Bernoulli **24**(2), 2018, 1531–1575
DOI: 10.3150/16-BEJ906

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

FRANÇOIS BACHOC

Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse. E-mail: francois.bachoc@math.univ-toulouse.fr

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 shown that, for irregular spatial sampling of observation points, Cross Validation can yield smaller prediction errors than Maximum Likelihood. Motivated by this observation, we provide a general asymptotic analysis of the misspecified case, for independent and uniformly distributed observation points. 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 Monte Carlo simulations, 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 independent and uniformly distributed observation points.

Keywords: covariance parameter estimation; cross validation; Gaussian processes; increasing-domain asymptotics; integrated square prediction error; Kullback–Leibler divergence; maximum likelihood

1. Introduction

Kriging models [37,48] 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 [39,40] and calibration [35] or global optimization [24].

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 [48], Chapter 6.8, the estimate in the Kriging equations. Usually, it is assumed that the covariance function belongs to a given parametric family (see [1] for a review of classical families). In this case, the estimation boils down to estimating the corresponding covariance parameters. For covariance, parameter estimation, Maximum Likelihood (ML) is the most studied and used method, while Cross Validation (CV) [49,55] is an alternative technique.

1350-7265 © 2018 ISI/BS

Consider first the case where the true covariance function of the Gaussian process belongs to the parametric family of covariance functions used for estimation, which we call the well-specified case. Then, it is shown in several references that ML should be preferred over CV. It is proved in [46] 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 [7], 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 [40], Chapter 3, or deterministic studies as in [34]. 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 [34] and having a free smoothness parameter in [40]. Finally, in situations similar to the well-specified case, ML-type methods have been shown to be preferable over CV-type methods in [47] 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, which 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 [10]), which de facto makes the covariance model misspecified if the Gaussian process has a different order of smoothness. In the misspecified case, [5] shows that, provided the spatial sampling of observation points is not too regular, CV can yield a smaller integrated square prediction error than ML. In a context of spline approximation methods, [47] and [25] also suggest that CV-type methods can provide smaller prediction errors than ML-type methods under misspecification.

In this paper, we primarily aim at showing, in agreement with the preceding discussion, that CV can provide asymptotically optimal integrated square prediction errors under misspecification. In this regard, the two most studied asymptotic frameworks in the Kriging literature are the increasing-domain and fixed-domain asymptotics [48], page 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 and confidence intervals. In this asymptotic framework, two types of covariance parameters can be distinguished: microergodic and non-microergodic covariance parameters. Following the definition in [48], a covariance parameter is microergodic if two covariance functions are orthogonal whenever they differ for it (as in [48], 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 and confidence intervals [43–45, 54]. On the contrary, there is a fair amount of literature on consistent estimation of microergodic covariance parameters [2,31,52–54]. Consistent estimation of microergodic parameters is shown, in some cases, to entail asymptotically optimal predictions and confidence intervals [36].

Nevertheless, a 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, [53] addresses ML for the tensorized exponential model only and [31] addresses ML for the Matérn $3/2$ covariance model only.

Hence, in this paper, we work under increasing-domain asymptotics, in which case results can be proved for fairly general covariance models [7,12,13,15,32]. In fact, generally speaking, under increasing-domain asymptotics, all (identifiable) covariance parameters have a strong asymptotic influence on predictions [7] and can be consistently estimated with asymptotic normality [7, 32]. 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. Note that, beside Kriging, increasing-domain asymptotics is largely considered in spatial statistics [21,28].

The increasing-domain asymptotic setting we consider in this paper consists of n independent observation points with uniform distribution on $[0, n^{1/d}]^d$, for $d \in \mathbb{N}^*$. 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. On the other hand, we prove in Theorem 3.3 that ML asymptotically minimizes, the Kullback–Leibler divergence from the true covariance function, defined at the observation vector. This latter finding does not provide information on the prediction errors of the Gaussian process at new points, stemming from ML. Thus, an asymptotic confirmation is given to the empirical finding of [5], that when the spatial sampling is not too regular, CV can provide smaller integrated square prediction errors than ML in the misspecified case.

On a more technical level, we provide increasing-domain asymptotic results for matrix-form estimation criteria with independent and uniformly distributed observation points. To the best of our knowledge, this type of situation has not been addressed in the existing literature.

We conclude this paper by Monte Carlo simulations, illustrating Theorems 3.3 and 3.4. The simulations highlight that, in some cases, 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. In the Monte Carlo simulations, we also present a case of non-Gaussian misspecification which, although out of the scope of Theorems 3.3 and 3.4, yields interesting conclusions.

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.

Finally, note that one should be cautious about inferring from this paper that CV is preferable over ML in the misspecified case. Indeed, there exist other prediction scores than the integrated square prediction error (see [17,18]) some of them also assessing the coverage of the confidence intervals obtained from the Kriging model. The main contribution of this paper is to provide rigorous results for CV, relatively to the integrated square prediction error only, which is nonetheless a largely considered criterion for comparing predictors.

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 $X_1, \dots, X_n \in \mathbb{R}^d$, for $n \in \mathbb{N}^*$. That is, for $i = 1, \dots, n$, we observe $y_i = Y(X_i) + \varepsilon_i$, where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$, Y and

$(X_1, \dots, X_n)$ 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 $(X_1, \dots, X_n)$ is specified and discussed in Condition 2.4 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 [9] or to Monte Carlo computer experiments [30]. 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 $X_1, \dots, X_n$ are two by two distinct. [This is the case in this paper, in an almost sure sense, see Condition 2.4.]

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, an existing practice (e.g., [11,34]) is to use the Gaussian covariance model, where $p = d + 1$, $\Theta \subset (0, \infty)^p$, $\theta = (\sigma^2, \ell_1, \dots, \ell_d)$ and $K_\theta(t) = \sigma^2 \exp(-\sum_{i=1}^d t_i^2 / \ell_i^2)$. With the Gaussian covariance model, all the covariance functions K_θ generate Gaussian process realizations that are almost surely infinitely differentiable (see the second part of Theorem 5 in [41]). Thus, the Gaussian model is de facto misspecified if the realizations of Y have only a finite order of differentiability. [Note that the use of the Gaussian covariance model is dis-advised in several references, see [48].] In theory, the Matérn model considered in Section 4 provides more flexibility 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 [10]).

In this paper, we are primarily interested in analyzing the misspecified case although the asymptotic results that are given in Section 3 are valid for both the well-specified and misspecified cases.

We let $X = (X_1, \dots, X_n)$ 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 define the random vector $y = (y_1, \dots, y_n)^t$ of size n by $y_i = Y(X_i) + \varepsilon_i$. 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 the covariance parameter θ.

2.2. Maximum likelihood and cross validation estimators

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

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

is the modified opposite log-likelihood.

Remark 2.1. For conciseness, we do not write explicitly the dependence of R_θ, R_0, y and 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 allow 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 $\hat{\theta}_{\mathrm{ML}}$ minimizing it. The same remark can be made for the CV criterion (2). We refer to Remark 2.1 in [7] for the existence of measurable minimizers of the ML and CV criteria.

Under several increasing-domain asymptotics settings, ML is consistent and asymptotically normal, with mean vector 0 and covariance matrix the inverse of the Fisher information matrix. This is shown in [32], 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 [12,13]. In [7], 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 $\hat{\theta}_{\mathrm{CV}} \in \operatorname{argmin}_\theta \mathrm{CV}_\theta$, with

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

where $\hat{y}_{i,\theta} := \mathbb{E}_{\theta|X}(y_i|y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$ 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 covariance function K_θ and variance δ_θ, given X, so that $\mathbb{E}_{\theta|X}(y_i|y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n) = \mathbb{E}_\theta(y_i|X, y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$.

Let $r_{i,\theta} = (K_\theta(X_i, X_1), \dots, K_\theta(X_i, X_{i-1}), K_\theta(X_i, X_{i+1}), \dots, K_\theta(X_i, X_n))^t$. 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 $y_{-i} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)^t$, we have $\hat{y}_{i,\theta} = r_{i,\theta}^t R_{i,\theta}^{-1} y_{-i}$.

The criterion (2) can be computed with a single matrix inversion, by means of virtual LOO formulas (see, e.g., [14,38]). These virtual LOO formulas yield, when writing $\operatorname{diag}(A)$ for the matrix obtained by setting to 0 all the off diagonal elements of a square matrix A,

$$\mathrm{CV}_\theta := \frac{1}{n} y^t R_\theta^{-1}\big(\operatorname{diag}\big(R_\theta^{-1}\big)\big)^{-2} R_\theta^{-1} y, \tag{3}$$

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

Finally, in [7] 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 with a randomly perturbed grid of observation points.

Remark 2.3. Note that, as follows from (3), CV_θ 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 the pair

$(K_\theta/K_\theta(0), \delta_\theta/K_\theta(0))$. In particular, the CV criterion (2) does not assess the validity of quantities like $\mathrm{var}_{\theta|X}(Y(t)|y)$, where $\mathrm{var}_{\theta|X}$ denotes the variance under parameter θ given X. Hence, the Kriging predictive confidence intervals obtained by CV can be unreliable and only the predictors $\mathbb{E}_{\theta|X}(Y(t)|y)$ of the values of Y at new points t are relevant. These predictors alone provide the same applicability as many regression techniques like kernel regression or neural network methods and can be used in a wide range of applications.

For some covariance models, any two different values of θ yield two different pairs $(K_\theta/K_\theta(0), \delta_\theta/K_\theta(0))$, and thus two different predictor functions $\mathbb{E}_{\theta|X}(Y(t)|y)$. For these covariance models, the CV criterion CV_θ is hence meant to estimate the full covariance parameter θ. One important instance of these models is when $\delta_\theta = \delta_1$ for all $\theta \in \Theta$ and where different values of θ yield different covariance functions K_θ. The Monte Carlo simulations of Section 4 lie in this framework.

For other covariance models, there can exist different values of θ yielding identical pairs $(K_\theta/K_\theta(0), \delta_\theta/K_\theta(0))$, and thus identical predictor functions $\mathbb{E}_{\theta|X}(Y(t)|y)$. The selection between these values of θ should thus be carried out based on criteria which do not involve only the leave-one-out conditional means $\hat{y}_{i,\theta}$. In these cases, it is possible to use the two-step estimation procedure proposed in [5], or the log predictive probability criterion ([37], Chapter 5, [49,55]). Both of these estimation methods take into account $\mathrm{var}_{\theta|X}(y_i|y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$.

In this paper, we shall not investigate these procedures aiming at distinguishing between values of θ yielding identical pairs $(K_\theta/K_\theta(0), \delta_\theta/K_\theta(0))$. Note that Theorem 3.4 below keeps the same interpretation and applicability also when different values of θ yield identical pairs $(K_\theta/K_\theta(0), \delta_\theta/K_\theta(0))$, see Remark 3.6.

2.3. Random spatial sampling

We consider an increasing-domain asymptotic framework where the observation points are independent and uniformly distributed, which constitutes the archetype of an irregular spatial sampling.

Condition 2.4. For all $n \in \mathbb{N}^*$, the observation points $X_1, \dots, X_n$ are random and follow independently the uniform distribution on $[0, n^{1/d}]^d$. The variables Y, $(X_1, \dots, X_n)$ and ε are mutually independent.

Condition 2.4 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., [56]), which is not the case here. In [26] and [27], the term increasing-domain is also used, when points are sampled randomly on a domain with volume proportional to n.

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 $|t| = \max_{i=1,\dots,d} |t_i|$.

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

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

In addition, for any $k \in \mathbb{N}$, for any two-by-two distinct points $x_1, \dots, x_k$, the matrix $(K_0(x_i - x_j))_{1 \leq i,j \leq k}$ is invertible. Finally, we have $\delta_0 \geq 0$.

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 $i_1, \dots, i_p \in \mathbb{N}$ so that $i_1 + \dots + i_p \leq p+1$, there exists $A_{i_1,\dots,i_p} < +\infty$ so that for all $t \in \mathbb{R}^d$, $\theta \in \Theta$,

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

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 $i_1, \dots, i_p \in \mathbb{N}$ so that $i_1 + \dots + i_p \leq p+1$, there exists $B_{i_1,\dots,i_p} < +\infty$ so that for all $\theta \in \Theta$,

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

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 [7], 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 [7] 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 [7]. 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. Indeed, the ML and CV criteria (1) and (3) involve the inverse covariance matrix R_θ^{-1}, and other inverses of covariance matrices obtained from K_θ. Having the upper bound $1/(\inf_{\theta \in \Theta} \delta_\theta)$ on the spectral norm of R_θ^{-1} is thus necessary in our proof. We remark that, on the other hand, the inverse covariance matrix R_0^{-1} does not appear in the expression of the ML and CV criteria, see (1) and (3). Thus, the asymptotic results given in this paper remain valid in the case $\delta_0 = 0$.

On a practical standpoint, the primary application cases addressed by this paper are measure errors, stochastic outputs and nugget effect, where $\delta_0 > 0$ and $\inf_{\theta \in \Theta} \delta_\theta > 0$, see the discussion in Section 2. The case $\delta_0 = 0$ and $\inf_{\theta \in \Theta} \delta_\theta > 0$ can also be relevant for applications. Indeed, 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 [3].

