indeed expect that the CV estimator minimizing CV_θ also asymptotically minimizes $E_{n,\theta}$. [The challenging part for proving Theorem 3.4 is to control the deviations of the criteria $E_{n,\theta}$ and CV_θ from their mean values, uniformly in θ.] This discussion is exactly the paradigm of CV, that uses the LOO errors as empirical versions of the actual prediction errors. On the other hand, if the observation points constitute for instance a regular grid, then the variable t in $E_{n,\theta}$ has close to nothing in common with them, so that Theorem 3.4 would generally not hold. This stresses that CV is generally not efficient for regular sampling of observation points, as discussed above.

In (5), we stress that $E_{n,\theta}$ and the observation vector y are defined with respect to the same Gaussian process Y. Thus, Theorem 3.4 gives a guarantee for the estimator $\hat{\theta}_{CV}$ relatively to the predictions it yields for the actual Gaussian process at hand.

Theorems 3.3 and 3.4, when considered together with the results in Bachoc (2014), draw the following conclusion, on comparing ML and CV for covariance parameter estimation: In the well-specified case, these two estimators can be compared under the common criterion of estimation error and ML is preferable over CV. On the other hand, in the misspecified case, there is not a unique comparison criterion. ML is asymptotically optimal for the normalized Kullback-Leibler divergence (which does not address prediction of the Gaussian process Y at new points) while, under the random sampling addressed here, CV is more efficient than ML for the prediction mean square error. This last finding also supports some references listed in Section 1 addressing specific situations similar to the misspecified case.

Beyond a comparison of ML and CV, Theorems 3.3 and 3.4 show that both estimators can not do significantly worse asymptotically than any other potential estimators for their respective error criteria. This last fact holds for ML under fairly general samplings of observation points, as we have discussed, but necessitates here a purely random sampling for CV.

Finally, the fact that ML and CV typically optimize different criteria in the misspecified case can serve as a practical guideline. That is, one can compute the estimated covariance parameters with both methods and compare the two estimates and the corresponding log-likelihood and LOO mean square error values. If the differences between ML and CV are large, then it could be a warning that the covariance model at hand can be inappropriate.

4 Monte Carlo simulation

We illustrate Theorems 3.3 and 3.4 in a Monte Carlo simulation. We consider the Matérn covariance model in dimension $d = 1$. A covariance function on $\mathbb{R}$ is Matérn (σ^2, ℓ, ν) when it is written

$$K_{\sigma^2,\ell,\nu}(t) = \frac{1}{\Gamma(\nu)\, 2^{\nu-1}} \left(2\sqrt{\nu}\frac{|t|}{\ell}\right)^{\nu} K_\nu \left(2\sqrt{\nu}\frac{|t|}{\ell}\right),$$

with Γ the Gamma function and K_ν the modified Bessel function of second order. The parameters σ^2, ℓ and ν are respectively the variance, correlation length and smoothness parameters. We refer to e.g Stein (1999) for a presentation of the Matérn covariance function.

In the simulation, the true covariance function of Y is Matérn $(\sigma_0^2, \ell_0, \nu_0)$ with $\sigma_0^2 = 1$, $\ell_0 = 3$ and $\nu_0 = 10$. This choice of ν_0 corresponds to a smooth Gaussian process and enables, as we see below, to illustrate Theorems 3.3 and 3.4 in a more striking manner. The true noise variance is $\delta_0 = 0.25^2$.

For the covariance and noise variance model, it is considered that the smoothness parameter ν_0 is known, that the noise variance δ_1 is fixed, and that the parameter $\theta = (\sigma^2, \ell)$ is estimated by ML or CV. For both ML and CV, the optimization is restricted to the domain $\Theta = [0.1^2, 10^2] \times [0.2, 10]$. [We experience that the conclusions of the Monte Carlo simulation are the same if a larger optimization domain is considered.] The well-specified case corresponds to $\delta_1 = \delta_0$ and the misspecified case corresponds to $\delta_1 = 0.1^2 \neq \delta_0$. This covariance model is representative of practical applications. Indeed, first it is common practice to fix the value of the smoothness parameter in the Matérn model, as is discussed in Section 2. Second, when using Gaussian process models on experimental or natural data, it can often occur that field experts provide an a priori value for the noise variance (see e.g. Bachoc et al. (2014)). The misspecified case we address corresponds to an underestimation of the noise variance, possibly because some sources of measurement errors have been neglected.

The Monte Carlo simulation is carried out as follows. For $n = 100$ and $N = 1000$ or $n = 500$ and $N = 200$ we repeat N data generations, estimations and quality criterion computations and average the results. More specifically, we simulate N independent realizations of the observation points, of the observation vector and of the Gaussian process on $[0, n]$, under the true covariance function and noise variance. For each of these N realizations, we compute the ML and CV estimates under the well-specified and misspecified models. For each of these estimates of the form $\hat{\theta} = (\hat{\sigma}^2, \hat{\ell})$, we compute the corresponding criteria $D_{n,\hat{\sigma}^2,\hat{\ell}}$ and $E_{n,\hat{\sigma}^2,\hat{\ell}}$.

In Figure 1 we report, for $n = 100$, for the well-specified and misspecified cases and for ML and CV, the histograms of the estimates $\hat{\ell}$, and of the values of the error criteria $D_{n,\hat{\sigma}^2,\hat{\ell}}$ and $E_{n,\hat{\sigma}^2,\hat{\ell}}$. In the well-specified case, the conclusions are in agreement with the main message of previous literature: Both estimators estimate the true $\ell_0 = 3$ with reasonable accuracy and have error criteria that are relatively small. By also considering Table 1, where the averages and standard deviations corresponding to Figures 1 and 2 are reported, we observe that ML performs better than CV in all aspects. The estimation error for ℓ and the normalized Kullback-Leibler divergence are significantly smaller for ML, while the integrated square prediction error is similar under ML and CV estimation, but nonetheless smaller for ML.

The conclusions are however radically different in the misspecified case, as is implied by Theorems 3.3 and 3.4. First, the ML estimates of ℓ are significantly smaller than in the well-specified case, and can even be equal to the lower-bound 0.2. The ML estimates of σ^2 are not reported in Figure 1 to save space and are close to 1, so that, approximately, the variance of the observations, as estimated by ML, is close to the true variance of the observations. The reason for these small estimates of ℓ by ML is the underestimation of the noise variance δ_0, coupled with the large smoothness parameter ν_0. Indeed, there exist pairs of closely spaced observation points for which the corresponding differences of observed values are large compared to δ_1, so that for values of ℓ that are larger than those computer by ML, the criterion (1) blows up, for all values of σ^2. [Using a value of σ^2 smaller or approximately equal to 1 does not counterbalance the damaging impact on (1) of these pairs of closely spaced observation points with large observed value differences. Increasing σ^2 over 1 is also not optimal for (1), since on a large scale, the observations do have variances close to 1.] This phenomenon for ML is all the more important when the smoothness parameter ν_0 is large, which is why we choose here the value $\nu_0 = 10$ to illustrate it. To summarize, ML gives an important weight to pairs of closely spaced observation points with large observation differences and consequently estimates small correlation lengths to explain, so to speak, these observation differences.

On the contrary for CV, if we consider only the predictions $\hat{y}_{\sigma^2,\ell}(t)$ at new points t and the LOO predictions $\hat{y}_{i,\sigma^2,\ell}$, with $(\ell, \sigma^2) \in \Theta$, then the situation is virtually the same as if the model was well-specified. Indeed, the covariance matrices and vectors obtained from σ^2, ℓ and δ_0 are equal to δ_0/δ_1 time those obtained from $\sigma^2\delta_1/\delta_0$, ℓ and δ_1, so that the corresponding predictions are identical. Hence, in Figure 1, the empirical distribution of $\hat{\ell}_{CV}$ is approximately the same between the well-specified and misspecified cases. In the misspecified case, we find that the empirical distribution of $\hat{\sigma}^2_{CV}$ (not reported in Figure 1 to save space) is δ_1/δ_0 time that of the well-specified case. Of course, although the CV predictions are not damaged by the misspecified δ_1, the CV estimations of other characteristics of the conditional distribution of Y given the observed data are damaged. [For example the confidence intervals for $Y(t)$ obtained from the CV estimates are significantly too small.]

The histograms of $E_{n,\hat{\sigma}^2,\hat{\ell}}$ and $D_{n,\hat{\sigma}^2,\hat{\ell}}$ for ML and CV in Figure 1 confirm the discussion on the estimated parameters. For ML which estimates small correlation lengths, the error criteria $E_{n,\hat{\sigma}^2_{ML},\hat{\ell}_{ML}}$ are significantly larger than in the well-specified case (by an approximate factor of 3 on average as seen in Table 1). The error criteria $D_{n,\hat{\sigma}^2_{ML},\hat{\ell}_{ML}}$ also increase and become larger than these of both ML and CV in the well-specified case. For CV, the error criteria $E_{n,\hat{\sigma}^2_{CV},\hat{\ell}_{CV}}$ are, as discussed, as small as in the well-specified case and approximately

n	Specification	Estimation	Average of $\hat{\ell}$	Standard deviation of $\hat{\ell}$	Average of $E_{n,\hat{\sigma}^2,\hat{\ell}}$	Average of $D_{n,\hat{\sigma}^2,\hat{\ell}}$
100	Well-specified	ML	3.031	0.370	0.073	0.023
	Well-specified	CV	3.447	1.159	0.085	0.222
	Misspecified	ML	1.090	0.553	0.255	1.025
	Misspecified	CV	3.525	1.324	0.087	3.563
500	Well-specified	ML	3.003	0.177	0.070	0.005
	Well-specified	CV	3.084	0.399	0.071	0.037
	Misspecified	ML	0.975	0.269	0.247	0.972
	Misspecified	CV	3.089	0.437	0.071	3.442

Table 1: For the settings of Figures 1 and 2, the averages and standard-deviations of $\hat{\ell}$, $D_{n,\hat{\sigma}^2,\hat{\ell}}$ and $E_{n,\hat{\sigma}^2,\hat{\ell}}$ are reported.

3 times smaller on average than for ML, illustrating Theorem 3.4. However, the error criteria $D_{n,\hat{\sigma}^2_{CV},\hat{\ell}_{CV}}$ are 3 times larger for CV than for ML, in the misspecified case, illustrating Theorem 3.3.

Finally, in Figure 2, the settings are the same as for Figure 1 but for $n = 500$. The conclusions on the comparison between ML and CV are the same as for $n = 100$. Then, the estimates of ℓ under ML and CV have less variance than for $n = 100$, and their histograms are approximately unimodal and symmetric. Finally, for ML and CV in the misspecified case, $E_{n,\hat{\sigma}^2,\hat{\ell}}$ and $D_{n,\hat{\sigma}^2,\hat{\ell}}$ keep the same averages between $n = 100$ and $n = 500$. In the well-specified case, $E_{n,\hat{\sigma}^2,\hat{\ell}}$ also keeps the same average, while $D_{n,\hat{\sigma}^2,\hat{\ell}}$ becomes very small. This is because $D_{n,\sigma_0^2,\ell_0} = 0$ in the well-specified case, while E_{n,σ_0^2,ℓ_0} is non-zero and should not vanish to 0 as $n \to \infty$, since the density of observation points in the prediction domain is constant with n.

5 Discussion

We prove that for independent and uniformly distributed observation points, CV asymptotically yields the smallest integrated square prediction error among all possible parameters. On the contrary ML is asymptotically optimal for the Kullback-Leibler divergence which does not provide information for prediction at new points. In the case addressed in the Monte Carlo simulation, it is clear that the covariance parameters estimated by ML and CV, and their corresponding performances for the integrated square prediction error and the Kullback-Leibler divergence are significantly different. Thus, the contrast between the well-specified case, where both estimators converge to the true covariance parameter, is striking.

For regular sampling, the principle of CV does not really have grounds while ML would typically still asymptotically minimize the Kullback-Leibler divergence. Note however that regular sampling can still damage the ML estimation and make model misspecification more difficult to detect (see Bachoc (2014) and the references discussed in its introduction and conclusion).

Finally, the results of the Monte Carlo simulation make it conceivable that, for independent and uniform observation points, the ML and CV estimators converge to optimal parameters, for respectively the Kullback-Leibler divergence and the integrated square prediction error, and are asymptotically normal. [These optimal parameters would be equal to the true ones in the well-specified case.] Considering asymptotic normality might require new techniques to account for the absence of minimum distance between different observation points. We leave this problem open to future research.

A Proofs

A.1 Notation

In all the appendix, we consider that Conditions 2.3, 3.1 and 3.2 hold. For a column vector v of size m, we let $||v||^2 = \sum_{i=1}^m v_i^2$ and $|v| = \max_{i=1,\ldots,m} |v_i|$. For a real $m \times m$ matrix A, we write as in Gray (2006),

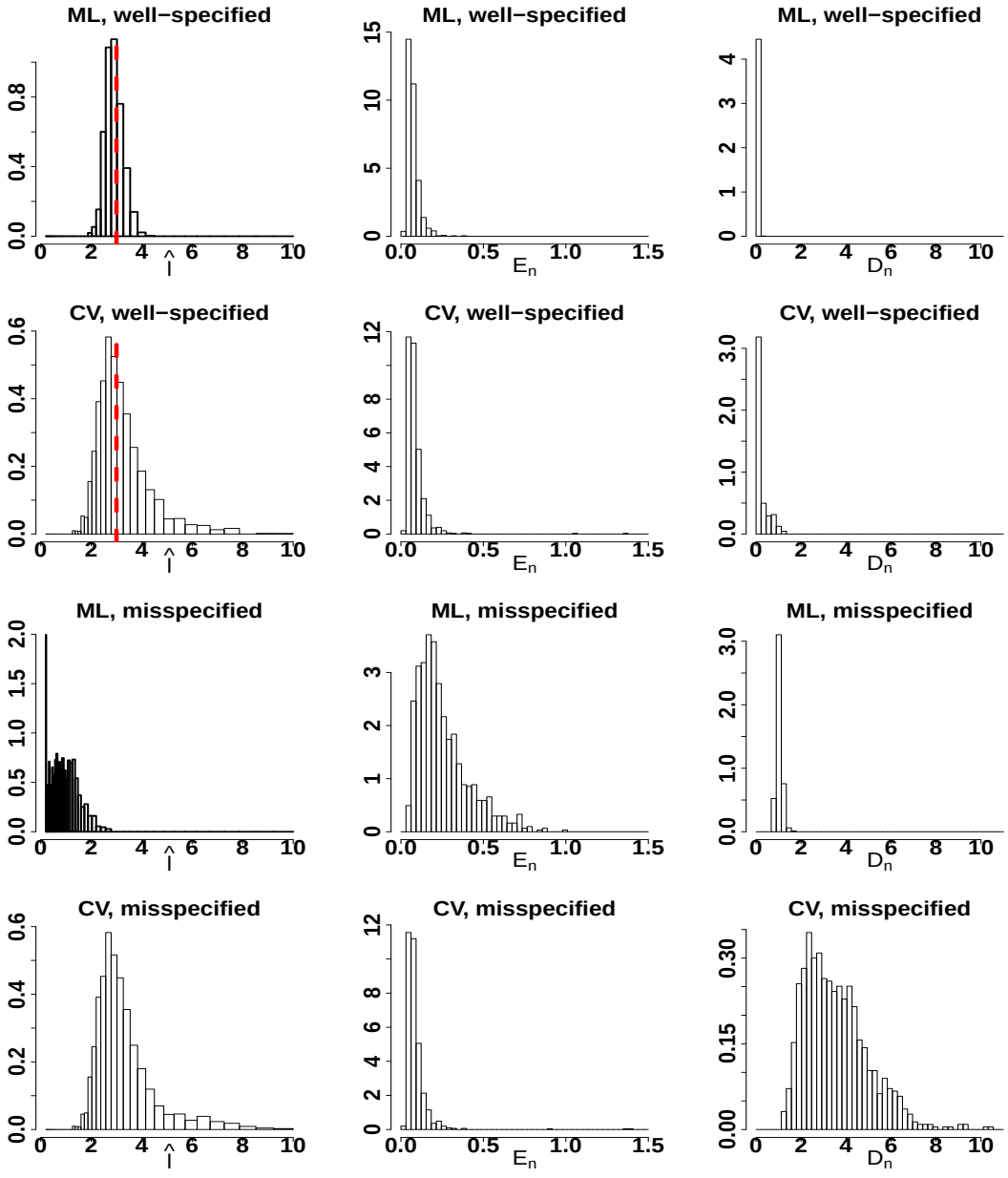

Figure 1: Simulation of $N = 1000$ independent realizations of $n = 100$ *iid* observation points with uniform distribution on $[0, n]$, of the Gaussian process Y on $[0, n]$ with Matérn ($\sigma_0^2 = 1, \ell_0 = 3, \nu_0 = 10$) covariance function, and of the corresponding observation vector with noise variance $\delta_0 = 0.25^2$. For each simulation, ν_0 is known, the noise variance is fixed to $\delta_1 = \delta_0$ (well-specified case) or $\delta_1 = 0.1^2 \neq \delta_0$ (misspecified case), σ^2 and ℓ are estimated by ML and CV and the corresponding error criteria $D_{n,\hat{\sigma}^2,\hat{\ell}}$ (normalized Kullback-Leibler divergence) and $E_{n,\hat{\sigma}^2,\hat{\ell}}$ (integrated square prediction error) are computed. The histograms of the N estimates of ℓ and of the N corresponding values of the error criteria are reported for ML and CV and in the well-specified and misspecified cases. In the well-specified case, the estimates are on average reasonably close to the true values, the error criteria are reasonably small and ML performs better than CV in all aspects. In the misspecified case, the ML and CV estimates of the correlation lengths are significantly different, ML performs better than CV for $D_{n,\hat{\sigma}^2,\hat{\ell}}$ and CV performs better than ML for $E_{n,\hat{\sigma}^2,\hat{\ell}}$.

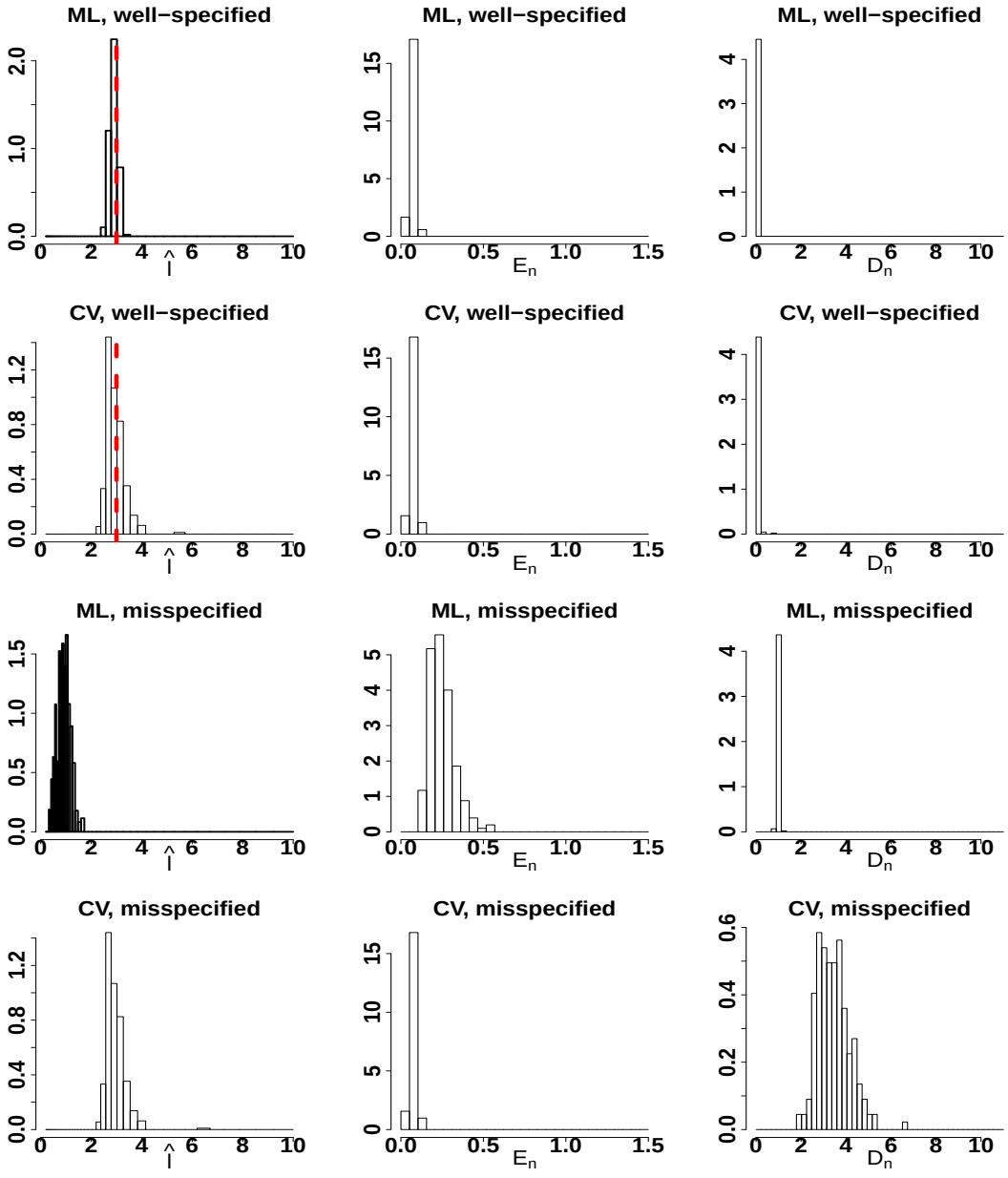

Figure 2: Same settings as for Figure 1 but with $N = 200$ and $n = 500$. The conclusions for the comparison of ML and CV are the same as for Figure 1. The variances of the estimates $\hat{\ell}$ and of the error criteria are smaller than in Figure 1 and the histograms of the estimates are approximately symmetric and unimodal. Between Figure 1 and Figure 2, all error criteria keep the same order of magnitude except $D_{n,\hat{\sigma}^2,\hat{\ell}}$ in the well-specified case which decreases and becomes very small for ML and CV (see also Table 1). This is because in the well-specified case $D_{n,\sigma_0^2,\ell_0} = 0$.

$|A|^2 = \frac{1}{m}\sum_{i,j=1}^m A_{i,j}^2$ and $||A||$ for the largest singular value of A. Both $|.|$ and $||.||$ are norms and $||.||$ is also a matrix norm.

For a sequence of real random variables z_n, we write $z_n \to_p 0$ and $z_n = o_p(1)$ when z_n converges to zero in probability. For a random variable A and a deterministic function $f(A)$, we may write $\mathbb{E}_A(f(A))$ for $\mathbb{E}(f(A))$. For two random variables A and B and a deterministic function $f(A,B)$ we may write $\mathbb{E}_{A|B}(f(A,B))$ for $\mathbb{E}(f(A,B)|B)$.

For a finite set E, we write $|E|$ for its cardinality. For a continuous set $E \subset \mathbb{R}^d$, we write $|E|$ for its Lebesgue measure. For two sets A, B in $\mathbb{R}^d$, we write $d(A,B) = \inf_{a\in A, b\in B}|a-b|$.

We write C_{sup} a generic non-negative finite constant (not depending on n, X, Y, ϵ and θ). The actual value of C_{sup} is of no interest and can change in the same sequence of equations. For instance, instead of writing, say, $a \leq 2b \leq 4c$, we shall write $a \leq C_{sup}b \leq C_{sup}c$. Similarly, we write C_{inf} a generic strictly positive constant (not depending on n, X, Y, ϵ and θ).

A.2 Proofs for Maximum Likelihood

Lemma A.1. *For $i = 1, ..., p$,*

$$\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}L_\theta\right|\right)$$

is bounded w.r.t n.