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 conditional Kullback–Leibler divergence of the latter distribution from the former is $(1/2)[\log(\det(R_\theta R_0^{-1})) + \operatorname{Tr}(R_0 R_\theta^{-1}) - n]$. We define the normalized Kullback–Leibler divergence $D_{n,\theta}$ by multiplying the above conditional Kullback–Leibler divergence by $2/n$. Hence, we have

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

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 the normalization with factor $2/n$ leading to $D_{n,\theta}$ is appropriate 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 [7], 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 $n \to +\infty$ [4]. This limit is twice the asymptotic Kullback information in [4] 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 2.4, 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 2.4, 3.1 and 3.2, we have, as $n \to \infty$,*

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

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 i.i.d. setting, ML asymptotically minimizes the Kullback–Leibler divergence (which does not depend on sample size) from the true distribution, within a misspecified parametric model [51]. 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 2.4 (e.g., for the randomly perturbed regular grid of [7]). Nevertheless, to the best of our knowledge, in the context of Condition 2.4, Theorem 3.3 is not a simple consequence of the existing literature, and an original proof is provided in the Appendix.

The Kullback–Leibler divergence is of course a central quality criterion for covariance parameters. Nevertheless, in the misspecified-case, Theorem 3.3 does not imply that ML is optimal for other common quality criteria, such as the integrated square prediction error introduced below. In addition, note that the Kullback–Leibler divergence addresses the distribution of the Gaussian process only at the observation points, thus providing no information on the inference of the values of Y at new points, obtained from $(K_\theta, \delta_\theta)$.

3.3. Cross validation

Let us recall the notation $\mathbb{E}_{\theta|X}(Y(t)|y) = \mathbb{E}_\theta(Y(t)|y, X)$ and 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

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

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. This criterion (or evaluations of it) is very commonly used, in particular for Gaussian process surrogate models of computer experiments (see, e.g., [19,33]). More generally, the square prediction error is largely considered to evaluate predictors, see [17].

It is natural to consider that the first objective of the CV estimator $\hat{\theta}_{\mathrm{CV}}$ is to yield a small $E_{n,\hat{\theta}_{\mathrm{CV}}}$. If the observation points $X_1, \dots, X_n$ 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 for example, [23] and [5]. If however the observation points $X_1, \dots, X_n$ are not regularly spaced, then it is shown numerically in [5] that the CV estimator $\hat{\theta}_{\mathrm{CV}}$ can yield a small $E_{n,\hat{\theta}_{\mathrm{CV}}}$ and, especially, smaller than $E_{n,\hat{\theta}_{\mathrm{ML}}}$. The following theorem, which is the main contribution of this paper, supports this conclusion under increasing-domain asymptotics.

Theorem 3.4. *Under Conditions 2.4, 3.1 and 3.2, we have, as $n \to \infty$,*

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

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 $n \to \infty$.

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 $\hat{\theta}_{\mathrm{CV}}$ relatively to the predictions it yields for the actual Gaussian process at hand. Theorem 3.4 does not only confirm that CV will not provide asymptotically larger integrated square prediction errors than ML, with independent and uniformly distributed observation points, it also shows that these integrated square prediction errors will be asymptotically minimal, over all possible estimators.

The setting of the proof of Theorem 3.4 combines independent and uniformly distributed observation points with the matrix-form estimation criteria (1) and (3). These criteria and their derivatives involve imbrications of covariance matrix derivatives and inverse covariance matrices, which can generally not be put in explicit matrix-free forms. To the best of our knowledge, this specific combination has not been addressed in the previous literature.

Indeed, on the one hand, when matrix-form criteria like (1) and (3) are treated, it is assumed, implicitly or explicitly that there exists a positive minimal distance between two different observation points. This is the case in [7]. Also, [32] and [12,13] 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. This minimal distance assumption does not hold with independent observation points. Instead clusters of closely spaced observation points may appear. As a consequence, the maximum eigenvalues of the covariance matrices and their derivatives are not upper bounded, even in probability, which brings new obstacles for the analysis of criteria like (1) and (3). In addition, considering random observation points with no underlying grid structure makes it more challenging to control the fluctuations of functions of (random) covariance matrices, compared to Proposition D.7 of [7], for instance.

On the other hand, when independent and uniformly distributed observation points are considered (see, e.g., [26,27,29]), the quantities of interest do not involve derivatives and inverse of $n \times n$ covariance 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.

Remark 3.5. 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 $(X_1, \dots, X_n)$. Hence, using the symmetry of $X_1, \dots, X_{n+1}$, 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.

Remark 3.6. Theorem 3.4 holds regardless of whether there is a unique θ minimizing CV_θ or not. In particular, in some cases it is possible to have $\theta = (s, \bar{\theta})$ and $\Theta = S \times \bar{\Theta}$, where CV_θ and $E_{n,\theta}$ actually depend only on $\bar{\theta}$. Then, in these cases $\hat{s}$ would be obtained separately from $\hat{\bar{\theta}}_{\mathrm{CV}}$, and Theorem 3.4 would read $E_{n,\hat{\bar{\theta}}_{\mathrm{CV}}} = \inf_{\bar{\theta} \in \bar{\Theta}} E_{n,\bar{\theta}} + o_p(1)$ and would not address $\hat{s}$.

One instance of the situation addressed above is when one has the decomposition $\theta = (s, \bar{\theta})$, $\Theta = [0, \infty) \times \bar{\Theta}$, with $\bar{\theta} = (\bar{\theta}_1, \bar{\theta}_2)$, $\bar{\Theta} = \bar{\Theta}_1 \times \bar{\Theta}_2$, $\bar{\theta}_1 \in \bar{\Theta}_1$, $\bar{\theta}_2 \in \bar{\Theta}_2$, and where $(K_\theta, \delta_\theta) = (s\bar{K}_{\bar{\theta}_1}, s\bar{\theta}_2)$, where $\bar{K}_{\bar{\theta}_1}$ is a correlation function. When setting $\bar{\Theta}_2 = \{0\}$, we find back the setting of [5], where $\hat{s}$ can be obtained from equation (7) in this reference. Note that Theorem 3.4 does not apply when $\bar{\Theta}_2 = \{0\}$.

4. Monte Carlo simulations

We illustrate Theorems 3.3 and 3.4 in several Monte Carlo simulations. First, we consider an illustrative one-dimensional case. Then we present several two-dimensional settings. Finally, we present a simulation study where the observed stochastic process is non-Gaussian. This last simulation is out of the scope of Theorems 3.3 and 3.4, but is a case of model misspecification providing interesting conclusions.

4.1. An illustrative one-dimensional case

We consider the Matérn covariance model in dimension $d = 1$ [37,40,48]. A covariance function on $\mathbb{R}$ is Matérn (σ^2, ℓ, ν) when it is written

$$K_{\mathrm{mat},\sigma^2,\ell,\nu}(t) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} \left(2\sqrt{\nu}\frac{|t|}{\ell}\right)^{\nu} K_\nu\left(2\sqrt{\nu}\frac{|t|}{\ell}\right), \tag{6}$$

with Γ the Gamma function and K_ν the modified Bessel function of the second kind with order ν. The parameters σ^2, ℓ and ν are respectively the variance, correlation length and smoothness parameters. Note that another parametrization of the Matérn covariance function exists, see, for instance, equation (32) in [48]. The parametrization (6) is the one used for instance in [22,37,40]. We find that this parameterization provides a good interpretation of the variance, correlation length and smoothness parameters. Indeed, the correlation length ℓ is only involved in the $|t|/\ell$ term in (6) and thus acts solely as a scale parameter. Note also the interesting fact that as $\nu \to \infty$,

$K_{\mathrm{mat},\sigma^2,\ell,\nu}(t) \to \sigma^2 e^{-t^2/\ell^2}$, as is obtained from the discussion following equation (4.14) in [37]. Hence, the value of ℓ, as a spatial correlation length, can be interpreted independently of ν. Further discussion of this interpretation is provided in Remark 2.28 in [6], as well as in Figures 4 and 5 in Chapter 2 of [48].

In the one-dimensional 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 $\delta_0 = 0.25^2$.

We consider that the noise variance is fixed (for all $\theta \in \Theta$, $\delta_\theta = \delta_1$), and that the covariance model is given by $\theta = (\sigma^2, \ell)$ and $K_{\sigma^2,\ell} = K_{\mathrm{mat},\sigma^2,\ell,\nu_0}$, so that the smoothness parameter ν_0 is known. 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.] Note that for CV, we have $(\hat{\sigma}^2_{\mathrm{CV}}, \hat{\ell}_{\mathrm{CV}}) \in \operatorname{argmin}_{(\sigma^2,\ell)\in\Theta} \mathrm{CV}_{(\sigma^2,\ell)}$, and that even for a fixed ℓ, different values of σ^2 may yield different values of $\mathrm{CV}_{(\sigma^2,\ell)}$ (as the ratio $K_{\sigma^2,\ell}(0)/\delta_1$ depends on σ^2 and has an impact on the conditional means).

The well-specified case corresponds to $\delta_1 = \delta_0$ and the misspecified case corresponds to $\delta_1 = 0.1^2 \neq \delta_0$. These settings are 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., [9]). 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 $n = 100, 500$ and $N = 2000$ we repeat N data generations, estimations and quality criterion computations and average the results. More specifically, we simulate N independent realizations of the n 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 Table 1 we report, for $n = 100$ and $n = 500$, for the well-specified and misspecified cases and for ML and CV, the averages and standard deviations of the estimates $\hat{\ell}$, and of the values of the error criteria $D_{n,\hat{\sigma}^2,\hat{\ell}}$ and $E_{n,\hat{\sigma}^2,\hat{\ell}}$.

Let us first discuss the case $n = 100$. 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. 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 (as can be seen in Figure 1 of the supplementary material [8]). The ML estimates of σ^2 are not reported in Table 1 for the sake of concision and are close to 1, so that, approximately, the variance of the observations, as estimated by ML,

Table 1. Simulation of $N = 2000$ independent realizations of $n = 100$ or $n = 500$ i.i.d. 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 averages and standard deviations of $\hat{\ell}$, $D_{n,\hat{\sigma}^2,\hat{\ell}}$ and $E_{n,\hat{\sigma}^2,\hat{\ell}}$ are reported. 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 $E_{n,\hat{\sigma}^2,\hat{\ell}}$

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	3.035	0.379	0.073	0.023
	Well-specified	CV	3.390	1.066	0.084	0.195
	Misspecified	ML	1.166	0.581	0.247	1.033
	Misspecified	CV	3.438	1.166	0.087	3.541
500	Well-specified	ML	3.007	0.157	0.070	0.004
	Well-specified	CV	3.076	0.382	0.072	0.034
	Misspecified	ML	1.011	0.278	0.238	0.968
	Misspecified	CV	3.076	0.386	0.072	3.407

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 computed 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, the empirical distribution of $\hat{\ell}_{\mathrm{CV}}$ is approximately the same between the well-specified and misspecified cases (see also Figure 1 in the supplementary material [8]). In the misspecified case, we find that the empirical distribution of $\hat{\sigma}^2_{\mathrm{CV}}$ (not reported in Table 1 for the sake of concision) 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 averages of $E_{n,\hat{\sigma}^2,\hat{\ell}}$ and $D_{n,\hat{\sigma}^2,\hat{\ell}}$ for ML and CV in Table 1 confirm the discussion on the estimated parameters in the misspecified case. For ML which estimates small correlation lengths, the average of the error criteria $E_{n,\hat{\sigma}^2_{\mathrm{ML}},\hat{\ell}_{\mathrm{ML}}}$ is approximately 3 times larger than in the well-specified case. The average of the error criteria $D_{n,\hat{\sigma}^2_{\mathrm{ML}},\hat{\ell}_{\mathrm{ML}}}$ also increases and becomes larger than that of both ML and CV in the well-specified case. For CV, the average of the error criteria $E_{n,\hat{\sigma}^2_{\mathrm{CV}},\hat{\ell}_{\mathrm{CV}}}$ is, as discussed, as small as in the well-specified case and approximately 3 times smaller than for ML, illustrating Theorem 3.4. However, the average of $D_{n,\hat{\sigma}^2_{\mathrm{CV}},\hat{\ell}_{\mathrm{CV}}}$ is 3 times larger for CV than for ML, in the misspecified case, illustrating Theorem 3.3.

Finally, for $n = 500$ in Table 1, the relative differences between ML and CV are the same as for $n = 100$. The estimates of ℓ under ML and CV have less variance than for $n = 100$.

We remark that, 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.

Finally, in Figures 1 and 2 of the supplementary material [8], we provide the histograms of the $N = 2000$ values of $\hat{\ell}$, $E_{n,\hat{\sigma}^2,\hat{\ell}}$ and $D_{n,\hat{\sigma}^2,\hat{\ell}}$. For $n = 100$, these histograms are unimodal, although the boundary of the optimization domain can be reached by $\hat{\ell}$, for ML. For $n = 500$, the histograms are more concentrated and become symmetric.

4.2. Two-dimensional settings

We now address the case $d = 2$. We proceed exactly as for $d = 1$, and consider different configurations of (K_0, δ_0) and $\{(K_\theta, \delta_\theta), \theta \in \Theta\}$. In all the configurations, we set $\theta = (\sigma^2, \ell)$, that is we estimate one variance parameter and one correlation length, common to the two spatial directions (isotropic case). Furthermore, as in the one-dimensional case above, we set, for all $\theta \in \Theta$, $\delta_\theta = \delta_1$, where $\delta_1 = \delta_0$ in the well-specified case and where δ_1 can differ from δ_0 in the misspecified case.