Proof of Lemma A.1.

$$\begin{aligned}
\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}L_\theta\right|\right) &= \mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{1}{n}\operatorname{Tr}\left(R_\theta^{-1}\frac{\partial}{\partial\theta_i}R_\theta\right) - \frac{1}{n}y^t R_\theta^{-1}\left(\frac{\partial}{\partial\theta_i}R_\theta\right)R_\theta^{-1}y\right|\right) \\
\text{(Cauchy-Schwarz:)} &\leq \mathbb{E}\left(\sup_{\theta\in\Theta}\sqrt{\left|R_\theta^{-1}\right|^2}\sqrt{\left|\frac{\partial}{\partial\theta_i}R_\theta\right|^2}\right) \qquad (6) \\
&\quad + \sqrt{\frac{1}{n}\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\left|R_\theta^{-1}y\right|\right|^2\right)}\sqrt{\frac{1}{n}\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\left|\left(\frac{\partial}{\partial\theta_i}R_\theta\right)R_\theta^{-1}y\right|\right|^2\right)} \qquad (7)
\end{aligned}$$

Now, $\left|R_\theta^{-1}\right|^2 \leq \left|\left|R_\theta^{-1}\right|\right|^2$ because of (2.19) in Gray (2006) and $\left|\left|R_\theta^{-1}\right|\right|^2$ is bounded uniformly in θ because of Lemma A.18. Also, $\mathbb{E}\left(\sup_{\theta\in\Theta}\left|[\partial/\partial\theta_i]R_\theta\right|^2\right)$ is bounded because of Condition 3.2 and of a simple case of Lemma A.17. So the right-hand side of (6) is bounded because of Jensen inequality. It remains to show that the term (7) is bounded. To show this, note first that

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\left|R_\theta^{-1}y\right|\right|^2\right) &\leq \frac{1}{n}\mathbb{E}\left(\sup_{\theta\in\Theta}||y||^2\left|\left|R_\theta^{-1}\right|\right|^2\right) \\
\text{(Lemma A.18:)} &\quad \frac{C_{sup}}{n}\mathbb{E}\left(||y||^2\right) \\
&= C_{sup}(K_0(0)+\delta_0),
\end{aligned}$$

is bounded. Thus, it remains to show that $\frac{1}{n}\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\left|\left([\partial/\partial\theta_i]R_\theta\right)R_\theta^{-1}y\right|\right|^2\right)$ is bounded. For this, we have

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\left|\left(\frac{\partial}{\partial\theta_i}R_\theta\right)R_\theta^{-1}y\right|\right|^2\right) \\
&= \frac{1}{n}\mathbb{E}\left(\sup_{\theta\in\Theta}y^tR_\theta^{-1}\left(\frac{\partial}{\partial\theta_i}R_\theta\right)^2R_\theta^{-1}y\right) \\
&\leq C_{sup}\sum_{i_1+...+i_p\leq p}\int_\Theta\frac{1}{n}\mathbb{E}\left(\left|\frac{\partial^{i_1}}{\partial\theta_1^{i_1}}...\frac{\partial^{i_p}}{\partial\theta_p^{i_p}}\left[y^tR_\theta^{-1}\left(\frac{\partial}{\partial\theta_i}R_\theta\right)^2R_\theta^{-1}y\right]\right|\right)d\theta \quad \text{(Lemma A.15).}
\end{aligned}$$

Thus, it suffices to show that, for fixed $i_1, ..., i_p \in \mathbb{N}$ so that $i_1 + ... + i_p \leq p$,

$$\frac{1}{n} \sup_{\theta \in \Theta} \mathbb{E}\left(\left| \frac{\partial^{i_1}}{\partial \theta_1^{i_1}} ... \frac{\partial^{i_p}}{\partial \theta_p^{i_p}} \left[y^t R_\theta^{-1} \left(\frac{\partial}{\partial \theta_i} R_\theta \right)^2 R_\theta^{-1} y \right] \right| \right)$$

is bounded. The above display is smaller than a fixed sum of terms of the form $(1/n) \sup_{\theta \in \Theta} \mathbb{E}\left(|y^t M_\theta y|\right)$, where the number of terms is independent of n and M_θ is of the form $N_{1,\theta} M_{1,\theta} ... M_{k,\theta} N_{k+1,\theta}$ with $N_{i,\theta} = I_n$ or $N_{i,\theta} = R_\theta^{-1}$ and with $M_{i,\theta}$ of the form $[\partial^{c_1}/\partial \theta_1^{c_1}] ... [\partial^{c_p}/\partial \theta_p^{c_p}] R_\theta$ with $c_1, ..., c_p \in \mathbb{N}$ and $c_1 + ... + c_p \leq p+1$. Hence, it is enough to show that any term of the form $\sup_{\theta \in \Theta} (1/n) \mathbb{E}\left(|y^t M_\theta y|\right)$ above is bounded. We have

$$\begin{aligned}
&\sup_{\theta \in \Theta} \frac{1}{n} \mathbb{E}\left(|y^t M_\theta y|\right) \\
&\leq \sup_{\theta \in \Theta} \frac{1}{n} \mathbb{E}\left(\left|y^t M_\theta y - \mathbb{E}\left(y^t M_\theta y \middle| X\right)\right|\right) + \sup_{\theta \in \Theta} \frac{1}{n} \mathbb{E}\left(\left|\mathbb{E}\left(y^t M_\theta y \middle| X\right)\right|\right) \\
&\leq \sup_{\theta \in \Theta} \sqrt{\mathbb{E}\left(\operatorname{var}\left[\frac{1}{n} y^t M_\theta y \middle| X\right]\right)} + \sup_{\theta \in \Theta} \frac{1}{n} \mathbb{E}\left(\left|\mathbb{E}\left(y^t M_\theta y \middle| X\right)\right|\right) \quad \text{(Jensen inequality)} \\
&= \sup_{\theta \in \Theta} \sqrt{\mathbb{E}\left(\frac{1}{2n^2} \operatorname{Tr}\left[R_0 \left\{M_\theta + M_\theta^t\right\} R_0 \left\{M_\theta + M_\theta^t\right\}\right]\right)} + \sup_{\theta \in \Theta} \frac{1}{n} \mathbb{E}\left(\left|\operatorname{Tr}\left[R_0 M_\theta\right]\right|\right) \\
&\leq \sup_{\theta \in \Theta} \sqrt{\frac{1}{2n} \mathbb{E}\left(\left|R_0 \left\{M_\theta + M_\theta^t\right\}\right|^2\right)} + \sup_{\theta \in \Theta} \sqrt{\mathbb{E}\left(|R_0|^2\right) \mathbb{E}\left(|M_\theta|^2\right)} \quad \text{(Cauchy-Schwarz).} \\
&\leq \sup_{\theta \in \Theta} \sqrt{\frac{1}{n} \mathbb{E}\left(|R_0 M_\theta|^2 + |R_0 M_\theta^t|^2\right)} + \sup_{\theta \in \Theta} \sqrt{\mathbb{E}\left(|R_0|^2\right) \mathbb{E}\left(|M_\theta|^2\right)}
\end{aligned}$$

In the display above, the first term goes to 0 because of Conditions 3.1 and 3.2 and Lemmas A.17 and A.18. The second term is bounded because of Lemmas A.17, A.18 and A.21. This completes the proof.

□

Corollary A.2. *For any $i = 1, ..., p$,*

$$\mathbb{E}\left(\sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_i} \mathbb{E}\left(L_\theta \middle| X\right) \right| \right) \quad \text{and} \quad \sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_i} D_{n,\theta} \right|$$

are bounded w.r.t n.

Proof of corollary A.2. Note first that $[\partial/\partial \theta_i] \mathbb{E}\left(L_\theta \middle| X\right) = [\partial/\partial \theta_i] D_{n,\theta}$. The corollary is then a consequence of the fact that, for fixed n, we have $(\partial/\partial \theta_i) \mathbb{E}\left(L_\theta \middle| X\right) = \mathbb{E}\left((\partial/\partial \theta_i) L_\theta \middle| X\right)$ and of $\sup_\theta |\mathbb{E}(.)| \leq \mathbb{E}(\sup_\theta |.|)$. □

Lemma A.3. *Consider a fixed $\theta \in \Theta$. Then*

$$\mathbb{E}\left(|L_\theta - \mathbb{E}(L_\theta|X)|\right) \to_{n \to \infty} 0.$$

Proof of Lemma A.3. We have, applying Jensen inequality twice

$$\mathbb{E}\left(|L_\theta - \mathbb{E}(L_\theta|X)|\right) \leq \mathbb{E}\left(\sqrt{var(L_\theta|X)}\right) \leq \sqrt{\mathbb{E}\left(var(L_\theta|X)\right)} = \sqrt{\mathbb{E}\left(\frac{2}{n^2} \operatorname{Tr}\left[R_0 R_\theta^{-1} R_0 R_\theta^{-1}\right]\right)}.$$

The eigenvalues of R_θ^{-1} are smaller than a finite constant C_{sup} for any n, X, θ from Lemma A.18. Thus, by applying Cauchy schwarz inequality and Lemmas A.17 and A.21,

$$\mathbb{E}\left(|L_\theta - \mathbb{E}(L_\theta|X)|\right) \leq \sqrt{\frac{2}{n}} \sqrt{\mathbb{E}\left(|R_0 R_\theta^{-1}|^2\right)} \leq \frac{C_{sup}}{\sqrt{n}}.$$

□

Proof of Theorem 3.3. We have

$$\begin{aligned}
&\sup_{\theta\in\Theta} |L_\theta - \log(\det(R_0)) - 1 - D_{n,\theta}| \\
&\leq \sup_{\theta\in\Theta} |L_\theta - \mathbb{E}(L_\theta|X)| + \sup_{\theta\in\Theta} |\mathbb{E}(L_\theta|X) - \log(\det(R_0)) - 1 - D_{n,\theta}| \\
&= \sup_{\theta\in\Theta} |L_\theta - \mathbb{E}(L_\theta|X)| .
\end{aligned}$$

The term in the above display goes to 0 in probability. Indeed, for fixed θ, the function of θ goes to 0 in probability because of Lemma A.3. The convergence of the supremum over θ to 0 is then a consequence of the fact that Θ is compact and of Lemma A.1 and corollary A.2.

Finally, since $\hat{\theta}_{ML}$ minimizes L_θ and so also $L_\theta - \log(\det(R_0)) - 1$, we conclude with, for any $\theta \in \Theta$,

$$\begin{aligned}
D_{n,\hat{\theta}_{ML}} - D_{n,\theta} &\leq L_{\hat{\theta}_{ML}} - L_\theta + 2\sup_{\theta\in\Theta} |L_\theta - \log(\det(R_0)) - 1 - D_{n,\theta}| \\
&\leq 2\sup_{\theta\in\Theta} |L_\theta - \log(\det(R_0)) - 1 - D_{n,\theta}| .
\end{aligned}$$

Hence $\sup_{\theta\in\Theta}\left(D_{n,\hat{\theta}_{ML}} - D_{n,\theta}\right) = o_p(1)$.

□

A.3 Proofs for Cross Validation

Lemma A.4. *For $i = 1, ..., p$,*

$$\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i} CV_\theta\right|\right)$$

is bounded w.r.t n.