Misspecification of δ

We consider the same case of misspecification of the noise variance as for the illustrative one-dimensional case. With $\|t\|_2 = (t_1^2 + t_2^2)^{1/2}$, we let $K_{\sigma^2,\ell}(t) = K_{\mathrm{mat},\sigma^2,\ell,10}(\|t\|_2)$, with $K_{\mathrm{mat},\sigma^2,\ell,\nu}$ as in (6). We let $K_0 = K_{\sigma_0^2,\ell_0}$ with $(\sigma_0^2, \ell_0) = (1, 4)$ and we let $\delta_0 = 0.25^2$. In the well-specified case, $\delta_1 = \delta_0$ while in the misspecified case $\delta_1 = 0.1^2 \neq \delta_0$. The obtained results are presented in Table 2. The conclusions and the interpretation are identical to the one-dimensional case.

Table 2. Similar settings as in Table 1 but with $d = 2$. The conclusions and the interpretation are identical to the one-dimensional case

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	4.014	0.600	0.021	0.026
	Well-specified	CV	4.525	1.564	0.024	0.123
	Misspecified	ML	1.279	0.385	0.112	1.120
	Misspecified	CV	4.637	1.754	0.024	3.725
500	Well-specified	ML	3.990	0.244	0.016	0.004
	Well-specified	CV	4.158	0.698	0.016	0.031
	Misspecified	ML	1.216	0.122	0.104	1.076
	Misspecified	CV	4.167	0.727	0.016	3.477

Misspecification of ν

We address a case where the smoothness parameter of the Matérn covariance function is misspecified. We let $\delta_1 = \delta_0 = 0.1^2$ in both the well-specified and misspecified cases. Let $K_{\mathrm{mat},\sigma^2,\ell,\nu}$ be as in (6). We let $K_{\sigma^2,\ell}(t) = K_{\mathrm{mat},\sigma^2,\ell,10}(\|t\|_2)$ in both the well-specified and misspecified cases. With $(\sigma_0^2, \ell_0) = (1, 4)$ we let $K_0(t) = K_{\mathrm{mat},\sigma_0^2,\ell_0,10}(\|t\|_2)$ in the well-specified case and $K_0(t) = K_{\mathrm{mat},\sigma_0^2,\ell_0,1}(\|t\|_2)$ in the misspecified case. Hence, misspecification consists in assuming that the observed Gaussian process is smoother than in reality.

The obtained results are presented in Table 3. Similarly to Tables 1 and 2, ML performs better than CV in all aspects in the well-specified case. Note that, in the well-specified case $E_{n,\hat{\sigma}^2,\hat{\ell}}$ is smaller than in Table 2 since δ_0 is smaller than in Table 2. In the misspecified case, the error criteria $E_{n,\hat{\sigma}^2,\hat{\ell}}$ and $D_{n,\hat{\sigma}^2,\hat{\ell}}$ are considerably larger than in the well-specified case. In agreement with Theorem 3.3, ML performs significantly better for the Kullback–Leibler divergence. Also, in agreement with Theorem 3.4, CV performs better than ML for the integrated square prediction error in the misspecified case. Nevertheless, the improvement brought by CV for the integrated square prediction error is more marginal than for Tables 1 and 2. Hence, an asymmetry appears here, where using CV instead of ML degrades the Kullback–Leibler divergence more importantly than it improves the integrated square prediction error (although of course the values of $E_{n,\hat{\sigma}^2,\hat{\ell}}$ and $D_{n,\hat{\sigma}^2,\hat{\ell}}$ have different interpretation). The fact that, in the misspecified case, CV improves $E_{n,\hat{\sigma}^2,\hat{\ell}}$ compared to ML more marginally than in Table 2 is not easy to interpret. One possible interpretation is that in Table 3, the Gaussian process to predict is not smooth ($\nu = 1$ in the Matérn model), so that the choice of the covariance parameters could have less impact on the integrated square prediction error.

Misspecification with spherical and Wendland 1 covariance functions

We let, with $a_+ = \max(a, 0)$

$$K_{\mathrm{sph},\sigma^2,\ell}(t) = \sigma^2\big(1 - \|t\|_2\big)_+^2\big(1 + \|t\|_2/2\big)$$

Table 3. Similar settings as in Table 1 but with $d = 2$ and where δ is always well-specified but where ν can be misspecified. The conclusions are similar to those of Tables 1 and 2, with the exception that CV provides a lesser improvement of $E_{n,\hat{\sigma}^2,\hat{\ell}}$ in the misspecified case

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	4.019	0.416	0.006	0.025
	Well-specified	CV	4.350	1.200	0.007	0.179
	Misspecified	ML	1.511	0.195	0.094	0.283
	Misspecified	CV	2.181	0.952	0.090	1.949
500	Well-specified	ML	3.994	0.175	0.004	0.004
	Well-specified	CV	4.162	0.642	0.004	0.053
	Misspecified	ML	1.484	0.086	0.085	0.256
	Misspecified	CV	1.759	0.201	0.075	1.588

and

$$K_{\text{wen1},\sigma^2,\ell}(t) = \sigma^2\left(1 - \|t\|_2\right)_+^4\left(1 + 4\|t\|_2\right)$$

be respectively, the spherical and Wendland 1 covariance functions (see, e.g., [15,16,50]). Both of these covariance functions are compactly supported. The spherical covariance function is not differentiable at the origin, while the Wendland 1 covariance function is. We let $\delta_0 = \delta_1 = 0.1^2$. We let $K_{\sigma^2,\ell}(t) = K_{\text{wen1},\sigma^2,\ell}(t)$ in both the well-specified and misspecified cases. We let $K_0(t) = K_{\text{wen1},1,6}(t)$ in the well-specified case and we let $K_0(t) = K_{\text{sph},1,6}(t)$ in the misspecified case.

The results obtained are presented in Table 4. The conclusions are the same as for Table 3. In articular CV marginally improves the integrated square prediction error in the misspecified case

Table 4. Similar settings as in Table 1 but with $d = 2$ and where the spherical and Wendland 1 covariance functions are investigated. The conclusions are the same as for Table 3

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	6.000	0.809	0.048	0.025
	Well-specified	CV	6.234	1.515	0.050	0.391
	Misspecified	ML	3.364	0.708	0.202	0.347
	Misspecified	CV	6.390	1.767	0.174	5.125
500	Well-specified	ML	5.996	0.345	0.041	0.004
	Well-specified	CV	6.095	0.878	0.041	0.065
	Misspecified	ML	3.361	0.349	0.191	0.325
	Misspecified	CV	6.653	1.103	0.160	4.578

compared to ML. Again, this could be due to the lack of smoothness of the Gaussian process to predict (having spherical covariance function).

4.3. A non-Gaussian case

In the situations treated above, the observed stochastic process Y is Gaussian, and, in the misspecified case, its covariance model (K_0, δ_0) does not belong to $\{(K_\theta, \delta_\theta); \theta \in \Theta\}$. We now address an other type of model misspecification, where $(K_0, \delta_0) \in \{(K_\theta, \delta_\theta); \theta \in \Theta\}$ but where Y is not a Gaussian process. We remark that this situation is not addressed in the theoretical results of Section 3.

We set $d = 2$. In both the well-specified and misspecified cases, $\delta_0 = 0.1^2$ and $\delta_\theta = 0.1^2$ for all $\theta \in \Theta$. Furthermore, in both cases, $\theta = (\sigma^2, \ell)$ and $K_\theta(t) = \sigma^2 \exp(-(\|t\|_2/\ell)^2)$. In the well-specified case, Y is a Gaussian process with covariance function K_{θ_0} with $\theta_0 = (\sigma_0^2, \ell_0) = (1, 4)$.

We base the misspecified case on a spectral decomposition of K_{θ_0}. Following Proposition 1 in [42], let $\beta = 4/\ell_0^2$, let for $i \in \mathbb{N}$, H_i be the ith order Hermite polynomial, let

$$\lambda_i = \sqrt{\frac{2}{1 + \beta + \sqrt{1 + 2\beta}}} \left(\frac{\beta}{1 + \beta + \sqrt{1 + 2\beta}} \right)^i$$

and for $x \in \mathbb{R}$

$$\phi_i(x) = \frac{(1 + 2\beta)^{1/8}}{\sqrt{2^i i!}} \exp\left(-\frac{x^2}{2} \frac{\sqrt{1 + 2\beta} - 1}{2} \right) H_i\left(\left(\frac{1}{4} + \frac{\beta}{2} \right)^{1/4} x \right).$$

Then, for any sequence $(z_{i,j})_{i,j \in \mathbb{N}}$ of i.i.d. random variables with mean 0 and variance 1, we have that, if for $t = (t_1, t_2) \in \mathbb{R}^2$

$$Y(t) = \sigma_0 \sum_{i,j=1}^{+\infty} \sqrt{\lambda_i \lambda_j} \phi_i(t_1) \phi_j(t_2) z_{i,j}, \tag{7}$$

then Y is a stochastic process with zero mean function and covariance function K_{θ_0} (as follows from Proposition 1 in [42]). Hence, in the misspecified case, Y is defined by (7), where the $z_{i,j}$ are of the form $(v_{i,j} - \Gamma(1 - 1/5))/(\Gamma(1 - 2/5) - \Gamma(1 - 1/5)^2)^{1/2}$ where $(v_{i,j})_{i,j \in \mathbb{N}}$ is composed of i.i.d. random variables with Frechet distribution with location parameter 0, scale parameter 1 and shape parameter 5. Hence, the $z_{i,j}$ indeed have mean 0 and variance 1, as follows from the expression of the moments of the Frechet distribution. In addition, the $z_{i,j}$ are non-Gaussian variables. Hence, the just-defined stochastic process Y is non-Gaussian with mean function zero and covariance function $\sigma_0^2 \exp(-(\|t\|_2/\ell_0)^2)$. Hence, we can speak of a true covariance parameter θ_0 even in the misspecified case, since in this case, setting $\theta = \theta_0$ yields the true covariance function of Y (even if this process is non-Gaussian).

We carry out the Monte Carlo simulation in the same way as for the previous cases. For each of the $N = 2000$ iterations of the Monte Carlo simulation, we sample the trajectory of Y, the observation points $(X_1, \dots, X_n)$ and the observation vector y. We compute $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{CV}}$ and the corresponding $E_{n,\hat{\theta}}$. However, the criterion $D_{n,\hat{\theta}}$ is specific to the Gaussian case, so that we replace it by

$$EL_{n,\theta} = \mathbb{E}_{\tilde{y}|X}\big(L_\theta(\tilde{y})\big),$$

with $L_\theta(\tilde{y})$ as in (1) with y replaced by $\tilde{y}$, and where the mean value is taken conditionally to the observation points X, and when conditionally to X, $\tilde{y}$ is an independent copy of y. We remark that maximizing $EL_{n,\theta}$ with respect to $\theta \in \Theta$ is equivalent to minimizing the Kullback–Leibler divergence of the Gaussian distribution of y assumed under covariance parameter θ, from the true non-Gaussian distribution of y. Hence, the larger $EL_{n,\theta}$ is, the better. Since Y has mean zero and covariance function K_{θ_0}, $EL_{n,\hat{\theta}}$ can be computed exactly for each of the N Monte Carlo iterations, since we have

$$EL_{n,\hat{\theta}} = \frac{1}{n}\log\big(\det(R_{\hat{\theta}})\big) + \frac{1}{n}\operatorname{Tr}\big(R_0 R_{\hat{\theta}}^{-1}\big),$$

with $R_0 = [K_{\theta_0}(X_i - X_j)]_{i,j=1,n}$.

The results obtained are presented in Table 5. We observe that both the ML and CV estimators seem to be consistent for estimating ℓ_0, in the well- and misspecified cases. In particular, the biases are very small, and the estimator variances decrease when n increases. Although, to our knowledge, there do not exist increasing-domain asymptotic results regarding ML or CV estimation in the non-Gaussian case, this apparent consistency can be interpreted. Indeed, for a given covariance function K_0, the mean values of the ML and CV criteria L_θ and CV_θ are the same, regardless of whether the stochastic process Y is Gaussian or not.

We also observe that the estimation error for ℓ_0 is larger in the misspecified case, for ML when $n = 500$. Otherwise, for CV when $n = 100$ and $n = 500$ and for ML when $n = 500$, the estimation errors for ℓ_0 are comparable in the well-specified and misspecified cases. The inte-

Table 5. Same type of Monte Carlo simulation as in Table 1, but with $d = 2$, where a Gaussian covariance function is used, and where the stochastic process Y is non-Gaussian in the misspecified case. Here we can speak of a true covariance parameter θ_0 even in the misspecified case. The CV and ML estimators both estimate θ_0 with reasonable accuracy. Despite the Gaussianity misspeficiation, ML performs here better than CV in all aspects

n	Specification	Estimation	Average of $\hat{\ell}$	Standard deviation of $\hat{\ell}$	Average of $E_{n,\hat{\sigma}^2,\hat{\ell}}$	Average of $EL_{n,\hat{\sigma}^2,\hat{\ell}}$
100	Well-specified	ML	4.007	0.371	0.0048	−2.506
	Well-specified	CV	4.312	1.005	0.0055	−2.358
	Misspecified	ML	4.005	0.395	0.0047	−2.492
	Misspecified	CV	4.244	0.969	0.0054	−2.346
500	Well-specified	ML	3.995	0.145	0.0031	−2.717
	Well-specified	CV	4.081	0.439	0.0032	−2.671
	Misspecified	ML	3.985	0.259	0.0024	−2.528
	Misspecified	CV	3.975	0.497	0.0024	−2.407

grated square prediction error $E_{n,\hat{\theta}}$ is smaller in the misspecified case than in the well-specified case. Finally, the criterion $EL_{n,\hat{\theta}}$ is smaller (hence more favourable) in the well-specified case. Hence, to summarize, this non-Gaussian misspecification appears to be less unilaterally harmful for inference on the process Y, compared to misspecification of the covariance model. Note nevertheless that, in the non-Gaussian case, it is more difficult to obtain confidence intervals for the values of Y, or to sample trajectories, conditionally to the observations.

We observe in Table 5 that ML performs better than CV for all the criteria displayed there. This is in contrasts with the case of covariance model misspecification, where CV performs better than ML for the integrated square prediction error. Our interpretation is that, when the covariance model is misspecified, ML is not always asymptotically optimal for the integrated square prediction error, while CV is, as is shown in Theorem 3.4. On the other hand, when the Gaussianity is misspecified, both the CV and ML estimators appear to converge to the true covariance parameters, and could thus be both asymptotically optimal for the integrated square prediction error.

5. Discussion

Theorems 3.3 and 3.4, together with the results of the simulation studies and the existing literature, draw the following conclusion.

In the well-specified case, any covariance parameter estimator can be evaluated relatively to the estimation error criterion. The ML estimator is thus generally optimal in the well-specified case. In the simulation study, ML performs better than CV for the estimation error, the conditional Kullback–Leibler divergence and the integrated square prediction error.

On the other hand, when the covariance model is misspecified, there is not a unique quality criterion for covariance parameter estimation. Different criteria are optimized by different covariance parameters and estimators. We prove that ML asymptotically minimizes the conditional Kullback–Leibler divergence. In the case of independent and uniformly distributed observation points, we prove that CV asymptotically minimizes the integrated square prediction error. Thus, CV is asymptotically optimal for the criterion it is designed for, provided the spatial sampling is in agreement with the CV principle. As shown in the simulation studies, the estimated covariance parameters and quality criterion values can differ radically between ML and CV when the covariance model is misspecified. In this regard, we point out that ML is not optimal relatively to all the common quality criteria, contrarily to the well-specified case. Note finally that our aim is not to provide a hierarchy between ML and CV, in case of covariance model misspecification.

The fact that ML and CV typically optimize different criteria in case of covariance model misspecification 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.

We would like to mention some avenues for future research. First, the results of the Monte Carlo simulations 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 independent and uniformly distributed observation points.

Second, consider the alternative CV estimator, maximizing the log predictive probability criterion ([37], Chapter 5, [49,55]). It would be interesting to see whether this estimator can be shown to minimize with respect to θ the quality criterion $\int_{[0,n^{1/d}]^d} d_\theta(t)\,dt$, where $d_\theta(t)$ is the conditional Kullback–Leibler divergence of the conditional distribution of $Y(t)$, given y and X, assumed under $(K_\theta, \delta_\theta)$, from the corresponding true conditional distribution obtained from (K_0, δ_0).

Finally, in the Monte Carlo simulations, we have addressed a misspecification setting where the covariance model is well-specified but where the observed stochastic process is non-Gaussian. We have obtained different conclusions compared to the case of covariance model misspecification. In particular, both the CV and ML estimators appear to be consistent for estimating the true covariance parameters. It would be interesting to obtain asymptotic results for ML and CV estimation for non-Gaussian processes, that would confirm these numerical observations.

Appendix: Proofs

The appendix proof section is organized as follows. In Section A.1, we present some technical notation that are used throughout the proofs. In Section A.2, we provide the outline of the proof of Theorem 3.3. In Section A.3, we give additional proofs, completing the proof of Theorem 3.3. Similarly, in Section A.4, we provide the outline of the proof of Theorem 3.4, and in Section A.5 we complete this proof. The proofs in Sections A.2 to A.5 rely on additional technical results that are stated in Section A.6 and proved in the supplementary material [8].

The purpose of this proof organization is that the reading of Sections A.2 and A.4 alone be sufficient to get a global understanding of the proofs of Theorems 3.3 and 3.4.

A.1. Notation

In all the appendix, we consider that Conditions 2.4, 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 $|v| = \max_{i=1,\dots,m} |v_i|$. For a real $m \times m$ matrix A, we write as in [20], $|A|^2 = \frac{1}{m}\sum_{i,j=1}^m A_{i,j}^2$ and $\|A\|$ for the largest singular value of A. Both $|\cdot|$ and $\|\cdot\|$ are norms and $\|\cdot\|$ 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, $a \le 2b \le 4c$, we shall write $a \le C_{\sup} b \le C_{\sup} c$. Similarly, we write $C_{\inf}$ a generic strictly positive constant (not depending on n, X, Y, ε and θ).

A.2. Outline of proof for Theorem 3.3

We have, for all $\theta \in \Theta$,

$$\begin{aligned}
D_{n,\hat{\theta}_{\mathrm{ML}}} - D_{n,\theta} &= L_{\hat{\theta}_{\mathrm{ML}}} + D_{n,\hat{\theta}_{\mathrm{ML}}} - L_{\hat{\theta}_{\mathrm{ML}}} - L_\theta - D_{n,\theta} + L_\theta \\
&= L_{\hat{\theta}_{\mathrm{ML}}} - L_\theta + D_{n,\hat{\theta}_{\mathrm{ML}}} + 1 + \frac{1}{n}\log\bigl(\det(R_0)\bigr) - L_{\hat{\theta}_{\mathrm{ML}}} + L_\theta \\
&\quad - D_{n,\theta} - 1 - \frac{1}{n}\log\bigl(\det(R_0)\bigr) \\
&\le L_{\hat{\theta}_{\mathrm{ML}}} - L_\theta + 2\sup_{\theta\in\Theta}\left| L_\theta - \frac{1}{n}\log\bigl(\det(R_0)\bigr) - 1 - D_{n,\theta}\right| \\
&\le 2\sup_{\theta\in\Theta}\left| L_\theta - \frac{1}{n}\log\bigl(\det(R_0)\bigr) - 1 - D_{n,\theta}\right|, \qquad (8)
\end{aligned}$$

where the last inequality follows from the fact that $\hat{\theta}_{\mathrm{ML}}$ minimizes L_θ.

In order to prove Theorem 3.3, it is thus sufficient to prove that the supremum in (8) goes to zero in probability as $n \to \infty$. Note that in the developments preceding (8), we have introduced the term $[1/n]\log(\det(R_0)) + 1$ precisely so that the random variable inside the supremum in (8) has mean zero (conditionally to X) when θ is fixed.

We have, using the triangle inequality and the definitions of L_θ and $D_{n,\theta}$ in (1) and (4),

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left| L_\theta - \frac{1}{n}\log\bigl(\det(R_0)\bigr) - 1 - D_{n,\theta}\right| \\
&\quad\le \sup_{\theta\in\Theta}\bigl|L_\theta - \mathbb{E}(L_\theta|X)\bigr| + \sup_{\theta\in\Theta}\left|\mathbb{E}(L_\theta|X) - \frac{1}{n}\log\bigl(\det(R_0)\bigr) - 1 - D_{n,\theta}\right| \\
&\quad\le \sup_{\theta\in\Theta}\bigl|L_\theta - \mathbb{E}(L_\theta|X)\bigr| + \sup_{\theta\in\Theta}\left|\mathbb{E}\left(\frac{1}{n}\log\bigl(\det(R_\theta)\bigr) + \frac{1}{n}y^t R_\theta^{-1} y \middle| X\right) - \frac{1}{n}\log\bigl(\det(R_0)\bigr) - 1\right. \\
&\qquad\left. - \frac{1}{n}\log\bigl(\det\bigl(R_\theta R_0^{-1}\bigr)\bigr) - \frac{1}{n}\operatorname{Tr}\bigl(R_0 R_\theta^{-1}\bigr) + 1\right|.
\end{aligned}$$

We now use (i) in Lemma A.16 to obtain

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left| L_\theta - \frac{1}{n}\log\bigl(\det(R_0)\bigr) - 1 - D_{n,\theta}\right| \\
&\quad\le \sup_{\theta\in\Theta}\bigl|L_\theta - \mathbb{E}(L_\theta|X)\bigr| + \sup_{\theta\in\Theta}\left|\frac{1}{n}\log\bigl(\det(R_\theta)\bigr) + \frac{1}{n}\operatorname{Tr}\bigl(R_0 R_\theta^{-1}\bigr)\right. \\
&\qquad\left. - \frac{1}{n}\log\bigl(\det(R_0)\bigr) - \frac{1}{n}\log\bigl(\det\bigl(R_\theta R_0^{-1}\bigr)\bigr) - \frac{1}{n}\operatorname{Tr}\bigl(R_0 R_\theta^{-1}\bigr)\right| \\
&\quad= \sup_{\theta\in\Theta}\bigl|L_\theta - \mathbb{E}(L_\theta|X)\bigr|.
\end{aligned}$$

We now aim at showing that $\sup_{\theta\in\Theta}|L_\theta - \mathbb{E}(L_\theta|X)| = o_p(1)$. With the following lemma (proved in Section A.3), we first have a convergence result for a fixed θ.

Lemma A.1. *Consider a fixed $\theta \in \Theta$. Then*

$$\mathbb{E}\big(\big|L_\theta - \mathbb{E}(L_\theta|X)\big|\big) \underset{n\to\infty}{\to} 0.$$

In order to obtain a uniform-in-θ convergence result from Lemma A.1, we show the following lemma and corollary (proofs in Section A.3), establishing a control of derivatives w.r.t. θ.

Lemma A.2. *For $i = 1, \dots, p$,*

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

is bounded w.r.t. n.

Corollary A.3. *For any $i = 1, \dots, p$,*

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

is bounded w.r.t. n.

We can now show that $\sup_{\theta\in\Theta}|L_\theta - \mathbb{E}(L_\theta|X)| = o_p(1)$. Indeed, for fixed θ, the term $|L_\theta - \mathbb{E}(L_\theta|X)|$ goes to 0 in probability because of Lemma A.1. Hence, using Lemma A.2 and Corollary A.3, together with Lemma A.15, we have $\sup_{\theta\in\Theta}|L_\theta - \mathbb{E}(L_\theta|X)| = o_p(1)$. This finishes the proof of Theorem 3.3, up to the complementary proofs in Section A.3.

A.3. Complement of proof for Theorem 3.3

Proof of Lemma A.1. We have, applying Jensen inequality twice and (ii) in Lemma A.16,

$$\mathbb{E}\big(\big|L_\theta - \mathbb{E}(L_\theta|X)\big|\big) \le \mathbb{E}\big(\sqrt{\operatorname{var}(L_\theta|X)}\big) \le \sqrt{\mathbb{E}\big(\operatorname{var}(L_\theta|X)\big)} = \sqrt{\mathbb{E}\left(\frac{2}{n^2}\operatorname{Tr}\big[R_0R_\theta^{-1}R_0R_\theta^{-1}\big]\right)}.$$

The eigenvalues of R_θ^{-1} are smaller than a finite constant $C_{\sup}$ for any n, X, θ from Lemma A.22. Thus, by applying Lemmas A.17, A.21 and A.25,

$$\mathbb{E}\big(\big|L_\theta - \mathbb{E}(L_\theta|X)\big|\big) \le \sqrt{\frac{2}{n}}\sqrt{\mathbb{E}\big(\big|R_0R_\theta^{-1}\big|^2\big)} \le \frac{C_{\sup}}{\sqrt{n}}. \qquad \square$$

Proof of Lemma A.2. We have, using the definition of L_θ in (1) and Lemma A.18,

$$\begin{aligned}
\mathbb{E}\Big(\sup_{\theta\in\Theta}\Big|\frac{\partial}{\partial\theta_i}L_\theta\Big|\Big) &= \mathbb{E}\Big(\sup_{\theta\in\Theta}\Big|\frac{1}{n}\operatorname{Tr}\Big(R_\theta^{-1}\frac{\partial}{\partial\theta_i}R_\theta\Big) \\
&\quad - \frac{1}{n}y^t R_\theta^{-1}\Big(\frac{\partial}{\partial\theta_i}R_\theta\Big)R_\theta^{-1}y\Big|\Big) \\
(\text{Lemma A.17 and Cauchy–Schwarz:}) &\leq \mathbb{E}\Big(\sup_{\theta\in\Theta}\sqrt{|R_\theta^{-1}|^2}\sqrt{\Big|\frac{\partial}{\partial\theta_i}R_\theta\Big|^2}\Big) \\
&\quad + \frac{1}{n}\mathbb{E}\Big(\sup_{\theta\in\Theta}\|R_\theta^{-1}y\|\Big\|\Big(\frac{\partial}{\partial\theta_i}R_\theta\Big)R_\theta^{-1}y\Big\|\Big) \\
&\leq \mathbb{E}\Big(\sup_{\theta\in\Theta}\sqrt{|R_\theta^{-1}|^2}\sqrt{\Big|\frac{\partial}{\partial\theta_i}R_\theta\Big|^2}\Big) \\
&\quad + \frac{1}{n}\mathbb{E}\Big(\Big[\sup_{\theta\in\Theta}\|R_\theta^{-1}y\|\Big]\Big[\sup_{\theta\in\Theta}\Big\|\Big(\frac{\partial}{\partial\theta_i}R_\theta\Big)R_\theta^{-1}y\Big\|\Big]\Big) \\
(\text{Cauchy–Schwarz:}) &\leq \mathbb{E}\Big(\sup_{\theta\in\Theta}\sqrt{|R_\theta^{-1}|^2}\sqrt{\Big|\frac{\partial}{\partial\theta_i}R_\theta\Big|^2}\Big) \qquad (9) \\
&\quad + \sqrt{\frac{1}{n}\mathbb{E}\Big(\sup_{\theta\in\Theta}\|R_\theta^{-1}y\|^2\Big)} \qquad (10) \\
&\quad \times \sqrt{\frac{1}{n}\mathbb{E}\Big(\sup_{\theta\in\Theta}\Big\|\Big(\frac{\partial}{\partial\theta_i}R_\theta\Big)R_\theta^{-1}y\Big\|^2\Big)}. \qquad (11)
\end{aligned}$$