Proof of Lemma A.4. We have

$$\begin{aligned}
\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i} CV_\theta\right|\right) &\leq \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n}\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}(y_k - \hat{y}_{k,\theta})^2\right|\right) \\
\text{(symmetry of } X_1, ..., X_n\text{:)} &= \mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}(y_1 - \hat{y}_{1,\theta})^2\right|\right) \\
\text{(Lemma A.15:)} &\leq C_{sup}\sum_{i_1+...+i_p\leq p}\int_\Theta \mathbb{E}\left(\left|\frac{\partial^{i_1}}{\partial\theta_1^{i_1}}...\frac{\partial^{i_p}}{\partial\theta_p^{i_p}}\frac{\partial}{\partial\theta_i}(y_1 - \hat{y}_{1,\theta})^2\right|\right) d\theta.
\end{aligned}$$

Let us consider a specific $i_1, ..., i_p$. Then $[\partial^{i_1}/\partial\theta_1^{i_1}]...[\partial^{i_p}/\partial\theta_p^{i_p}][\partial/\partial\theta_i](y_1 - \hat{y}_{1,\theta})^2$ is a weighted sum (weights and number of terms depending only on $i_1, ..., i_p$), so that the terms are of the two following forms:

$$(y_1 - \hat{y}_{1,\theta})\left(\frac{\partial^{k_1}}{\partial\theta_1^{k_1}}...\frac{\partial^{k_p}}{\partial\theta_p^{k_p}}\frac{\partial}{\partial\theta_i}\hat{y}_{1,\theta}\right) \quad \text{or} \quad \left(\frac{\partial^{k_1}}{\partial\theta_1^{k_1}}...\frac{\partial^{k_p}}{\partial\theta_p^{k_p}}\frac{\partial}{\partial\theta_i}\hat{y}_{1,\theta}\right)\left(\frac{\partial^{l_1}}{\partial\theta_1^{l_1}}...\frac{\partial^{l_p}}{\partial\theta_p^{l_p}}\frac{\partial}{\partial\theta_i}\hat{y}_{1,\theta}\right).$$

Thus, we just have to show that the mean values of the absolute values of the terms of the form above (for $k_1 + ... + k_p \leq p$ and $l_1 + ... + l_p \leq p$) are bounded uniformly in $\theta \in \Theta$. By using Cauchy-Schwarz inequality, these means of absolute values are smaller than either

$$\sqrt{\mathbb{E}\left((y_1 - \hat{y}_{1,\theta})^2\right)}\sqrt{\mathbb{E}\left(\left(\frac{\partial^{k_1}}{\partial\theta_1^{k_1}}...\frac{\partial^{k_p}}{\partial\theta_p^{k_p}}\frac{\partial}{\partial\theta_i}\hat{y}_{1,\theta}\right)^2\right)}$$

or

$$\sqrt{\mathbb{E}\left(\left(\frac{\partial^{k_1}}{\partial\theta_1^{k_1}}...\frac{\partial^{k_p}}{\partial\theta_p^{k_p}}\frac{\partial}{\partial\theta_i}\hat{y}_{1,\theta}\right)^2\right)}\sqrt{\mathbb{E}\left(\left(\frac{\partial^{l_1}}{\partial\theta_1^{l_1}}...\frac{\partial^{l_p}}{\partial\theta_p^{l_p}}\frac{\partial}{\partial\theta_i}\hat{y}_{1,\theta}\right)^2\right)}.$$

Now, $\mathbb{E}\left((y_1-\hat{y}_{1,\theta})^2\right) \leq 2\mathbb{E}(y_1^2)+2\mathbb{E}(\hat{y}_{1,\theta}^2)$. The term $\mathbb{E}(y_1^2)$ is bounded uniformly in θ. Thus, finally, it remains to show that for any $a_1+...+a_p \leq p+1$, $\sup_{\theta\in\Theta}\mathbb{E}\left(\left([\partial^{a_1}/\partial\theta_1^{a_1}]...[\partial^{a_p}/\partial\theta_p^{a_p}]\hat{y}_{1,\theta}\right)^2\right)$ is bounded. For that, we have $\hat{y}_{1,\theta}=r_{1,\theta}^t R_{1,\theta}^{-1}y_{-1}$. Thus, $[\partial^{a_1}/\partial\theta_1^{a_1}]...[\partial^{a_p}/\partial\theta_p^{a_p}]\hat{y}_{1,\theta}$ is a fixed sum of weighted terms of the form $w_\theta^t M_\theta y_{-1}$, where w_θ is of the form $[\partial^{b_1}/\partial\theta_1^{b_1}]...[\partial^{b_p}/\partial\theta_p^{b_p}]r_{1,\theta}$ ($b_1+...+b_p \leq p+1$) and M_θ is of the form $R_{1,\theta}^{-1}M_{1,\theta}...R_{1,\theta}^{-1}M_{k,\theta}R_{1,\theta}^{-1}$. Finally, k is smaller than a finite constant C_{sup} (function of p) and $M_{i,\theta}$ is of the form $[\partial^{c_1}/\partial\theta_1^{c_1}]...[\partial^{c_p}/\partial\theta_p^{c_p}]R_{1,\theta}$, with $c_1+...+c_p \leq p+1$. Thus, it is sufficient to show that a generic $\sup_{\theta\in\Theta}\mathbb{E}\left(\left(w_\theta^t M_\theta y_{-1}\right)^2\right)$, as previously defined, is bounded.

Then,

$$\begin{aligned}
\sup_{\theta\in\Theta}\mathbb{E}\left(\left(w_\theta^t M_\theta y_{-1}\right)^2\right) &= \sup_{\theta\in\Theta}\mathbb{E}_X\mathbb{E}_{y|X}\left(y_{-1}^t M_\theta^t w_\theta w_\theta^t M_\theta y_{-1}\right)\\
&= \sup_{\theta\in\Theta}\mathbb{E}_X \operatorname{Tr}\left(R_{1,0}M_\theta^t w_\theta w_\theta^t M_\theta\right)\\
&\leq \sup_{\theta\in\Theta}\mathbb{E}_X\left[\sum_{i,j=2}^n\left|\left(M_\theta R_{1,0}M_\theta^t\right)_{i,j}\right|\left|\left(w_\theta w_\theta^t\right)_{i,j}\right|\right] \qquad (8)\\
&= \sup_{\theta\in\Theta}\left[\sum_{i,j=2}^n\mathbb{E}_{X_2,...,X_n}\left(\left|\left(M_\theta R_{1,0}M_\theta^t\right)_{i,j}\right|\mathbb{E}_{X_1|X_2,...,X_n}\left|\left(w_\theta w_\theta^t\right)_{i,j}\right|\right)\right].
\end{aligned}$$

Now, because of Conditions 2.3 and 3.2,

$$\begin{aligned}
\mathbb{E}_{X_1|X_2,...,X_n}\left|\left(w_\theta w_\theta^t\right)_{i,j}\right| &\leq \frac{C_{sup}}{n}\int_{[0,n^{1/d}]^d}\frac{1}{1+|X_i-x_1|^{d+1}}\frac{1}{1+|X_j-x_1|^{d+1}}dx_1\\
\text{(Lemma A.16:)} &\leq \frac{1}{n}\frac{C_{sup}}{1+|X_i-X_j|^{d+1}}.
\end{aligned}$$

So,

$$\begin{aligned}
\sup_{\theta\in\Theta}\mathbb{E}\left(\left(w_\theta^t M_\theta y_{-1}\right)^2\right) &\leq C_{sup}\frac{1}{n}\sup_{\theta\in\Theta}\left[\sum_{i,j=2}^n\mathbb{E}_{X_2,...,X_n}\left(\left|\left(M_\theta R_{1,0}M_\theta^t\right)_{i,j}\right|\frac{1}{1+|X_i-X_j|^{d+1}}\right)\right]\\
\text{(Cauchy-Schwarz:)} &\leq \sup_{\theta\in\Theta}\left[\sqrt{\mathbb{E}\left\{\left|M_\theta R_{1,0}M_\theta^t\right|^2\right\}}\sqrt{\mathbb{E}\left\{\frac{1}{n}\sum_{i,j=2}^n\left(\frac{1}{1+|X_i-X_j|^{d+1}}\right)^2\right\}}\right].
\end{aligned}$$

The supremum over θ of the second term above is bounded because of Lemma A.17. The supremum over θ of the first term above is bounded because of Lemmas A.17, A.18 and A.21.

□

Corollary A.5. *For any $i=1,...,p$,*

$$\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}\mathbb{E}\left(CV_\theta|\,X\right)\right|\right) \quad \textit{and} \quad \sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}\mathbb{E}\left(CV_\theta\right)\right|$$

are bounded w.r.t n.

Proof of corollary A.5. The corollary is a consequence of Lemma A.4, $\sup_\theta|\mathbb{E}(.)| \leq \mathbb{E}(\sup_\theta|.|)$ and of the fact that, for fixed n, we have $(\partial/\partial\theta_i)\mathbb{E}\left(CV_\theta|\,X\right)=\mathbb{E}\left((\partial/\partial\theta_i)CV_\theta|\,X\right)$ and $(\partial/\partial\theta_i)\mathbb{E}(CV_\theta)=\mathbb{E}((\partial/\partial\theta_i)CV_\theta)$. □

Lemma A.6. *For any fixed $\theta\in\Theta$ we have*

$$\mathbb{E}\left(|CV_\theta-\mathbb{E}(CV_\theta|X)|\right)\to_{n\to\infty}0.$$

Proof of Lemma A.6. From (3), we have $CV_\theta = y^t M_\theta y$, with $M_\theta = (1/n) R_\theta^{-1} diag(R_\theta^{-1})^{-2} R_\theta^{-1}$. Because of Lemma A.18, the eigenvalues of M_θ are bounded uniformly in n, X, θ by a finite constant C_{sup}. Thus, the proof of the lemma is exactly the same as that of Lemma A.3, with R_θ^{-1} replaced by M_θ.

□

Definition A.7. *Consider a fixed $\theta \in \Theta$. Consider two functions of n: $n_2(n) \in \mathbb{N}^*$ and $\Delta(n) \geq 0$, that we write n_2 and Δ for simplicity, so that, for any $n \in \mathbb{N}^*$, n_2 can be written $n_2 = N_2^d$, with $N_2 \in \mathbb{N}^*$, and so that $n = n_2 \Delta$. Let, for $i = 1, ..., N_2 - 1$, $c_i = [((i-1)/N_2) n^{1/d}, (i/N_2) n^{1/d})$. Let $c_{N_2} = [((N_2 - 1)/N_2) n^{1/d}, n^{1/d}]$. Let, for $x \in [0, n^{1/d}]$, $i(x)$ be the unique $i \in \{1, ..., N_2\}$ so that $x \in c_i$. Let, for $t = (t_1, ..., t_d)^t \in [0, n^{1/d}]^d$, $C(t) = \prod_{j=1}^d c_{i(t_j)}$. Define the non-stationary covariance function $\tilde{K}_\theta(t_1, t_2) = K_\theta(t_1, t_2) \mathbf{1}_{C(t_1) = C(t_2)}$. Define $\tilde{R}_\theta$, $\tilde{R}_{i,\theta}$, $\tilde{r}_{i,\theta}$, $\tilde{\hat{y}}_{i,\theta}$, $\tilde{CV}_\theta$ similarly to R_θ, $R_{i,\theta}$, $r_{i,\theta}$, $\hat{y}_{i,\theta}$, CV_θ but with K_θ replaced by $\tilde{K}_\theta$. Furthermore, let us write the n_2 aforementioned sets of the form $\prod_{j=1}^d c_{i_j}$, for $i_1, ..., i_d \in \{1, ..., N_2\}$, as the sets $C_1, ..., C_{n_2}$. [The specific one-to-one correspondence we use between $\{1, ..., N_2\}^d$ and $\{1, ..., n_2\}$ is of no interest. Note that this one-to-one correspondence depends on n. The sets $C_1, ..., C_{n_2}$ also depend of n, but we drop this dependence in the notation for simplicity.]*

Let N_i be the random number of observation points in C_i and let X^i be the random N_i-tuple obtained from X by keeping only the observation points that are in C_i and by preserving the order of the indices in X. Let y^i be the column vector of size N_i, composed by the components y_j of y for which X_j is in C_i (preserving the order of indexes). Let $\bar{R}_{i,\theta}$ and $\bar{R}_{i,0}$ be the covariance matrices, under $(K_\theta, \delta_\theta)$ and (K_0, δ_0), of y^i, given X.

Finally, for $1 \leq i, j \leq n_2$, let v_i and w_j be two $N_i \times 1$ and $N_j \times 1$ vectors and M^{ij} be a $N_i \times N_j$ matrix. Then we use the convention that, when $N_i = 0$, $|M^{ij}| = ||M^{ij}|| = 0$, $||v_i|| = |v_i| = 0$ and $v_i^t M^{ij} w_j = 0$. Furthermore, if $i = j$ and M^{ii} is invertible when $N_i \geq 1$, we use the convention that $v_i^t (M^{ii})^{-1} w_i = 0$ when $N_i = 0$. [These conventions enable to write equalities or inequalities involving matrices and vectors of size N_i, N_j or $N_i \times N_j$, that hold regardless of whether N_i or N_j are zero or not. As can be checked along the proofs involving Definition A.7, these relations boil down to trivial relations (e.g. $0 = 0$) when $N_i = 0$ or $N_j = 0$. This way of proceeding considerably simplifies the exposition in these proofs.]