Now, $|R_\theta^{-1}|^2 \leq \|R_\theta^{-1}\|^2$ because of (2.19) in [20] and $\|R_\theta^{-1}\|^2$ is bounded uniformly in θ because of Lemma A.22. Also, $\mathbb{E}(\sup_{\theta\in\Theta}|[\partial/\partial\theta_i]R_\theta|^2)$ is bounded because of Condition 3.2 and of a simple case of Lemma A.21. So the right-hand side of (9) is bounded because of Jensen inequality. It remains to show that the product of the terms (10) and (11) is bounded. To show this, note first that

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\Big(\sup_{\theta\in\Theta}\|R_\theta^{-1}y\|^2\Big) &\leq \frac{1}{n}\mathbb{E}\Big(\sup_{\theta\in\Theta}\|y\|^2\|R_\theta^{-1}\|^2\Big) \\
(\text{Lemma A.22:}) &\leq \frac{C_{\sup}}{n}\mathbb{E}\big(\|y\|^2\big) \\
&= C_{\sup}\big(K_0(0)+\delta_0\big),
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}\biggl(\sup_{\theta\in\Theta}\biggl\|\biggl(\frac{\partial}{\partial\theta_i}R_\theta\biggr)R_\theta^{-1}y\biggr\|^2\biggr)\\
&\quad=\frac{1}{n}\mathbb{E}\biggl(\sup_{\theta\in\Theta}y^tR_\theta^{-1}\biggl(\frac{\partial}{\partial\theta_i}R_\theta\biggr)^2R_\theta^{-1}y\biggr)\\
&\quad\le C_{\mathrm{sup}}\sum_{i_1+\cdots+i_p\le p}\int_\Theta\frac{1}{n}\mathbb{E}\biggl(\biggl|\frac{\partial^{i_1}}{\partial\theta_1^{i_1}}\cdots\frac{\partial^{i_p}}{\partial\theta_p^{i_p}}\biggl[y^tR_\theta^{-1}\biggl(\frac{\partial}{\partial\theta_i}R_\theta\biggr)^2R_\theta^{-1}y\biggr]\biggr|\biggr)\,d\theta \qquad \text{(Lemma A.19).}
\end{aligned}$$

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

$$\frac{1}{n}\sup_{\theta\in\Theta}\mathbb{E}\biggl(\biggl|\frac{\partial^{i_1}}{\partial\theta_1^{i_1}}\cdots\frac{\partial^{i_p}}{\partial\theta_p^{i_p}}\biggl[y^tR_\theta^{-1}\biggl(\frac{\partial}{\partial\theta_i}R_\theta\biggr)^2R_\theta^{-1}y\biggr]\biggr|\biggr)$$

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

$$\begin{aligned}
&\sup_{\theta\in\Theta}\frac{1}{n}\mathbb{E}\bigl(\bigl|y^tM_\theta y\bigr|\bigr)\\
&\quad\le\sup_{\theta\in\Theta}\frac{1}{n}\mathbb{E}\bigl(\bigl|y^tM_\theta y-\mathbb{E}\bigl(y^tM_\theta y|X\bigr)\bigr|\bigr)+\sup_{\theta\in\Theta}\frac{1}{n}\mathbb{E}\bigl(\bigl|\mathbb{E}\bigl(y^tM_\theta y|X\bigr)\bigr|\bigr)\\
&\quad\le\sup_{\theta\in\Theta}\sqrt{\mathbb{E}\biggl(\mathrm{var}\biggl[\frac{1}{n}y^tM_\theta y\bigg|X\biggr]\biggr)}+\sup_{\theta\in\Theta}\frac{1}{n}\mathbb{E}\bigl(\bigl|\mathbb{E}\bigl(y^tM_\theta y|X\bigr)\bigr|\bigr) \qquad \text{(Jensen inequality)}\\
&\quad=\sup_{\theta\in\Theta}\sqrt{\mathbb{E}\biggl(\frac{1}{2n^2}\mathrm{Tr}\bigl[R_0\bigl\{M_\theta+M_\theta^t\bigr\}R_0\bigl\{M_\theta+M_\theta^t\bigr\}\bigr]\biggr)}+\sup_{\theta\in\Theta}\frac{1}{n}\mathbb{E}\bigl(\bigl|\mathrm{Tr}[R_0M_\theta]\bigr|\bigr)\\
&\qquad\qquad \text{(using Lemma A.16)}\\
&\quad\le\sup_{\theta\in\Theta}\sqrt{\frac{1}{2n}\mathbb{E}\bigl(\bigl|R_0\bigl\{M_\theta+M_\theta^t\bigr\}\bigr|^2\bigr)}+\sup_{\theta\in\Theta}\sqrt{\mathbb{E}\bigl(|R_0|^2\bigr)\mathbb{E}\bigl(|M_\theta|^2\bigr)} \qquad \text{(Lemma A.17)}\\
&\quad\le\sup_{\theta\in\Theta}\sqrt{\frac{1}{n}\mathbb{E}\bigl(|R_0M_\theta|^2+\bigl|R_0M_\theta^t\bigr|^2\bigr)}+\sup_{\theta\in\Theta}\sqrt{\mathbb{E}\bigl(|R_0|^2\bigr)\mathbb{E}\bigl(|M_\theta|^2\bigr)}.
\end{aligned}$$

In the display above, the first term goes to 0 because of Conditions 3.1 and 3.2 and Lemmas A.21 and A.22. The second term is bounded because of Lemmas A.21, A.22 and A.25. This completes the proof. □

Proof of Corollary A.3. The corollary is a consequence of Lemma A.2 and of the fact that, for fixed n, we have $(\partial/\partial\theta_i)\mathbb{E}(L_\theta|X) = \mathbb{E}((\partial/\partial\theta_i)L_\theta|X)$ and of $\sup_\theta|\mathbb{E}(\cdot)| \le \mathbb{E}(\sup_\theta|\cdot|)$. □

A.4. Outline of proof for Theorem 3.4

We have,

$$\begin{aligned}
E_{n,\hat{\theta}_{\mathrm{CV}}} - E_{n,\theta} &= \mathrm{CV}_{\hat{\theta}_{\mathrm{CV}}} + E_{n,\hat{\theta}_{\mathrm{CV}}} - \mathrm{CV}_{\hat{\theta}_{\mathrm{CV}}} - \mathrm{CV}_\theta - E_{n,\theta} + \mathrm{CV}_\theta \\
&= \mathrm{CV}_{\hat{\theta}_{\mathrm{CV}}} + E_{n,\hat{\theta}_{\mathrm{CV}}} - \mathrm{CV}_{\hat{\theta}_{\mathrm{CV}}} + \delta_0 - \mathrm{CV}_\theta - E_{n,\theta} - \delta_0 + \mathrm{CV}_\theta \\
&\le \mathrm{CV}_{\hat{\theta}_{\mathrm{CV}}} - \mathrm{CV}_\theta + 2\sup_{\theta\in\Theta}|\mathrm{CV}_\theta - \delta_0 - E_{n,\theta}| \\
&\le 2\sup_{\theta\in\Theta}|\mathrm{CV}_\theta - \delta_0 - E_{n,\theta}|,
\end{aligned}$$

where the last inequality holds since $\hat{\theta}_{\mathrm{CV}}$ minimizes CV_θ.

In order to prove Theorem 3.4, it is thus sufficient to show that

$$\sup_{\theta\in\Theta}|\mathrm{CV}_\theta - \delta_0 - E_{n,\theta}| = o_p(1).$$

We first show in Lemma A.4 (proved in Section A.5) that, for fixed θ, the random variable inside the supremum above has almost zero mean value.

Lemma A.4. *Consider a fixed $\theta \in \Theta$. Then*

$$\mathbb{E}(\mathrm{CV}_\theta) - \mathbb{E}(E_{n,\theta}) - \delta_0$$

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

As in the proof of Theorem 3.3, we aim at obtaining a uniform-in-θ convergence by controlling the derivatives w.r.t. θ of $\mathbb{E}(\mathrm{CV}_\theta) - \mathbb{E}(E_{n,\theta}) - \delta_0$. All the results concerning the control of derivatives w.r.t. θ, that will be used throughout the proof of Theorem 3.4, are gathered in Lemma A.5 and Corollaries A.6 and A.7 below (these results are proved in Section A.5).

Lemma A.5. *For $i = 1, \ldots, p$,*

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

is bounded w.r.t. n.

Corollary A.6. *For any $i = 1, \ldots, p$,*

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

are bounded w.r.t. n.

Corollary A.7. *For any $i = 1, \dots, p$,*

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

are bounded w.r.t. n.

Using Lemma A.4 and Corollaries A.6 and A.7, together with the fact that simple convergence of a sequence of functions with uniformly bounded Lipschitz norms on a compact set implies uniform convergence, we have

$$\sup_{\theta\in\Theta}\left|\mathbb{E}(\mathrm{CV}_\theta) - \mathbb{E}(E_{n,\theta}) - \delta_0\right| = o(1).$$

Hence, we have

$$\begin{aligned}\sup_{\theta\in\Theta}|\mathrm{CV}_\theta - \delta_0 - E_{n,\theta}| &= \sup_{\theta\in\Theta}\left|\mathrm{CV}_\theta - \delta_0 - E_{n,\theta} - \left(\mathbb{E}(\mathrm{CV}_\theta) - \delta_0 - \mathbb{E}(E_{n,\theta})\right)\right| + o(1)\\ &\le \sup_{\theta\in\Theta}\left|\mathrm{CV}_\theta - \mathbb{E}(\mathrm{CV}_\theta)\right| + \sup_{\theta\in\Theta}\left|E_{n,\theta} - \mathbb{E}(E_{n,\theta})\right| + o(1).\end{aligned}$$

Hence, in order to prove Theorem 3.4, it is now sufficient to prove

$$\sup_{\theta\in\Theta}\left|\mathrm{CV}_\theta - \mathbb{E}(\mathrm{CV}_\theta)\right| = o_p(1) \tag{12}$$

and

$$\sup_{\theta\in\Theta}\left|E_{n,\theta} - \mathbb{E}(E_{n,\theta})\right| = o_p(1). \tag{13}$$

We first address (12). For a fixed θ, the quantity CV_θ has, so to speak, two sources of randomness: X and Y, ε. It proves useful to address these two sources of randomness separately, by writing

$$\sup_{\theta\in\Theta}\left|\mathrm{CV}_\theta - \mathbb{E}(\mathrm{CV}_\theta)\right| \le \sup_{\theta\in\Theta}\left|\mathrm{CV}_\theta - \mathbb{E}(\mathrm{CV}_\theta|X)\right| + \sup_{\theta\in\Theta}\left|\mathbb{E}(\mathrm{CV}_\theta|X) - \mathbb{E}(\mathrm{CV}_\theta)\right|. \tag{14}$$

The first supremum in the right-hand side in (14) is treated by Lemma A.8 (proved in Section A.5).

Lemma A.8. *For any fixed $\theta \in \Theta$, we have*

$$\mathbb{E}\left(\left|\mathrm{CV}_\theta - \mathbb{E}(\mathrm{CV}_\theta|X)\right|\right) \underset{n\to\infty}{\to} 0.$$

Using Lemma A.8, Lemma A.5 and Corollary A.6, together with Lemma A.15, we obtain

$$\sup_{\theta\in\Theta}\left|\mathrm{CV}_\theta - \mathbb{E}(\mathrm{CV}_\theta|X)\right| = o_p(1). \tag{15}$$

We now aim at showing that, for any fixed $\theta \in \Theta$, $\mathbb{E}(\mathrm{CV}_\theta|X) - \mathbb{E}(\mathrm{CV}_\theta) = o_p(1)$. In other words, we aim at controlling the fluctuations of the random variable $\mathbb{E}(\mathrm{CV}_\theta|X)$. In order to control these fluctuations, we show that $\mathbb{E}(\mathrm{CV}_\theta|X)$ can be well approximated by a quantity of the form $\mathbb{E}(\tilde{\mathrm{CV}}_\theta|X)$, which itself is more amenable to an asymptotic analysis. The precise definition of $\mathbb{E}(\tilde{\mathrm{CV}}_\theta|X)$ is given in Definition A.14 in Section A.5.

To summarize, in Definition A.14, for each $n \in \mathbb{N}$, we define n_2 subsets $C_1, \dots, C_{n_2}$ that constitute a partition of $[0, n^{1/d}]^d$ (n_2 depends on n but we do not write the dependence explicitly for concision). Then, for $\theta \in \Theta$, we define the non-stationary covariance function $\tilde{K}_\theta(t_1, t_2)$ by $\tilde{K}_\theta(t_1, t_2) = K_\theta(t_1, t_2)$ is t_1 and t_2 belong to the same subset C_j for some $j = 1, \dots, n_2$, and by $\tilde{K}_\theta(t_1, t_2) = 0$ if $t_1 \in C_j$ and $t_2 \in C_k$ with $j \neq K$. This new covariance function $\tilde{K}_\theta$ enables to define $\tilde{\mathrm{CV}}_\theta$ and $\tilde{E}_{n,\theta}$, in the same way as CV_θ and $E_{n,\theta}$ are defined from K_θ.

The benefit of the covariance function $\tilde{K}_\theta$ is that it yields a block-diagonal covariance matrix $\tilde{R}_\theta = (\tilde{K}_\theta(X_i - X_j))_{i,j=1,\dots,n}$ (the blocks correspond to observation points in the same subset C_j). Thus, the matrix $\tilde{R}_\theta^{-1}(\mathrm{diag}(\tilde{R}_\theta^{-1}))^{-2}\tilde{R}_\theta^{-1}$, corresponding for $\tilde{\mathrm{CV}}_\theta$ to the matrix in (3) is block diagonal, so that $\tilde{\mathrm{CV}}_\theta$ can be written as $\frac{1}{n_2}\sum_{i=1,\dots,n_2} V_i$, where the n_2 random variables $\mathbb{E}(V_1|X), \dots, \mathbb{E}(V_{n_2}|X)$ can be shown to be weakly dependent. This writing of $\mathbb{E}(\tilde{\mathrm{CV}}_\theta|X)$ as an average helps us showing that $\mathbb{E}(\tilde{\mathrm{CV}}_\theta|X)$ is concentrated around its mean. Note that this construction of $\tilde{K}_\theta$ is similar to that used in the proof of Proposition D.7 in [7]. We primary use Definition A.14 to prove Lemmas A.10, A.12 and A.13 below.

We first show that $\tilde{\mathrm{CV}}_\theta$ provides a good approximation of CV_θ in Lemma A.9 (proved in Section A.5).

Lemma A.9. *Consider a fixed* $\theta \in \Theta$. *In the context of Definition* A.14, *if* $n_2 = o(n)$,

$$\mathbb{E}\big(|\mathrm{CV}_\theta - \tilde{\mathrm{CV}}_\theta|\big) \underset{n\to\infty}{\to} 0.$$

Lemma A.9 directly implies that also $\mathbb{E}(|\mathbb{E}(\mathrm{CV}_\theta|X) - \mathbb{E}(\tilde{\mathrm{CV}}_\theta|X)|) \to_{n\to\infty} 0$. We use this fact to prove Lemma A.10, where the proof, provided in Section A.5, is based on showing that $\mathrm{var}[\mathbb{E}(\tilde{\mathrm{CV}}_\theta|X)] \to_{n\to\infty} 0$.

Lemma A.10. *For any fixed* $\theta \in \Theta$,

$$\mathbb{E}\big(\big|\mathbb{E}(\mathrm{CV}_\theta) - \mathbb{E}[\mathrm{CV}_\theta|X]\big|\big)$$

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

Using Lemma A.10 and Corollary A.6, together with Lemma A.15, we obtain

$$\sup_{\theta\in\Theta}\big|\mathbb{E}(\mathrm{CV}_\theta|X) - \mathbb{E}(\mathrm{CV}_\theta)\big| = o_p(1).$$

Hence, by recalling (14) and (15), we prove (12).

We now address (13). Similarly as for CV_θ, we write

$$\sup_{\theta\in\Theta}\big|E_{n,\theta} - \mathbb{E}(E_{n,\theta})\big| \le \sup_{\theta\in\Theta}\big|E_{n,\theta} - \mathbb{E}(E_{n,\theta}|X)\big| + \sup_{\theta\in\Theta}\big|\mathbb{E}(E_{n,\theta}|X) - \mathbb{E}(E_{n,\theta})\big| \tag{16}$$

and treat the two terms in the right-hand side of (16) separately. We first recall that the definition of $\tilde{K}_\theta$ (see Definition A.14) provides a corresponding integrated square prediction error $\tilde{E}_{n,\theta}$. We first show in Lemma A.11 (proved in Section A.5) that $\tilde{E}_{n,\theta}$ provides a good approximation of $E_{n,\theta}$.

Lemma A.11. *Let, with the notation of Definition A.14, $\tilde{E}_{n,\theta}$ be defined as $E_{n,\theta}$, with K_θ replaced by $\tilde{K}_\theta$. Fix $\theta \in \Theta$. Then, if $n_2 = o(n)$,*

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

Lemma A.11 enables to prove Lemma A.12 below; the proof of Lemma A.12 is given in Section A.5, and is based on showing that $\mathbb{E}(\mathrm{var}(\tilde{E}_{n,\theta}|X)) \to_{n\to\infty} 0$.

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

$$\mathbb{E}\big(\big|E_{n,\theta} - \mathbb{E}(E_{n,\theta}|X)\big|\big) \underset{n\to\infty}{\to} 0.$$

Similarly, we prove Lemma A.13 by using Lemma A.11 and by showing that $\mathrm{var}(\mathbb{E}(\tilde{E}_{n,\theta}|X)) \to_{n\to\infty} 0$ (see the proof in Section A.5).

Lemma A.13. *For any fixed $\theta \in \Theta$,*

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

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

From Lemmas A.12 and A.13, from Corollary A.7, and from Lemma A.15, we have that the two terms in the right-hand side of (16) go to zero in probability as $n \to \infty$. Hence (13) is proved, and since (12) is also proved, the proof of Theorem 3.4 is complete.

A.5. Complement of proof for Theorem 3.4

Proof of Lemma A.4. Recall the expression

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

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

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

where we remind that $\hat{y}_\theta(X_{n+1}) = r_\theta^t(X_{n+1}) R_\theta^{-1} y$, with $r_\theta(X_{n+1}) = (K_\theta(X_1, X_{n+1}), \dots, K_\theta(X_n, X_{n+1}))^t$.

Let us also consider a Gaussian variable ε_{n+1} with mean 0 and variance δ_0, and so that X_{n+1} and ε_{n+1} are independent of X, Y and ε. By symmetry of the roles of $X_1, \dots, X_{n+1}$ and $\varepsilon_1, \dots \varepsilon_{n+1}$, we have

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

with $\hat{y}_{n-1,\theta}(X_{n+1}) = \check{r}^t_{n-1,\theta}\check{R}^{-1}_{n-1,\theta}y$, with $\check{r}_{n-1,\theta} = (K(X_1, X_{n+1}), \dots, K(X_{n-1}, X_{n+1}), 0)^t$ and

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

Hence,

$$\begin{aligned}
&\big|\mathbb{E}(\mathrm{CV}_\theta) - \mathbb{E}(E_{n,\theta}) - \delta_0\big| \\
&\quad = \big|\mathbb{E}\big(\big[Y(X_{n+1}) - \hat{y}_{n-1,\theta}(X_{n+1})\big]^2 - \big[Y(X_{n+1}) - \hat{y}_\theta(X_{n+1})\big]^2\big)\big| \\
&\quad = \big|\mathbb{E}\big(\big[\hat{y}_\theta(X_{n+1}) - \hat{y}_{n-1,\theta}(X_{n+1})\big]\big[2Y(X_{n+1}) - \hat{y}_{n-1,\theta}(X_{n+1}) - \hat{y}_\theta(X_{n+1})\big]\big)\big| \\
&\quad \le \sqrt{\mathbb{E}\big(\big[r^t_\theta(X_{n+1})R^{-1}_\theta y - \check{r}^t_{n-1,\theta}\check{R}^{-1}_{n-1,\theta}y\big]^2\big)} \qquad (17) \\
&\qquad \times \sqrt{\mathbb{E}\big(\big[2Y(X_{n+1}) - \check{r}^t_{n-1,\theta}\check{R}^{-1}_{n-1,\theta}y - r^t_\theta(X_{n+1})R^{-1}_\theta y\big]^2\big)},
\end{aligned}$$

where the last inequality is obtained by applying the Cauchy–Schwarz inequality. Using $(a + b + c)^2 \le 3(a^2 + b^2 + c^2)$, the mean value inside the second square root in (17) is smaller than

$$3\big(\mathbb{E}\big(\big[2Y(X_{n+1})\big]^2\big) + \mathbb{E}\big(\big[\check{r}^t_{n-1,\theta}\check{R}^{-1}_{n-1,\theta}y\big]^2\big) + \mathbb{E}\big(\big[r^t_\theta(X_{n+1})R^{-1}_\theta y\big]^2\big)\big).$$

In the above display, $\mathbb{E}([2Y(X_{n+1})]^2) = 4K_0(0)$ is bounded, and the two other summands can be shown to be bounded with techniques similar to but simpler than in the proof of Lemma A.5 below (see the proof of Lemma A.5 from the sequence of equations in (18) to the end).

Finally, we can show that the first square root in (17) goes to zero with techniques similar to, but simpler than, those used to prove (19) in the proof of Lemma A.9 below. This completes the proof. □

Proof of Lemma A.5. We have

$$\begin{aligned}
\mathbb{E}\bigg(\sup_{\theta\in\Theta}\bigg|\frac{\partial}{\partial\theta_i}\mathrm{CV}_\theta\bigg|\bigg) &\le \mathbb{E}\bigg(\frac{1}{n}\sum_{k=1}^n \sup_{\theta\in\Theta}\bigg|\frac{\partial}{\partial\theta_i}(y_k - \hat{y}_{k,\theta})^2\bigg|\bigg) \\
(\text{symmetry of } X_1, \dots, X_n\text{:}) &= \mathbb{E}\bigg(\sup_{\theta\in\Theta}\bigg|\frac{\partial}{\partial\theta_i}(y_1 - \hat{y}_{1,\theta})^2\bigg|\bigg) \\
(\text{Lemma A.19:}) &\le C_{\sup}\sum_{i_1+\cdots+i_p\le p}\int_\Theta \mathbb{E}\bigg(\bigg|\frac{\partial^{i_1}}{\partial\theta_1^{i_1}}\cdots\frac{\partial^{i_p}}{\partial\theta_p^{i_p}}\frac{\partial}{\partial\theta_i}(y_1 - \hat{y}_{1,\theta})^2\bigg|\bigg)\,d\theta.
\end{aligned}$$

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

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

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

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

or

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

Now, $\mathbb{E}((y_1 - \hat{y}_{1,\theta})^2) \le 2\mathbb{E}(y_1^2) + 2\mathbb{E}(\hat{y}_{1,\theta}^2)$. The term $\mathbb{E}(y_1^2)$ is bounded uniformly in θ. Thus, finally, it remains to show that for any $a_1 + \cdots + a_p \le p + 1$, $\sup_{\theta\in\Theta} \mathbb{E}(([\partial^{a_1}/\partial\theta_1^{a_1}] \cdots [\partial^{a_p}/\partial\theta_p^{a_p}]\hat{y}_{1,\theta})^2)$ is bounded. For that, we have $\hat{y}_{1,\theta} = r_{1,\theta}^t R_{1,\theta}^{-1} y_{-1}$. Thus, using Lemma A.18, $[\partial^{a_1}/\partial\theta_1^{a_1}] \cdots [\partial^{a_p}/\partial\theta_p^{a_p}]\hat{y}_{1,\theta}$ is a fixed sum of weighted terms of the form $w_\theta^t M_\theta y_{-1}$, where w_θ is of the form $[\partial^{b_1}/\partial\theta_1^{b_1}] \cdots [\partial^{b_p}/\partial\theta_p^{b_p}] r_{1,\theta}$ ($b_1 + \cdots + b_p \le p + 1$) and M_θ is of the form $R_{1,\theta}^{-1} M_{1,\theta} \cdots R_{1,\theta}^{-1} M_{k,\theta} R_{1,\theta}^{-1}$. Finally, k is smaller than a finite constant $C_{\sup}$ (function of p) and $M_{i,\theta}$ is of the form $[\partial^{c_1}/\partial\theta_1^{c_1}] \cdots [\partial^{c_p}/\partial\theta_p^{c_p}] R_{1,\theta}$, with $c_1 + \cdots + c_p \le p + 1$. Thus, it is sufficient to show that a generic $\sup_{\theta\in\Theta} \mathbb{E}((w_\theta^t M_\theta y_{-1})^2)$, as previously defined, is bounded.