Lemma A.8. *Consider a fixed $\theta \in \Theta$. In the context of Definition A.7, if $n_2 = o(n)$,*

$$\mathbb{E}\left(\left|CV_\theta - \tilde{CV}_\theta\right|\right) \to_{n \to \infty} 0.$$

Proof of Lemma A.8. Assume that $n_2 = o(n)$, or equivalently that $\Delta \to_{n \to \infty} \infty$. We have

$$\begin{aligned}
\mathbb{E}\left(\left|CV_\theta - \tilde{CV}_\theta\right|\right) &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left(\left|(y_i - \hat{y}_{i,\theta})^2 - (y_i - \tilde{\hat{y}}_{i,\theta})^2\right|\right) \\
\text{(symmetry:)} &= \mathbb{E}\left(\left|(y_1 - \hat{y}_{1,\theta})^2 - (y_1 - \tilde{\hat{y}}_{1,\theta})^2\right|\right) \\
&= \mathbb{E}\left(\left|(y_1 - r_{1,\theta}^t R_{1,\theta}^{-1} y_{-1})^2 - (y_1 - \tilde{r}_{1,\theta}^t \tilde{R}_{1,\theta}^{-1} y_{-1})^2\right|\right) \\
&= \mathbb{E}\left(\left|\tilde{r}_{1,\theta}^t \tilde{R}_{1,\theta}^{-1} y_{-1} - r_{1,\theta}^t R_{1,\theta}^{-1} y_{-1}\right| \left|2y_1 - r_{1,\theta}^t R_{1,\theta}^{-1} y_{-1} - \tilde{r}_{1,\theta}^t \tilde{R}_{1,\theta}^{-1} y_{-1}\right|\right) \\
\text{(Cauchy-Schwarz:)} &\leq \sqrt{\mathbb{E}\left(\left(\tilde{r}_{1,\theta}^t \tilde{R}_{1,\theta}^{-1} y_{-1} - r_{1,\theta}^t R_{1,\theta}^{-1} y_{-1}\right)^2\right)} \\
&\quad \sqrt{\mathbb{E}\left(\left(2y_1 - r_{1,\theta}^t R_{1,\theta}^{-1} y_{-1} - \tilde{r}_{1,\theta}^t \tilde{R}_{1,\theta}^{-1} y_{-1}\right)^2\right)}
\end{aligned}$$

Now, the second square root in the above display is bounded, because of $(a+b+c)^2 \leq 3\left(a^2 + b^2 + c^2\right)$ and of arguments similar to but simpler than those given in the proof of Lemma A.4. Thus it only remains to show that

$$\mathbb{E}\left(\left(\tilde{r}_{1,\theta}^t \tilde{R}_{1,\theta}^{-1} y_{-1} - r_{1,\theta}^t R_{1,\theta}^{-1} y_{-1}\right)^2\right) \to_{n \to \infty} 0.$$

For this,

$$\begin{aligned}\mathbb{E}\left(\left(\tilde{r}_{1,\theta}^t\tilde{R}_{1,\theta}^{-1}y_{-1}-r_{1,\theta}^tR_{1,\theta}^{-1}y_{-1}\right)^2\right) &\leq 2\mathbb{E}\left(\left(\tilde{r}_{1,\theta}^t(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})y_{-1}\right)^2\right) \\ &\quad +2\mathbb{E}\left(\left((\tilde{r}_{1,\theta}-r_{1,\theta})^tR_{1,\theta}^{-1}y_{-1}\right)^2\right).\end{aligned} \tag{9}$$

We show separately that both terms in the right-hand side of (9) converge to 0. For the first term,

$$\begin{aligned}\mathbb{E}\left(\left(\tilde{r}_{1,\theta}^t(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})y_{-1}\right)^2\right) &= \mathbb{E}\left(\mathrm{Tr}\left[R_{1,0}(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})\tilde{r}_{1,\theta}\tilde{r}_{1,\theta}^t(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})\right]\right) \\ &\leq \sum_{i,j=2}^n\mathbb{E}\left(|(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})R_{1,0}(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})|_{i,j}|\tilde{r}_{1,\theta}\tilde{r}_{1,\theta}^t|_{i,j}\right).\end{aligned}$$

Hence, by the same arguments as after (8) in the proof of Lemma A.4, we obtain

$$\begin{aligned}\left[\mathbb{E}\left(\left(\tilde{r}_{1,\theta}^t(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})y_{-1}\right)^2\right)\right]^2 &\leq C_{sup}\mathbb{E}\left(|(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})R_{1,0}(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})|^2\right) \\ &\leq C_{sup}\mathbb{E}\left(\left\{||\tilde{R}_{1,\theta}^{-1}||+||R_{1,\theta}^{-1}||\right\}|R_{1,0}(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})|^2\right) \\ \text{(Lemmas A.18 and A.19:)} \quad &\leq \frac{C_{sup}}{n}\mathbb{E}\left(\mathrm{Tr}\left[(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})R_{1,0}^2(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})\right]\right) \\ \text{(Cauchy-Schwarz:)} \quad &\leq C_{sup}\sqrt{\mathbb{E}\left(|(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})^2|^2\right)}\sqrt{\mathbb{E}\left(|R_{1,0}^2|^2\right)}.\end{aligned}$$

From Lemma A.17, $\mathbb{E}\left(|R_{1,0}^2|^2\right)$ is bounded, so it remains to show that $\mathbb{E}\left(|(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})^2|^2\right)$ converges to 0. For this,

$$\begin{aligned}\mathbb{E}\left(|(\tilde{R}_{1,\theta}^{-1}-R_{1,\theta}^{-1})^2|^2\right) &= \mathbb{E}\left(|(\tilde{R}_{1,\theta}^{-1}(R_{1,\theta}-\tilde{R}_{1,\theta})R_{1,\theta}^{-1})^2|^2\right) \\ \text{(Lemma A.18:)} \quad &\leq C_{sup}\mathbb{E}\left(|(R_{1,\theta}-\tilde{R}_{1,\theta})R_{1,\theta}^{-1}\tilde{R}_{1,\theta}^{-1}(R_{1,\theta}-\tilde{R}_{1,\theta})|^2\right) \\ &= C_{sup}\frac{1}{n}\mathbb{E}\left(\mathrm{Tr}\left[(R_{1,\theta}-\tilde{R}_{1,\theta})^2R_{1,\theta}^{-1}\tilde{R}_{1,\theta}^{-1}(R_{1,\theta}-\tilde{R}_{1,\theta})^2\tilde{R}_{1,\theta}^{-1}R_{1,\theta}^{-1}\right]\right) \\ \text{(Cauchy-Schwarz:)} \quad &\leq C_{sup}\sqrt{\mathbb{E}\left(|(R_{1,\theta}-\tilde{R}_{1,\theta})^2R_{1,\theta}^{-1}\tilde{R}_{1,\theta}^{-1}|^2\right)}\sqrt{\mathbb{E}\left(|(R_{1,\theta}-\tilde{R}_{1,\theta})^2\tilde{R}_{1,\theta}^{-1}R_{1,\theta}^{-1}|^2\right)}.\end{aligned}$$

Hence, with Lemmas A.18, A.19 and A.22, we conclude that the first term of the right hand side of (9) goes to 0. Let us now show that the second term of the right hand side of (9) goes to 0. We have,

$$\begin{aligned}&\mathbb{E}\left(\left((\tilde{r}_{1,\theta}-r_{1,\theta})^tR_{1,\theta}^{-1}y_{-1}\right)^2\right) \\ &= \mathbb{E}\left(\mathrm{Tr}\left(R_{1,\theta}^{-1}R_{1,0}R_{1,\theta}^{-1}(\tilde{r}_{1,\theta}-r_{1,\theta})(\tilde{r}_{1,\theta}-r_{1,\theta})^t\right)\right) \\ &\leq \sum_{i,j=2}^n\mathbb{E}\left(\left|[R_{1,\theta}^{-1}R_{1,0}R_{1,\theta}^{-1}]_{i,j}\right|\frac{1}{n}\int_{[0,n^{1/d}]^d}\frac{1}{1+|X_i-x_1|^{d+1}}\frac{1}{1+|X_j-x_1|^{d+1}}\mathbf{1}_{C(X_i)\neq C(x_1)}\mathbf{1}_{C(X_j)\neq C(x_1)}dx_1\right),\end{aligned}$$

where the last line is obtained similarly to after (8) in the proof of Lemma A.4. Thus we have, with the notation

and result of Lemma A.23,

$$\begin{aligned}\mathbb{E}\left(\left((\tilde{r}_{1,\theta}-r_{1,\theta})^t R_{1,\theta}^{-1}y_{-1}\right)^2\right) &\leq C_{sup}\frac{1}{n}\sum_{i,j=2}^{n}\mathbb{E}\left(\left|[R_{1,\theta}^{-1}R_{1,0}R_{1,\theta}^{-1}]_{i,j}\right|\frac{1}{1+|X_i-X_j|^{d+1}}f(D_\Delta(X_i,X_j))\right)\\ \text{(Cauchy-Schwarz:)} &\leq \sqrt{\mathbb{E}\left(|R_{1,\theta}^{-1}R_{1,0}R_{1,\theta}^{-1}|^2\right)}\\ &\quad\sqrt{\frac{1}{n}\sum_{i,j=2}^{n}\mathbb{E}\left[\left(\frac{1}{1+|X_i-X_j|^{d+1}}\right)^2 f^2(D_\Delta(X_i,X_j))\right]}.\end{aligned}$$

From Lemmas A.17, A.18 and A.21, the first $\sqrt{.}$ in the above display is bounded. Thus it remains to show that the second $\sqrt{.}$ goes to 0. For this, noting that $f^2(t)\leq C_{sup}f(t)$ and distinguishing the case $i=j$ from the case $i\neq j$,

$$\begin{aligned}&\frac{1}{n}\sum_{i,j=2}^{n}\mathbb{E}\left[\left(\frac{1}{1+|X_i-X_j|^{d+1}}\right)^2 f(D_\Delta(X_1,X_j))\right]\\ &\leq \frac{C_{sup}}{n}\int_{[0,n^{1/d}]^d}f(D_\Delta(x))dx+\frac{C_{sup}}{n}\int_{[0,n^{1/d}]^d}dx_1\int_{[0,n^{1/d}]^d}dx_2\frac{1}{1+|x_1-x_2|^{d+1}}f(D_\Delta(x_1,x_2))\\ &=\frac{C_{sup}}{n}\int_{[0,n^{1/d}]^d}f(D_\Delta(x))dx+o(1)\quad\text{(Lemma A.24).}\end{aligned}\tag{10}$$

Now, for any $\epsilon>0$, there is a finite T so that $f(T)\leq\epsilon$, and by defining $E_n=\{x\in[0,n^{1/d}]^d; D_\Delta(x)\leq T\}$, we have $|E_n|=o(n)$, as can be seen easily, and

$$\frac{1}{n}\int_{[0,n^{1/d}]^d}f(D_\Delta(x))dx\leq f(0)\frac{|E_n|}{n}+\epsilon.$$

This finally shows that the second term of the right hand side of (9) goes to 0 which finishes the proof.

□

Lemma A.9. *For any fixed $\theta\in\Theta$,*

$$\mathbb{E}\left(\left|\mathbb{E}\left(CV_\theta\right)-\mathbb{E}\left[CV_\theta|X\right]\right|\right)$$

goes to 0 as $n\to\infty$.

Proof of Lemma A.9. Fix $\theta\in\Theta$. Because of Lemma A.8 and of $|\mathbb{E}(.)|\leq\mathbb{E}(|.|)$, it is sufficient to show that there exists a sequence $\Delta\to+\infty$ so that the lemma holds with CV_θ replaced by $\tilde{CV}_\theta$. Then, because of $\left(\mathbb{E}(.)\right)^2\leq\mathbb{E}\left((.)^2\right)$, it is sufficient to show $var\left(\mathbb{E}\left[\tilde{CV}_\theta\middle|X\right]\right)\to_{n\to\infty}0$.

Let $C_1,...,C_{n_2}$ be as in Definition A.7. Define, for $k=1,...,n_2$,

$$f_k(X)=\frac{1}{\Delta}\sum_{X_i\in C_k}\mathbb{E}\left(\left[y_i-\tilde{\hat{y}}_{i,\theta}\right]^2\middle|X\right).$$

[Note that, following the discussion in Definition A.7, we have $f_k(X)=0$ if $N_k=0$ and $f_k(X)=K_0(0)+\delta_0$ if $N_k=1$.] Then $\mathbb{E}\left(\tilde{CV}_\theta|X\right)=(1/n_2)\sum_{k=1}^{n_2}f_k(X)$. Let $\bar{R}_{k,\theta}$ and $\bar{R}_{k,0}$ be as in Definition A.7. Because of the definition of $\tilde{K}$ and by (3), we have

$$f_k(X)=\frac{1}{\Delta}\,\mathrm{Tr}\left(\bar{R}_{k,0}\bar{R}_{k,\theta}^{-1}diag(\bar{R}_{k,\theta}^{-1})^{-2}\bar{R}_{k,\theta}^{-1}\right).$$

The functions $f_k(X)$ satisfy the conditions of Lemma A.25. Furthermore, by using the notation N_k of Lemma A.25, we have

$$\begin{aligned}
\mathbb{E}\left(f_k^2(X)\middle|\, N_k = N\right) &= \frac{1}{\Delta^2}\mathbb{E}\left(\left[\mathrm{Tr}\left(\bar{R}_{k,0}\bar{R}_{k,\theta}^{-1}diag(\bar{R}_{k,\theta}^{-1})^{-2}\bar{R}_{k,\theta}^{-1}\right)\right]^2\middle| N_k = N\right)\\
\text{(Cauchy-Schwarz:)} &\leq \frac{1}{\Delta^2}\mathbb{E}\left(N^2|\bar{R}_{k,0}|^2|\bar{R}_{k,\theta}^{-1}diag(\bar{R}_{k,\theta}^{-1})^{-2}\bar{R}_{k,\theta}^{-1}|^2\middle| N_k = N\right)\\
\text{(Lemma A.20:)} &\leq C_{sup}\frac{N^2}{\Delta^2}\mathbb{E}\left(|\bar{R}_{k,0}|^2\middle| N_k = N\right)\\
\text{(Condition 3.1 and Lemma A.25:)} &\leq C_{sup}\frac{N^2}{\Delta^2}\left(1+\frac{N}{\Delta^2}\int_{[0,\Delta^{1/d}]^d}\int_{[0,\Delta^{1/d}]^d}\frac{1}{1+|x_1-x_2|^{d+1}}dx_1dx_2\right)\\
&\leq C_{sup}\left(\frac{N^2}{\Delta^2}+\frac{N^3}{\Delta^3}\right)\\
&\leq C_{sup}\left(1+\frac{N^4}{\Delta^4}\right).
\end{aligned}$$

Thus, because of Lemma A.26, there exists a sequence $\Delta \to_{n\to\infty} \infty$ so that $var\left(\mathbb{E}\left[\tilde{CV}_\theta\middle| X\right]\right) \to_{n\to\infty} 0$, which completes the proof. □

Lemma A.10. *Let, with the notation of Definition A.7, $\tilde{E}_{n,\theta}$ be defined as $E_{n,\theta}$, with K_θ replaced by $\tilde{K}_\theta$. Fix $\theta \in \Theta$. Then, if $n_2 = o(n)$,*

$$\mathbb{E}\left(\left|E_{n,\theta} - \tilde{E}_{n,\theta}\right|\right) \to_{n\to\infty} 0.$$

Proof of Lemma A.10. We have, by letting $\tilde{\hat{y}}_\theta(t)$ be as $\hat{y}_\theta(t)$, with K_θ replaced by $\tilde{K}_\theta$.

$$\begin{aligned}
\mathbb{E}\left(\left|E_{n,\theta} - \tilde{E}_{n,\theta}\right|\right) &= \mathbb{E}\left(\left|\frac{1}{n}\int_{[0,n^{1/d}]^d}\left[Y(t)-\hat{y}_\theta(t)\right]^2 dt - \frac{1}{n}\int_{[0,n^{1/d}]^d}\left[Y(t)-\tilde{\hat{y}}_\theta(t)\right]^2 dt\right|\right)\\
&\leq \mathbb{E}\left(\frac{1}{n}\int_{[0,n^{1/d}]^d}\left|\left[Y(t)-\hat{y}_\theta(t)\right]^2 - \left[Y(t)-\tilde{\hat{y}}_\theta(t)\right]^2\right| dt\right)
\end{aligned}$$

The variable t in the integral above is formally equivalent to a new observation point X_{n+1}, so that $X_1, ..., X_{n+1}$ are independent and uniformly distributed on $[0, n^{1/d}]^d$. Thus,

$$\mathbb{E}\left(\left|E_{n,\theta} - \tilde{E}_{n,\theta}\right|\right) \leq \mathbb{E}\left(\left|\left(Y(X_{n+1})-\hat{y}_\theta(X_{n+1})\right)^2 - \left(Y(X_{n+1})-\tilde{\hat{y}}_\theta(X_{n+1})\right)^2\right|\right).$$

The rest of the proof is carried out as in Lemma A.8, the only difference being that there are $n+1$ observation points instead of n.

□

Lemma A.11. *For any fixed $\theta \in \Theta$ we have*

$$\mathbb{E}\left(|E_{n,\theta} - \mathbb{E}(E_{n,\theta}|X)|\right) \to_{n\to\infty} 0.$$

Proof of Lemma A.11. Because of Lemma A.10 and using $|\mathbb{E}(.)| \leq \mathbb{E}(|.|)$ and $\mathbb{E}^2(.) \leq \mathbb{E}((.)^2)$, it is sufficient to show that there exists a sequence $\Delta \to_{n\to\infty}$ so that

$$\mathbb{E}\left(var\left(\tilde{E}_{n,\theta}\middle| X\right)\right)$$

goes to 0 as $n \to \infty$.

Let us use the notation $C_1, ..., C_{n_2}$ of Definition A.7. Let, for $t \in \mathbb{R}^d$ and $v = (v_1, ..., v_m) \in (\mathbb{R}^d)^m$, $r_\theta(t, v) = (K_\theta(t, v_1), ..., K_\theta(t, v_m))^t$. We define $r_0(t, v)$ similarly. Let y^i, $\bar{R}_{i,\theta}$ and $\bar{R}_{i,0}$ be as in Definition A.7. Let for $i \neq j$, $R_0(X^i, X^j) = \left[K_0((X^i)_k, (X^j)_l)\right]_{k=1,...,N_i; l=1,...,N_j}$. Let $R_0(X^i, X^i) = \left[K_0((X^i)_k, (X^i)_l)\right]_{k,l=1,...,N_i} + \delta_0 I_{N_i}$. Then,

$$\tilde{E}_{n,\theta} = \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{1}{\Delta} \int_{C_i} dt_i \left[Y(t_i) - r_\theta^t(t_i, X^i) \bar{R}_{i,\theta}^{-1} y^i \right]^2$$

Hence, using the relation $cov\left(A^2, B^2\right) = 2\left(cov(A, B)\right)^2$, for two centered Gaussian variables A and B, we obtain

$$\begin{aligned}
&var\left(\tilde{E}_{n,\theta} \middle| X \right) \\
&= \frac{2}{n_2} \sum_{i=1}^{n_2} \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{1}{\Delta^2} \int_{C_i} dt_i \int_{C_j} dt_j cov^2 \left(\left[Y(t_i) - r_\theta^t(t_i, X^i) \bar{R}_{i,\theta}^{-1} y^i \right], \left[Y(t_j) - r_\theta^t(t_j, X^j) \bar{R}_{j,\theta}^{-1} y^j \right] \middle| X \right) \\
&= \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{1}{\Delta^2} \int_{C_i} dt_i \int_{C_j} dt_j \Big\{ K_0(t_i, t_j) - r_\theta^t(t_i, X^i) \bar{R}_{i,\theta}^{-1} r_0(t_j, X^i) - r_\theta^t(t_j, X^j) \bar{R}_{j,\theta}^{-1} r_0(t_i, X^j) \\
&\quad + r_\theta^t(t_i, X^i) \bar{R}_{i,\theta}^{-1} R_0(X^i, X^j) \bar{R}_{j,\theta}^{-1} r_\theta(t_j, X^j) \Big\}^2 .
\end{aligned}$$

Now, we use $(a_1 + a_2 + a_3 + a_4)^2 \leq 4(a_1^2 + a_2^2 + a_3^2 + a_4^2)$. Hence we obtain

$$\mathbb{E}\left(var\left(\tilde{E}_{n,\theta} \middle| X \right) \right) \leq C_{sup} \left(T_1 + T_2 + T_3 + T_4 \right), \tag{11}$$

where T_1, T_2, T_3, T_4 are defined and treated below, and with $T_2 = T_3$ by symmetry.

For T_1,

$$\begin{aligned}
T_1 &= \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{1}{\Delta^2} \int_{C_i} dt_i \int_{C_j} dt_j K_0^2(t_i, t_j) \\
\text{(Condition 3.1:)} &\leq \frac{C_{sup}}{n_2} \sum_{i=1}^{n_2} \frac{1}{n_2} \frac{1}{\Delta^2} \int_{C_i} dt_i \int_{\mathbb{R}^d} dt \left(\frac{1}{1 + |t_i - t|^{d+1}} \right)^2 \\
&\leq C_{sup} \frac{1}{n_2 \Delta}.
\end{aligned} \tag{12}$$

For T_2, using Cauchy-Schwarz and Lemma A.20,

$$\begin{aligned}
&T_2 \\
&= \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{1}{\Delta^2} \int_{C_i} dt_i \int_{C_j} dt_j \mathbb{E}\left[\left(r_\theta^t(t_i, X^i) \bar{R}_{i,\theta}^{-1} r_0(t_j, X^i) \right)^2 \right] \\
&\leq C_{sup} \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{1}{\Delta^2} \int_{C_i} dt_i \int_{C_j} dt_j \mathbb{E}\left[||r_\theta(t_i, X^i)||^2 ||r_0(t_j, X^i)||^2 \right].
\end{aligned}$$

Now, using the notation N_i of Lemma A.25 and Conditions 3.1 and 3.2,

$$
\begin{aligned}
&T_2 \\
&\leq C_{sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i}dt_i\int_{C_j}dt_j\mathbb{E}\left[N_i^2\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\}^4\right] \\
&\leq C_{sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\Delta^2\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\} \quad \text{(Lemma A.27)} \\
&\leq C_{sup}\frac{\Delta^2}{n_2}\max_{i=1,\ldots,n_2}\sum_{j=1}^{n_2}\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\} \\
&\leq C_{sup}\frac{\Delta^2}{n_2} \quad \text{(Lemma A.28, and because we will set } \Delta\to_{n\to\infty}\infty\text{).}
\end{aligned} \tag{13}
$$

For T_4 in (11), using Cauchy-Schwarz and Lemma A.20,

$$
\begin{aligned}
&T_4 \\
&= \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i}dt_i\int_{C_j}dt_j\mathbb{E}\left[\left(r_\theta^t(t_i,X^i)\bar{R}_{i,\theta}^{-1}R_0(X^i,X^j)\bar{R}_{j,\theta}^{-1}r_\theta(t_j,X^j)\right)^2\right] \\
&\leq C_{sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i}dt_i\int_{C_j}dt_j\sqrt{\mathbb{E}\left[||r_\theta^t(t_i,X^i)||^4\right]}\sqrt{\mathbb{E}\left[||R_0(X^i,X^j)\bar{R}_{j,\theta}^{-1}r_\theta(t_j,X^j)||^4\right]}.
\end{aligned} \tag{14}
$$

Using Condition 3.1, Lemma A.20 and Lemma A.29, we obtain

$$
\begin{aligned}
||R_0(X^i,X^j)\bar{R}_{j,\theta}^{-1}r_\theta(t_j,X^j)||^2 &\leq C_{sup}N_iN_j\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\}^2||\bar{R}_{j,\theta}^{-1}r_\theta(t_j,X^j)||^2 \\
&\leq C_{sup}N_iN_j^2\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\}^2.
\end{aligned}
$$

Hence, going back to (14),

$$
\begin{aligned}
&T_4 \\
&\leq C_{sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i}dt_i\int_{C_j}dt_j\sqrt{\mathbb{E}\left[N_i^2\right]}\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\}^2\sqrt{\mathbb{E}\left[N_i^2N_j^4\right]} \\
&\leq C_{sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i}dt_i\int_{C_j}dt_j\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\}^2\sqrt{\mathbb{E}\left[N_i^2\right]}\sqrt{\sqrt{\mathbb{E}\left[N_i^4\right]}\sqrt{\mathbb{E}\left[N_j^8\right]}} \\
&\leq C_{sup}\frac{\Delta^4}{n_2} \quad \text{(Lemmas A.27 and A.28).}
\end{aligned} \tag{15}
$$

Hence, from (12), (13) and (15), we can set $\Delta = n^{1/6}$ to complete the proof.