Then,

$$\begin{aligned}
\sup_{\theta\in\Theta} \mathbb{E}\big((w_\theta^t M_\theta y_{-1})^2\big) &= \sup_{\theta\in\Theta} \mathbb{E}_X \mathbb{E}_{y|X}\big(y_{-1}^t M_\theta^t w_\theta w_\theta^t M_\theta y_{-1}\big) \\
&= \sup_{\theta\in\Theta} \mathbb{E}_X \operatorname{Tr}\big(R_{1,0} M_\theta^t w_\theta w_\theta^t M_\theta\big) \\
&\le \sup_{\theta\in\Theta} \mathbb{E}_X \left[\sum_{i,j=2}^{n} \big|(M_\theta R_{1,0} M_\theta^t)_{i,j}\big| \big|(w_\theta w_\theta^t)_{i,j}\big|\right] \\
&= \sup_{\theta\in\Theta} \left[\sum_{i,j=2}^{n} \mathbb{E}_{X_2,\dots,X_n}\Big(\big|(M_\theta R_{1,0} M_\theta^t)_{i,j}\big| \mathbb{E}_{X_1|X_2,\dots,X_n}\big|(w_\theta w_\theta^t)_{i,j}\big|\Big)\right].
\end{aligned} \tag{18}$$

Now, because of Conditions 2.4 and 3.2,

$$\mathbb{E}_{X_1|X_2,\dots,X_n}\left|\left(w_\theta w_\theta^t\right)_{i,j}\right| \le \frac{C_{\sup}}{n}\int_{[0,n^{1/d}]^d}\frac{1}{1+|X_i-x_1|^{d+1}}\frac{1}{1+|X_j-x_1|^{d+1}}\,dx_1$$

$$(\text{Lemma A.20:}) \le \frac{1}{n}\frac{C_{\sup}}{1+|X_i-X_j|^{d+1}}.$$

So,

$$\sup_{\theta\in\Theta}\mathbb{E}\left(\left(w_\theta^t M_\theta y_{-1}\right)^2\right) \le C_{\sup}\frac{1}{n}\sup_{\theta\in\Theta}\left[\sum_{i,j=2}^{n}\mathbb{E}_{X_2,\dots,X_n}\left(\left|\left(M_\theta R_{1,0}M_\theta^t\right)_{i,j}\right|\frac{1}{1+|X_i-X_j|^{d+1}}\right)\right]$$

$$(\text{Cauchy–Schwarz:}) \le \sup_{\theta\in\Theta}\left[\sqrt{\mathbb{E}\left\{\left|M_\theta R_{1,0}M_\theta^t\right|^2\right\}}\sqrt{\mathbb{E}\left\{\frac{1}{n}\sum_{i,j=2}^{n}\left(\frac{1}{1+|X_i-X_j|^{d+1}}\right)^2\right\}}\right].$$

The supremum over θ of the second term above is bounded because of Lemma A.21. The supremum over θ of the first term above is bounded because of Lemmas A.21, A.22 and A.25. □

Proof of Corollary A.6. The corollary is a consequence of Lemma A.5, $\sup_\theta|\mathbb{E}(\cdot)| \le \mathbb{E}(\sup_\theta|\cdot|)$ and of the fact that, for fixed n, we have $(\partial/\partial\theta_i)\mathbb{E}(\mathrm{CV}_\theta|X) = \mathbb{E}((\partial/\partial\theta_i)\mathrm{CV}_\theta|X)$ and $(\partial/\partial\theta_i)\mathbb{E}(\mathrm{CV}_\theta) = \mathbb{E}((\partial/\partial\theta_i)\mathrm{CV}_\theta)$. □

Proof of Corollary A.7. We have

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

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

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

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

Hence, by considering t as a new random observation point X_{n+1} as in the proof of Lemma A.4, we show the first bound of the lemma as in the proof of Lemma A.5, 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.6. □

Proof of Lemma A.8. From (3), we have $\mathrm{CV}_\theta = y^t M_\theta y$, with $M_\theta = (1/n)R_\theta^{-1}\operatorname{diag}(R_\theta^{-1})^{-2}\times R_\theta^{-1}$. Because of Lemma A.22, 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.1, with R_θ^{-1} replaced by M_θ. □

Definition A.14. 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 $i = 1, \dots, N_2 - 1$, $c_i = [((i-1)/N_2)n^{1/d}, (i/N_2)n^{1/d}]$. 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 $i \in \{1, \dots, N_2\}$ so that $x \in c_i$. Let, for $t = (t_1, \dots, t_d)^t \in [0, n^{1/d}]^d$, $C(t) = \prod_{j=1}^d c_{i(t_j)}$. 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 $\tilde{R}_\theta$, $\tilde{R}_{i,\theta}$, $\tilde{r}_{i,\theta}$, $\tilde{\hat{y}}_{i,\theta}$, $\widetilde{\mathrm{CV}}_\theta$ 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 $i_1, \dots, i_d \in \{1, \dots, N_2\}$, as the sets $C_1, \dots, C_{n_2}$. [The specific one-to-one correspondence we use between $\{1, \dots, N_2\}^d$ and $\{1, \dots, n_2\}$ is of no interest. Note that this one-to-one correspondence depends on n. The sets $C_1, \dots, C_{n_2}$ 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 indices). 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.14, 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.]

Proof of Lemma A.9. Assume that $n_2 = o(n)$, or equivalently that $\Delta \to_{n \to \infty} \infty$. We have

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

Now, the second square root in the above display is bounded, because of $(a+b+c)^2 \leq 3(a^2 + b^2 + c^2)$ and of arguments similar to but simpler than those given in the proof of Lemma A.5.

Thus, it only remains to show that

$$\mathbb{E}\big(\big(\tilde{r}^t_{1,\theta}\tilde{R}^{-1}_{1,\theta}y_{-1} - r^t_{1,\theta}R^{-1}_{1,\theta}y_{-1}\big)^2\big) \underset{n\to\infty}{\to} 0. \tag{19}$$

For this,

$$\begin{aligned}\mathbb{E}\big(\big(\tilde{r}^t_{1,\theta}\tilde{R}^{-1}_{1,\theta}y_{-1} - r^t_{1,\theta}R^{-1}_{1,\theta}y_{-1}\big)^2\big) &\le 2\mathbb{E}\big(\big(\tilde{r}^t_{1,\theta}\big(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta}\big)y_{-1}\big)^2\big) \\ &\quad + 2\mathbb{E}\big(\big((\tilde{r}_{1,\theta} - r_{1,\theta})^t R^{-1}_{1,\theta}y_{-1}\big)^2\big).\end{aligned} \tag{20}$$

We show separately that both terms in the right-hand side of (20) converge to 0. For the first term,

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

Hence, by the same arguments as after (18) in the proof of Lemma A.5, we obtain

$$\begin{aligned}\big[\mathbb{E}\big(\big(\tilde{r}^t_{1,\theta}\big(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta}\big)y_{-1}\big)^2\big)\big]^2 &\le C_{\sup}\mathbb{E}\big(\big|\big(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta}\big)R_{1,0}\big(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta}\big)\big|^2\big) \\ &\le C_{\sup}\mathbb{E}\big(\big\{\big\|\tilde{R}^{-1}_{1,\theta}\big\| + \big\|R^{-1}_{1,\theta}\big\|\big\}\big|R_{1,0}\big(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta}\big)\big|^2\big) \\ \text{(Lemmas A.22 and A.23:)} &\le \frac{C_{\sup}}{n}\mathbb{E}\big(\mathrm{Tr}\big[\big(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta}\big)R^2_{1,0}\big(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta}\big)\big]\big) \\ \text{(Lemma A.17:)} &\le C_{\sup}\sqrt{\mathbb{E}\big(\big|\big(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta}\big)^2\big|^2\big)}\sqrt{\mathbb{E}\big(\big|R^2_{1,0}\big|^2\big)}.\end{aligned}$$

From Lemma A.21, $\mathbb{E}(|R^2_{1,0}|^2)$ is bounded, so it remains to show that $\mathbb{E}(|(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta})^2|^2)$ converges to 0. For this,

$$\begin{aligned}\mathbb{E}\big(\big|\big(\tilde{R}^{-1}_{1,\theta} - R^{-1}_{1,\theta}\big)^2\big|^2\big) &= \mathbb{E}\big(\big|\big(\tilde{R}^{-1}_{1,\theta}(R_{1,\theta} - \tilde{R}_{1,\theta})R^{-1}_{1,\theta}\big)^2\big|^2\big) \\ \text{(Lemma A.22:)} &\le C_{\sup}\mathbb{E}\big(\big|(R_{1,\theta} - \tilde{R}_{1,\theta})R^{-1}_{1,\theta}\tilde{R}^{-1}_{1,\theta}(R_{1,\theta} - \tilde{R}_{1,\theta})\big|^2\big) \\ &= C_{\sup}\frac{1}{n}\mathbb{E}\big(\mathrm{Tr}\big[(R_{1,\theta} - \tilde{R}_{1,\theta})^2R^{-1}_{1,\theta}\tilde{R}^{-1}_{1,\theta}(R_{1,\theta} - \tilde{R}_{1,\theta})^2\tilde{R}^{-1}_{1,\theta}R^{-1}_{1,\theta}\big]\big) \\ \text{(Lemma A.17:)} &\le C_{\sup}\sqrt{\mathbb{E}\big(\big|(R_{1,\theta} - \tilde{R}_{1,\theta})^2R^{-1}_{1,\theta}\tilde{R}^{-1}_{1,\theta}\big|^2\big)}\sqrt{\mathbb{E}\big(\big|(R_{1,\theta} - \tilde{R}_{1,\theta})^2\tilde{R}^{-1}_{1,\theta}R^{-1}_{1,\theta}\big|^2\big)}.\end{aligned}$$

Hence, with Lemmas A.22, A.23 and A.26, we conclude that the first term of the right-hand side of (20) goes to 0. Let us now show that the second term of the right-hand side of (20) goes to 0.

We have,

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

where the last line is obtained similarly to after (18) in the proof of Lemma A.5. Thus we have, with the notation and result of Lemma A.27,

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

From Lemmas A.21, A.22 and A.25, the first $\sqrt{\cdot}$ in the above display is bounded. Thus it remains to show that the second $\sqrt{\cdot}$ goes to 0. For this, noting that $f^2(t) \le C_{\sup} f(t)$ and distinguishing the case $i = j$ from the case $i \neq j$,

$$
\begin{aligned}
&\frac{1}{n} \sum_{i,j=2}^{n} \mathbb{E}\bigg[\bigg(\frac{1}{1+|X_i - X_j|^{d+1}}\bigg)^2 f\big(D_\Delta(X_i, X_j)\big)\bigg] \\
&\quad \le \frac{C_{\sup}}{n} \int_{[0,n^{1/d}]^d} f\big(D_\Delta(x)\big)\, dx \\
&\qquad + \frac{C_{\sup}}{n} \int_{[0,n^{1/d}]^d} dx_1 \int_{[0,n^{1/d}]^d} dx_2 \frac{1}{1+|x_1 - x_2|^{d+1}} f\big(D_\Delta(x_1, x_2)\big) \\
&\quad = \frac{C_{\sup}}{n} \int_{[0,n^{1/d}]^d} f\big(D_\Delta(x)\big)\, dx + o(1) \qquad \text{(Lemma A.28).} \qquad (21)
\end{aligned}
$$

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

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

This finally shows that the second term of the right-hand side of (20) goes to 0 which finishes the proof. □

Proof of Lemma A.10. Fix $\theta \in \Theta$. Because of Lemma A.9 and of $|\mathbb{E}(\cdot)| \le \mathbb{E}(|\cdot|)$, it is sufficient to show that there exists a sequence $\Delta \to +\infty$ so that the lemma holds with CV_θ replaced by $\widetilde{\mathrm{CV}}_\theta$. Then, because of $(\mathbb{E}(\cdot))^2 \le \mathbb{E}((\cdot)^2)$, it is sufficient to show $\mathrm{var}(\mathbb{E}[\widetilde{\mathrm{CV}}_\theta|X]) \to_{n\to\infty} 0$.

Let $C_1, \dots, C_{n_2}$ be as in Definition A.14. Define, for $k = 1, \dots, n_2$,

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

[Note that, following the discussion in Definition A.14, we have $f_k(X) = 0$ if $N_k = 0$ and $f_k(X) = K_0(0) + \delta_0$ if $N_k = 1$.] Then $\mathbb{E}(\widetilde{\mathrm{CV}}_\theta|X) = (1/n_2)\sum_{k=1}^{n_2} f_k(X)$. Let $\bar{R}_{k,\theta}$ and $\bar{R}_{k,0}$ be as in Definition A.14. Because of the definition of $\tilde{K}$ and by (3), we have

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

The functions $f_k(X)$ satisfy the conditions of Lemma A.29. Furthermore, by using the notation N_k of Lemma A.29, we have

$$\begin{aligned}
\mathbb{E}\bigl(f_k^2(X)|N_k = N\bigr) &= \frac{1}{\Delta^2}\mathbb{E}\bigl(\bigl[\operatorname{Tr}\bigl(\bar{R}_{k,0}\bar{R}_{k,\theta}^{-1}\operatorname{diag}\bigl(\bar{R}_{k,\theta}^{-1}\bigr)^{-2}\bar{R}_{k,\theta}^{-1}\bigr)\bigr]^2|N_k = N\bigr)\\
\text{(Lemma A.17:)} &\le \frac{1}{\Delta^2}\mathbb{E}\bigl(N^2|\bar{R}_{k,0}|^2\bigl|\bar{R}_{k,\theta}^{-1}\operatorname{diag}\bigl(\bar{R}_{k,\theta}^{-1}\bigr)^{-2}\bar{R}_{k,\theta}^{-1}\bigr|^2|N_k = N\bigr)\\
\text{(Lemma A.24:)} &\le C_{\sup}\frac{N^2}{\Delta^2}\mathbb{E}\bigl(|\bar{R}_{k,0}|^2|N_k = N\bigr)\\
&\text{(Condition 3.1 and Lemma A.29:)}\\
&\le C_{\sup}\frac{N^2}{\Delta^2}\biggl(1 + \frac{N}{\Delta^2}\int_{[0,\Delta^{1/d}]^d}\int_{[0,\Delta^{1/d}]^d}\frac{1}{1+|x_1 - x_2|^{d+1}}\,dx_1\,dx_2\biggr)\\
&\le C_{\sup}\biggl(\frac{N^2}{\Delta^2} + \frac{N^3}{\Delta^3}\biggr)\\
&\le C_{\sup}\biggl(1 + \frac{N^4}{\Delta^4}\biggr).
\end{aligned}$$

Thus, because of Lemma A.30, there exists a sequence $\Delta \to_{n\to\infty} \infty$ so that $\mathrm{var}(\mathbb{E}[\widetilde{\mathrm{CV}}_\theta|X]) \to_{n\to\infty} 0$, which completes the proof. □

Proof of Lemma A.11. We have, by letting $\tilde{\hat{y}}_\theta(t)$ be as $\hat{y}_\theta(t)$, with K_θ replaced by $\tilde{K}_\theta$.