□

Lemma A.12. *For any fixed $\theta\in\Theta$,*

$$\mathbb{E}\left(\left|\mathbb{E}\left(E_{n,\theta}\right)-\mathbb{E}\left[E_{n,\theta}|X\right]\right|\right)$$

goes to 0 as $n\to\infty$.

Proof of Lemma A.12. Fix $\theta \in \Theta$. Because of Lemma A.10 and of $|\mathbb{E}(.)| \leq \mathbb{E}(|.|)$, it is sufficient to show that there exists a sequence $\Delta \to +\infty$ so that the lemma holds with $E_{n,\theta}$ replaced by $\tilde{E}_{n,\theta}$. Then, because of $(\mathbb{E}(.))^2 \leq \mathbb{E}\left((.)^2\right)$, it is sufficient to show $var\left(\mathbb{E}\left[\tilde{E}_{n,\theta}\middle| X\right]\right) \to_{n\to\infty} 0$.

Let $C_1, ..., C_{n_2}$ be as in Definition A.7 and let $\tilde{\hat{y}}_\theta(t)$ be as in the proof of Lemma A.10. Define, for $k = 1, ..., n_2$,

$$g_k(X) = \frac{1}{\Delta}\int_{C_k} dt_k \mathbb{E}\left(\left[Y(t_k) - \tilde{\hat{y}}_\theta(t_k)\right]^2 \middle| X\right).$$

[Note that, following the discussion in Definition A.7, we have $g_k(x) = K_0(0)$ if $N_k = 0$.] Then $\mathbb{E}\left(\tilde{E}_{n,\theta}|X\right) = (1/n_2)\sum_{k=1}^{n_2} g_k(X)$. Following the notation of Lemma A.11 we have,

$$\begin{aligned}
&g_k(X) = \\
&\frac{1}{\Delta}\int_{C_k} dt_k \mathbb{E}\left(\left[Y(t_k) - r_\theta^t(t_k, X^k)\bar{R}_{k,\theta}^{-1} y^k\right]^2 \middle| X\right) \\
&\leq 2\frac{1}{\Delta}\int_{C_k} dt_k \left(K_0(0) + r_\theta^t(t_k, X^k)\bar{R}_{k,\theta}^{-1}\bar{R}_{k,0}\bar{R}_{k,\theta}^{-1} r_\theta(t_k, X^k)\right) \\
&\leq C_{sup} + 2\frac{1}{\Delta}\int_{C_k} dt_k ||r_\theta^t(t_k, X^k)|| ||\bar{R}_{k,0}\bar{R}_{k,\theta}^{-1} r_\theta(t_k, X^k)|| \quad \text{(Lemma A.20)} \\
&\leq C_{sup} + C_{sup}\frac{1}{\Delta}\int_{C_k} dt_k \sqrt{N_k} N_k ||\bar{R}_{k,\theta}^{-1} r_\theta(t_k, X^k)|| \quad \text{(Conditions 3.1 and 3.2 and Lemma A.29)} \\
&\leq C_{sup}\left(1 + N_k^2\right) \quad \text{(Lemma A.20).}
\end{aligned}$$

Hence $\mathbb{E}\left(g_k^2(X)\middle| N_k = N\right) \leq C_{sup}(1 + N^4)$, so that we can complete the proof with Lemma A.26.

□

Lemma A.13. *Consider a fixed $\theta \in \Theta$. Then*

$$\mathbb{E}(CV_\theta) - \mathbb{E}(E_{n,\theta}) - \delta_0$$

goes to 0 as $n \to \infty$.

Proof of Lemma A.13. Let us consider a random observation point X_{n+1} with uniform distribution on $[0, n^{1/d}]^d$. Let us also consider a Gaussian variable ϵ_{n+1} with mean 0 and variance δ_0. Consider that X_{n+1} and ϵ_{n+1} are independent and independent of X, Y and ϵ. With the same argument as in the proof of Lemma A.10, we have

$$\mathbb{E}\left(E_{n,\theta}\right) = \mathbb{E}\left(\left[Y(X_{n+1}) - \hat{y}_\theta(X_{n+1})\right]^2\right),$$

where we remind that $\hat{y}_\theta(X_{n+1}) = r_\theta^t(X_{n+1})R_\theta^{-1} y$, with $r_\theta(X_{n+1}) = (K(X_1, X_{n+1}), ..., K(X_n, X_{n+1}))^t$. Now, by symmetry of the roles of $X_1, ..., X_{n+1}$ and $\epsilon_1, ...\epsilon_{n+1}$, we have

$$\mathbb{E}\left(CV_\theta\right) = \mathbb{E}\left(\left[Y(X_{n+1}) - \hat{y}_{n-1,\theta}(X_{n+1})\right]^2\right) + \delta_0,$$

with $\hat{y}_{n-1,\theta}(X_{n+1}) = r_{n-1,\theta}^t \tilde{R}_{n-1,\theta}^{-1} y$, with $r_{n-1,\theta} = (K(X_1, X_{n+1}), ..., K(X_{n-1}, X_{n+1}), 0)^t$ and

$$\tilde{R}_{n-1,\theta} = \begin{pmatrix} (K_\theta(X_i, X_j))_{i,j=1,...,(n-1)} + \delta_\theta I_{n-1} & 0 \\ 0 & 1 \end{pmatrix}.$$

Hence, using Cauchy-Schwarz

$$\begin{aligned}
&|\mathbb{E}(CV_\theta) - \mathbb{E}(E_{n,\theta}) - \delta_0| \\
&\leq \sqrt{\mathbb{E}\left(\left[r_\theta^t(X_{n+1})R_\theta^{-1} y - r_{n-1,\theta}^t \tilde{R}_{n-1,\theta}^{-1} y\right]^2\right)}\sqrt{\mathbb{E}\left(\left[r_\theta^t(X_{n+1})R_\theta^{-1} y + r_{n-1,\theta}^t \tilde{R}_{n-1,\theta}^{-1} y - 2Y(X_{n+1})\right]^2\right)}. \qquad (16)
\end{aligned}$$

The second term in (16) is shown to be bounded with techniques similar to but simpler than in the proof of Lemma A.4. The first term in (16) is shown to go to zero with techniques similar to but simpler than in the proof of Lemma A.8.

□

Corollary A.14. *For any $i = 1, ..., p$,*

$$\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}E_{n,\theta}\right|\right), \quad \mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}\mathbb{E}\left(E_{n,\theta}|\,X\right)\right|\right) \quad \textit{and} \quad \sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}\mathbb{E}\left(E_{n,\theta}\right)\right|$$

are bounded w.r.t n.

Proof of corollary A.14. We have

$$\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}E_{n,\theta}\right|\right) = \mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}\frac{1}{n}\int_{[0,n^{1/d}]^d}\left[Y(t)-\hat{y}_\theta(t)\right]^2 dt\right|\right).$$

For fixed n we can exchange derivative and integration, so we obtain

$$\begin{aligned}\mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}E_{n,\theta}\right|\right) &= \mathbb{E}\left(\sup_{\theta\in\Theta}\left|\frac{1}{n}\int_{[0,n^{1/d}]^d}\frac{\partial}{\partial\theta_i}\left[Y(t)-\hat{y}_\theta(t)\right]^2 dt\right|\right)\\ &\leq \mathbb{E}\left(\frac{1}{n}\int_{[0,n^{1/d}]^d}\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}\left[Y(t)-\hat{y}_\theta(t)\right]^2\right| dt\right).\end{aligned}$$

Hence, by considering t as a new random observation point X_{n+1} as in the proof of Lemma A.10, we show the first bound of the lemma as in the proof of Lemma A.4, the only difference being that there are $n+1$ observation points instead of n. The second and third bounds are proved as in the proof of corollary A.5.

□

Proof of Theorem 3.4. We have

$$\begin{aligned}&\sup_{\theta\in\Theta}|CV_\theta-\delta_0-E_{n,\theta}|\\ &\leq \sup_{\theta\in\Theta}|CV_\theta-\mathbb{E}(CV_\theta|X)|+\sup_{\theta\in\Theta}|E(CV_\theta|X)-\mathbb{E}(CV_\theta)|+\sup_{\theta\in\Theta}|\mathbb{E}(CV_\theta)-\delta_0-\mathbb{E}(E_{n,\theta})|\\ &\quad+\sup_{\theta\in\Theta}|\mathbb{E}(E_{n,\theta})-\mathbb{E}(E_{n,\theta}|X)|+\sup_{\theta\in\Theta}|\mathbb{E}(E_{n,\theta}|X)-E_{n,\theta}|.\end{aligned}$$

The five terms in the right-hand size of the above equation go to 0 in probability. Indeed, for fixed θ, the functions of θ go to 0 in probability because of Lemmas A.6, A.9, A.11, A.12 and A.13. The convergence of the supremums over θ to 0 is then a consequence of the fact that Θ is compact and of Lemma A.4 and corollaries A.5 and A.14. Finally, since $\hat{\theta}_{CV}$ minimizes $CV_\theta+\delta_0$, we conclude with, for any $\theta\in\Theta$

$$\begin{aligned}E_{n,\hat{\theta}_{CV}}-E_{n,\theta} &\leq CV_{\hat{\theta}_{CV}}-CV_\theta+2\sup_{\theta\in\Theta}|CV_\theta-\delta_0-E_{n,\theta}|\\ &\leq 2\sup_{\theta\in\Theta}|CV_\theta-\delta_0-E_{n,\theta}|.\end{aligned}$$

Hence

$$\sup_{\theta\in\Theta}\left(E_{n,\hat{\theta}_{CV}}-E_{n,\theta}\right)=o_p(1).$$

□

A.4 Technical results

The following technical results are proved in the supplementary material.

Lemma A.15. *Consider a fixed number n of observation points. Consider a function $f_\theta(X,y)$ that is p times continuously differentiable w.r.t θ for any X,y and so that, for $i_1+...+i_p \leq p$,*

$$\sup_\theta \left|(\partial^{i_1}/\partial\theta_1^{i_1})...(\partial^{i_p}/\partial\theta_p^{i_p})f_\theta(X,y)\right|$$

has finite mean value w.r.t X and y. Then, there exists a constant C_{sup} (depending only of Θ) so that

$$\mathbb{E}\left(\sup_{\theta\in\Theta}|f_\theta(X,y)|\right) \leq C_{sup}\sum_{i_1+...+i_p\leq p}\int_\Theta \mathbb{E}\left(\left|\frac{\partial^{i_1}}{\partial\theta_1^{i_1}}...\frac{\partial^{i_p}}{\partial\theta_p^{i_p}}f_\theta(X,y)\right|\right)d\theta.$$

Lemma A.16. *There exists a finite constant C_{sup} so that, for any $a,b\in\mathbb{R}^d$,*

$$\int_{\mathbb{R}^d}\frac{1}{1+|a-c|^{d+1}}\frac{1}{1+|b-c|^{d+1}}dc \leq C_{sup}\frac{1}{1+|a-b|^{d+1}}.$$

Lemma A.17. *Let $0<C_{inf}\leq C_{sup}<\infty$ be fixed independently of n. Let s_n be a function of n so that $s_n\in\mathbb{N}^*$ and $C_{inf}n\leq s_n\leq C_{sup}n$. Consider s_n observation points $\bar{X}_1,...,\bar{X}_{s_n}$, independent and uniformly distributed on $[0,n^{1/d}]^d$. Let $A_1,...,A_k$ be k sequences of $s_n\times s_n$ random matrices so that, for $l=1,...,k$, $(A_l)_{i,j}$ depends only on $\bar{X}_i$ and $\bar{X}_j$ and satisfies $|(A_l)_{i,j}|\leq 1/(1+|\bar{X}_i-\bar{X}_j|^{d+1})$. Then $\mathbb{E}_X\left(|A_1...A_k|^2\right)$ is bounded w.r.t. n.*

Lemma A.18. *The supremum over n, θ and X of the eigenvalues of R_θ^{-1}, $R_{1,\theta}^{-1}$, $diag(R_\theta^{-1})$, $diag(R_{1,\theta}^{-1})$, $diag(R_\theta^{-1})^{-1}$ and $diag(R_{1,\theta}^{-1})^{-1}$ is smaller than a constant $C_{sup}<+\infty$.*

Lemma A.19. *Lemma A.18 also holds when K_θ is replaced by $\tilde{K}_\theta$ of Definition A.7.*

Lemma A.20. *Lemma A.18 also holds when R_θ is replaced by $\bar{R}_{k,\theta}$ of Definition A.7.*

Lemma A.21. *Let $k\in\mathbb{N}$. Let $A_{1,\theta},...,A_{k,\theta}$ be k sequences of symmetric random matrices (functions of X and θ) so that, for any $m\in\mathbb{N}$, $a_1,...,a_m\in\{1,...,k\}$, $\sup_{\theta\in\Theta}\mathbb{E}_X\left|A_{a_1,\theta}...A_{a_m,\theta}\right|^2$ is bounded (w.r.t n). Let $B_{1,\theta},...,B_{k+1,\theta}$ be $k+1$ sequences of random symmetric non-negative matrices (functions of X and θ) so that $\sup_\theta||B_{1,\theta}||,...,\sup_\theta||B_{k+1,\theta}||$ are bounded (w.r.t n and X). Then*

$$\sup_{\theta\in\Theta}\mathbb{E}_X\left|B_{1,\theta}A_{1,\theta}B_{2,\theta}...B_{k,\theta}A_{k,\theta}B_{k+1,\theta}\right|^2$$

is bounded w.r.t n.

Lemma A.22. *Consider a fixed $\theta\in\Theta$. With the notation of Definition A.7, we have, when $n_2=o(n)$,*

$$\mathbb{E}\left(\left|(R_{1,\theta}-\tilde{R}_{1,\theta})^2\right|^2\right)\to_{n\to\infty}0.$$

Lemma A.23. *Let $C(t)$ be as in Definition A.7. Define, for $T\geq 0$, $f(T)=\int_{\mathbb{R}^d\setminus[-T,T]^d}1/(1+|t|^{d+1})dt$. Define, for $x\in[0,n^{1/d}]^d$, $D_\Delta(x)=\inf_{t\in\mathbb{R}^d\setminus C(x)}|x-t|$. Define $D_\Delta(x_1,...,x_m)=\min_{i=1,...,m}D_\Delta(x_i)$. Then, there exists a finite constant C_{sup} so that, for any n, for any $x_1,x_2\in[0,n^{1/d}]^d$,*

$$\int_{\mathbb{R}^d}\frac{1}{1+|x_1-x|^{d+1}}\frac{1}{1+|x_2-x|^{d+1}}\mathbf{1}_{C(x)\neq C(x_1)}\mathbf{1}_{C(x)\neq C(x_2)}dx \leq C_{sup}f(D_\Delta(x_1,x_2))\frac{1}{1+|x_1-x_2|^{d+1}}.$$

Lemma A.24. *Use the notation n_2,Δ, $C(t)$, $f(T)$ and $D_\Delta(x_1,x_2)$ of Definition A.7 and Lemma A.23. Then, when $n_2=o(n)$,*

$$\frac{1}{n}\int_{[0,n^{1/d}]^d}dx_1\int_{[0,n^{1/d}]^d}dx_2\frac{1}{1+|x_1-x_2|^{d+1}}f(D_\Delta(x_1,x_2))\to_{n\to+\infty}0.$$

Lemma A.25. *Use the notation n_2, Δ and $C_1, ..., C_{n_2}$ of Definition A.7. Let, for $i = 1, ..., n_2$, $X_1^i, ..., X_{N_i}^i$ be the N_i components of X that are in C_i (so that the order of their indexes in X is preserved). Then*

i) For $i = 1, ..., n_2$, N_i follows a binomial $B(n, 1/n_2)$ distribution. For any $i, j = 1, ..., n_2; i \neq j$, conditionally to $N_i = k_i$, N_j follows a binomial $B(n - k_i, 1/(n_2 - 1))$ distribution.

ii) Conditionally to $N_i = k_i$, $X_1^i, ..., X_{k_i}^i$ are independent and uniformly distributed on C_i.

iii) For $1 \leq i \neq j \leq n_2$, conditionally to $N_i = k_i, N_j = k_j$, the sets of random variables $(X_1^i, ..., X_{k_i}^i)$ and $(X_1^j, ..., X_{k_j}^j)$ are independent, and their components are independent and uniformly distributed on C_i and C_j respectively.

Consider n_2 real-valued functions $f_1, ..., f_{n_2}$ of X that can be written $f_i(X) = \bar{f}(N_i, X_1^i, ..., X_{N_i}^i)$, and so that, for any $t \in \mathbb{R}^d$, $x_1, ..., x_N \in \mathbb{R}^d$, $\bar{f}(N, x_1 + t, ..., x_N + t) = \bar{f}(N, x_1, ..., x_N)$. Then

iv) The variables $f_1(X), ..., f_{n_2}(X)$ have the same distribution. The couples $(f_i(X), f_j(X))$, for $1 \leq i \neq j \leq n_2$, have the same distribution.

Lemma A.26. *Use the notation of Lemma A.25, and consider n_2 functions $f_1, ..., f_{n_2}$ that satisfy the conditions of Lemma A.25. Assume that there exist fixed even natural numbers q, l and a finite constant C_{sup} (independent of n and X) so that $\mathbb{E}\left(f_i^2(X)|N_i = k\right) \leq C_{sup}(1 + k^q + k^{q+l}/\Delta^l)$. Then, if $\Delta \to_{n\to\infty} +\infty$ and $\Delta = O(n^{1/(2q+5)})$,*

$$var\left(\frac{1}{n_2}\sum_{i=1}^{n_2} f_i(X)\right) \to_{n\to\infty} 0.$$

Lemma A.27. *Let N follow the binomial distribution $B(n, 1/n_2)$, with $n/n_2 = \Delta \to_{n\to\infty} +\infty$. Then, for any $k \in \mathbb{N}$, there exists a finite constant C_{sup}, independent of n, so that*

$$\mathbb{E}\left(N^k\right) \leq C_{sup}\Delta^k.$$

Lemma A.28. *Let n_2, Δ and $C_1, ..., C_{n_2}$ be as in Definition A.7. Assume that Δ is lower bounded, as a function of n. Then, there exists a finite constant C_{sup} so that for any n, $i \in \{1, ..., n_2\}$,*

$$\sum_{j=1}^{n_2} \frac{1}{1 + d(C_i, C_j)^{d+1}} \leq C_{sup}.$$

Lemma A.29. *Let A be a real $m_1 \times m_2$ matrix and b be a m_2-dimensional real column vector. Then*

$$||Ab||^2 \leq m_1 m_2 \left(\max_{i,j} A_{i,j}^2\right) ||b||^2.$$

Acknowledgements

The author thanks Josselin Garnier and Benedikt Pötscher for constructive discussions.

Supplementary material

In the supplementary material, we give the proof of the lemmas stated in Section A.4.

References

Abrahamsen, P. (1997). A review of Gaussian random fields and correlation functions. Technical report, Norwegian computing center.

Anderes, E. (2010). On the consistent separation of scale and variance for Gaussian random fields. The Annals of Statistics, 38:870–893.

Andrianakis, I. and Challenor, P. G. (2012). The effect of the nugget on Gaussian process emulators of computer models. Computational Statistics and Data Analysis, 56:4215–4228.

Azencott, R. and Dacunha-Castelle, D. (1986). Series of Irregular Observations: Forecasting and Model Building. Springer-Verlag New York.

Bachoc, F. (2013). Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model mispecification. Computational Statistics and Data Analysis, 66:55–69.

Bachoc, F. (2014). Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of gaussian processes. Journal of Multivariate Analysis, 125:1–35.

Bachoc, F., Bois, G., Garnier, J., and Martinez, J. (2014). Calibration and improved prediction of computer models by universal Kriging. Nuclear Science and Engineering, 176(1):81–97.

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., and Richet, Y. (2014). Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. Technometrics.

Cressie, N. and Lahiri, S. (1993). The asymptotic distribution of REML estimators. Journal of Multivariate Analysis, 45:217–233.

Cressie, N. and Lahiri, S. (1996). Asymptotics for reml estimation of spatial covariance parameters. Journal of Statistical Planning and Inference, 50:327–341.

Dubrule, O. (1983). Cross validation of Kriging in a unique neighborhood. Mathematical Geology, 15:687–699.

Gray, R. M. (2006). Toeplitz and circulant matrices: A review. Foundations and Trends® in Communications and Information Theory, 2(3):155–239.

Iooss, B., Boussouf, L., Feuillard, V., and Marrel, A. (2010). Numerical studies of the metamodel fitting and validation processes. International Journal of Advances in Systems and Measurements, 3:11–21.

Jones, D., Schonlau, M., and Welch, W. (1998). Efficient global optimization of expensive black box functions. Journal of Global Optimization, 13:455–492.

Kou, S. C. (2003). On the efficiency of selection criteria in spline regression. Probability Theory and Related Fields, 127:153–176.

Lahiri, S. N. (2003). Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs. Sankhyã: The Indian Journal of Statistics, 65:356–388.

Lahiri, S. N. and Mukherjee, K. (2004). Asymptotic distributions of M-estimators in a spatial regression model under some fixed and stochastic spatial sampling designs. Annals of the Institute of Statistical Mathematics, 56:225–250.

Lahiri, S. N. and Zhu, J. (2006). Resampling methods for spatial regression models under a class of stochastic designs. The Annals of Statistics, 34(4):1774–1813.

Le Gratiet, L. and Garnier, J. (2014). Asymptotic analysis of the learning curve for gaussian process regression. Machine Learning, pages 1–27.

Loh, W. (2005). Fixed domain asymptotics for a subclass of Matérn type Gaussian random fields. The Annals of Statistics, 33:2344–2394.

Mardia, K. and Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71:135–146.

Martin, J. and Simpson, T. (2004). On the use of Kriging models to approximate deterministic computer models. In DETC'04 ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference Salt Lake City, Utah USA, September 28 - October 2, 2004.

Paulo, R., Garcia-Donato, G., and Palomo, J. (2012). Calibration of computer models with multivariate output. Computational Statistics and Data Analysis, 56:3959–3974.

Rasmussen, C. and Williams, C. (2006). Gaussian Processes for Machine Learning. The MIT Press, Cambridge.

Ripley, B. (1981). Spatial Statistics. Wiley, New York.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. Statistical Science, 4:409–423.

Santner, T., Williams, B., and Notz, W. (2003). The Design and Analysis of Computer Experiments. Springer, New York.

Stein, M. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. The Annals of Statistics, 16:55–63.

Stein, M. (1990a). Bounds on the efficiency of linear predictions using an incorrect covariance function. The Annals of Statistics, 18:1116–1138.

Stein, M. (1990b). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. The Annals of Statistics, 18:1139–1157.

Stein, M. (1990c). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. The Annals of Statistics, 18:850–872.

Stein, M. (1999). Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Stein, M. L. (1993). Spline smoothing with an estimated order parameter. The Annals of Statistics, 21:1522–1544.

Sundararajan, S. and Keerthi, S. (2001). Predictive approaches for choosing hyperparameters in Gaussian processes. Neural Computation, 13:1103–1118.

Vazquez, E. (2005). Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications. PhD thesis, Université Paris XI Orsay. Available at http://tel.archives-ouvertes.fr/tel-00010199/en.

White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica, 50:1–25.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. Journal of Multivariate Analysis, 36:280–296.

Ying, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. The Annals of Statistics, 21:1567–1590.

Zhang, H. (2004). Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. Journal of the American Statistical Association, 99:250–261.

Zhang, H. and Wang, Y. (2010). Kriging and cross validation for massive spatial data. Environmetrics, 21:290–304.

Zhang, H. and Zimmerman, D. (2005). Toward reconciling two asymptotic frameworks in spatial statistics. Biometrika, 92:921–936.

Zheng, Y. and Zhu, J. (2012). On the asymptotics of maximum likelihood estimation for spatial linear models on a lattice. Sankhya, 74-A:29–56.