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

As for the proof of Lemma A.4, the variable t in the integral above is formally equivalent to a new observation point X_{n+1}, so that $X_1, \dots, X_{n+1}$ are independent and uniformly distributed on $[0, n^{1/d}]^d$. Thus,

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

The rest of the proof is carried out as in the proof of Lemma A.9, the only difference being that there are $n+1$ observation points instead of n. □

Proof of Lemma A.12. Because of Lemma A.11 and using $|\mathbb{E}(\cdot)| \le \mathbb{E}(|\cdot|)$ and $\mathbb{E}^2(\cdot) \le \mathbb{E}((\cdot)^2)$, it is sufficient to show that there exists a sequence $\Delta \to_{n\to\infty}$ so that

$$\mathbb{E}\big(\mathrm{var}(\tilde{E}_{n,\theta}|X)\big)$$

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

Let us use the notation $C_1, \dots, C_{n_2}$ of Definition A.14. Let, for $t \in \mathbb{R}^d$ and $v = (v_1, \dots, v_m) \in (\mathbb{R}^d)^m$, $r_\theta(t, v) = (K_\theta(t, v_1), \dots, K_\theta(t, v_m))^t$. We define $r_0(t, v)$ similarly. Let y^i, $\bar{R}_{i,\theta}$ and $\bar{R}_{i,0}$ be as in Definition A.14. Let for $i \neq j$, $R_0(X^i, X^j) = [K_0((X^i)_k, (X^j)_l)]_{k=1,\dots,N_i; l=1,\dots,N_j}$. Let $R_0(X^i, X^i) = [K_0((X^i)_k, (X^i)_l)]_{k,l=1,\dots,N_i} + \delta_0 I_{N_i}$. Then,

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

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

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

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

$$\mathbb{E}\big(\operatorname{var}(\tilde{E}_{n,\theta}|X)\big) \le C_{\sup}(T_1+T_2+T_3+T_4), \tag{22}$$

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,

$$\begin{aligned} T_1 &= \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i} dt_i \int_{C_j} dt_j K_0^2(t_i,t_j) \\ (\text{Condition 3.1:}) &\le \frac{C_{\sup}}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\frac{1}{\Delta^2}\int_{C_i} dt_i \int_{\mathbb{R}^d} dt \left(\frac{1}{1+|t_i-t|^{d+1}}\right)^2 \\ &\le C_{\sup}\frac{1}{n_2\Delta}. \end{aligned} \tag{23}$$

For T_2, using Cauchy–Schwarz and Lemma A.24,

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

Now, using the notation N_i of Lemma A.29 and Conditions 3.1 and 3.2,

$$\begin{aligned} T_2 &\le C_{\sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i} dt_i \int_{C_j} dt_j \mathbb{E}\left[N_i^2\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\}^4\right] \\ &\le C_{\sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\Delta^2\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\} \qquad (\text{Lemma A.31}) \\ &\le C_{\sup}\frac{\Delta^2}{n_2}\max_{i=1,\dots,n_2}\sum_{j=1}^{n_2}\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\} \\ &\le C_{\sup}\frac{\Delta^2}{n_2} \qquad \Big(\text{Lemma A.32, and because we will set } \Delta \underset{n\to\infty}{\to} \infty\Big). \end{aligned} \tag{24}$$

For T_4 in (22), using Cauchy–Schwarz and Lemma A.24,

$$\begin{aligned} T_4 &= \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i} dt_i \int_{C_j} dt_j \mathbb{E}\big[\big(r_\theta^t(t_i,X^i)\bar{R}_{i,\theta}^{-1}R_0(X^i,X^j)\bar{R}_{j,\theta}^{-1}r_\theta(t_j,X^j)\big)^2\big] \\ &\le C_{\sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i} dt_i \int_{C_j} dt_j \sqrt{\mathbb{E}\big[\big\|r_\theta^t(t_i,X^i)\big\|^4\big]} \\ &\quad\times\sqrt{\mathbb{E}\big[\big\|R_0(X^i,X^j)\bar{R}_{j,\theta}^{-1}r_\theta(t_j,X^j)\big\|^4\big]}. \end{aligned} \tag{25}$$

Using Condition 3.1, Lemma A.24 and Lemma A.33, we obtain

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

Hence, going back to (25),

$$\begin{aligned}T_4 &\le C_{\sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i} dt_i \int_{C_j} dt_j \sqrt{\mathbb{E}[N_i^2]}\left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\}^2 \sqrt{\mathbb{E}[N_i^2 N_j^4]} \\ &\le C_{\sup}\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{\Delta^2}\int_{C_i} dt_i \int_{C_j} dt_j \left\{\frac{1}{1+d(C_i,C_j)^{d+1}}\right\}^2 \\ &\quad\times \sqrt{\mathbb{E}[N_i^2]}\sqrt{\sqrt{\mathbb{E}[N_i^4]}\sqrt{\mathbb{E}[N_j^8]}} \\ &\le C_{\sup}\frac{\Delta^4}{n_2} \qquad \text{(Lemmas A.31 and A.32).}\end{aligned} \tag{26}$$

Recall that by definition $n = n_2\Delta$. Hence, we can set $\Delta = n^{1/6}$, so that $n_2 = n^{5/6}$, $1/(n_2\Delta) \to_{n\to\infty} 0$, $\Delta^2/n_2 = n^{-3/6} \to_{n\to\infty} 0$ and $\Delta^4/n_2 = n^{-1/6} \to_{n\to\infty} 0$. Hence, from (23), (24) and (26), the proof is completed. □

Proof of Lemma A.13. Fix $\theta \in \Theta$. Because of Lemma A.11 and of $|\mathbb{E}(\cdot)| \le \mathbb{E}(|\cdot|)$, 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 $(\mathbb{E}(\cdot))^2 \le \mathbb{E}((\cdot)^2)$, it is sufficient to show $\mathrm{var}(\mathbb{E}[\tilde{E}_{n,\theta}|X]) \to_{n\to\infty} 0$.

Let $C_1, \dots, C_{n_2}$ be as in Definition A.14 and let $\tilde{\hat{y}}_\theta(t)$ be as in the proof of Lemma A.11. Define, for $k = 1, \dots, n_2$,

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

[Note that, following the discussion in Definition A.14, we have $g_k(x) = K_0(0)$ if $N_k = 0$.] Then $\mathbb{E}(\tilde{E}_{n,\theta}|X) = (1/n_2)\sum_{k=1}^{n_2} g_k(X)$. Following the notation of Lemma A.12 we have,

$$\begin{aligned}g_k(X) &= \frac{1}{\Delta}\int_{C_k} dt_k \mathbb{E}([Y(t_k) - r_\theta^t(t_k, X^k)\bar{R}^{-1}_{k,\theta} y^k]^2|X) \\ &\le 2\frac{1}{\Delta}\int_{C_k} dt_k (K_0(0) + r_\theta^t(t_k, X^k)\bar{R}^{-1}_{k,\theta}\bar{R}_{k,0}\bar{R}^{-1}_{k,\theta} r_\theta(t_k, X^k)) \\ &\le C_{\sup} + 2\frac{1}{\Delta}\int_{C_k} dt_k \|r_\theta^t(t_k, X^k)\| \|\bar{R}_{k,0}\bar{R}^{-1}_{k,\theta} r_\theta(t_k, X^k)\| \qquad \text{(Lemma A.24)}\end{aligned}$$

$$\le C_{\sup} + C_{\sup}\frac{1}{\Delta}\int_{C_k} dt_k\sqrt{N_k}N_k\left\|\bar{R}_{k,\theta}^{-1}r_\theta\left(t_k, X^k\right)\right\|$$

(Conditions 3.1 and 3.2 and Lemma A.33)

$$\le C_{\sup}\left(1+N_k^2\right) \qquad \text{(Lemma A.24).}$$

Hence $\mathbb{E}(g_k^2(X)|N_k = N) \le C_{\sup}(1 + N^4)$, so that we can complete the proof with Lemma A.30. □

A.6. Technical results

The following technical results are proved in the supplementary material [8].

Lemma A.15. *Consider a function $f_\theta(X, y)$ that is continuously differentiable w.r.t. θ for any X, y. Assume that for all $\theta \in \Theta$, $f_\theta(X, y) = o_p(1)$ and that, for $i = 1, \dots, p$, $\sup_{\theta\in\Theta} |[\partial/(\partial\theta_i)]f_\theta(X, y)| = O_p(1)$. Then*

$$\sup_{\theta\in\Theta}\left|f_\theta(X, y)\right| = o_p(1).$$

Lemma A.16. *Let z be a $k \times 1$ random vector with mean vector 0 and covariance matrix Σ. Let A be a fixed $k \times k$ matrix. Then*

(i) $\mathbb{E}(z^t Az) = \mathrm{Tr}(A\Sigma)$.
(ii) *If, in addition, z is a Gaussian vector then we have* $\mathrm{var}(z^t Az) = 2\,\mathrm{Tr}(A\Sigma A\Sigma)$.

Lemma A.17. *Let A and B be two $m \times m$ matrices. Then, we have*

$$\frac{1}{m}\left|\mathrm{Tr}(AB)\right| \le \sqrt{|A|^2}\sqrt{|B|^2}.$$

In addition, when the matrices A and B are random, we have

$$\frac{1}{m}\mathbb{E}\left(\left|\mathrm{Tr}(AB)\right|\right) \le \sqrt{\mathbb{E}\left[|A|^2\right]}\sqrt{\mathbb{E}\left[|B|^2\right]}.$$

Lemma A.18. *Let M_θ be a $m \times m$ matrix that is a differentiable function of $\theta \in \Theta$ and that is invertible for all $\theta \in \Theta$. Then we have, for $i = 1, \dots, p$,*

$$\frac{\partial}{\partial\theta_i}\left[\log\left(\det(M_\theta)\right)\right] = \mathrm{Tr}\left(M_\theta^{-1}\left[\frac{\partial}{\partial\theta_i}M_\theta\right]\right)$$

and

$$\frac{\partial}{\partial\theta_i}\left[M_\theta^{-1}\right] = -M_\theta^{-1}\left[\frac{\partial}{\partial\theta_i}M_\theta\right]M_\theta^{-1}.$$

Lemma A.19. *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 $i_1 + \cdots + i_p \le p$,*

$$\sup_\theta \bigl| \bigl(\partial^{i_1}/\partial\theta_1^{i_1}\bigr) \cdots \bigl(\partial^{i_p}/\partial\theta_p^{i_p}\bigr) f_\theta(X, y) \bigr|$$

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

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

Lemma A.20. *There exists a finite constant $C_{\sup}$ so that, for any $a, b \in \mathbb{R}^d$,*

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

Lemma A.21. *Let $0 < C_{\inf} \le C_{\sup} < \infty$ be fixed independently of n. Let s_n be a function of n so that $s_n \in \mathbb{N}^*$ and $C_{\inf} n \le s_n \le C_{\sup} n$. Consider s_n observation points $\bar{X}_1, \dots, \bar{X}_{s_n}$, independent and uniformly distributed on $[0, n^{1/d}]^d$. Let $A_1, \dots, A_k$ be k sequences of $s_n \times s_n$ random matrices so that, for $l = 1, \dots, k$, $(A_l)_{i,j}$ depends only on $\bar{X}_i$ and $\bar{X}_j$ and satisfies $|(A_l)_{i,j}| \le 1/(1 + |\bar{X}_i - \bar{X}_j|^{d+1})$. Then $\mathbb{E}_X(|A_1 \cdots A_k|^2)$ is bounded w.r.t. n.*

Lemma A.22. *The supremum over n, θ and X of the eigenvalues of R_θ^{-1}, $R_{1,\theta}^{-1}$,* $\operatorname{diag}(R_\theta^{-1})$, $\operatorname{diag}(R_{1,\theta}^{-1})$, $\operatorname{diag}(R_\theta^{-1})^{-1}$ *and* $\operatorname{diag}(R_{1,\theta}^{-1})^{-1}$ *is smaller than a constant $C_{\sup} < +\infty$.*

Lemma A.23. *Lemma A.22 also holds when K_θ is replaced by $\tilde{K}_\theta$ of Definition A.14.*

Lemma A.24. *Lemma A.22 also holds when R_θ is replaced by $\bar{R}_{k,\theta}$ of Definition A.14.*

Lemma A.25. *Let $k \in \mathbb{N}$. Let $A_{1,\theta}, \dots, A_{k,\theta}$ be k sequences of symmetric random matrices (functions of X and θ) so that, for any $m \in \mathbb{N}$, $a_1, \dots, a_m \in \{1, \dots, k\}$, $\sup_{\theta\in\Theta} \mathbb{E}_X |A_{a_1,\theta} \cdots A_{a_m,\theta}|^2$ is bounded (w.r.t. n). Let $B_{1,\theta}, \dots, B_{k+1,\theta}$ be $k + 1$ sequences of random symmetric non-negative matrices (functions of X and θ) so that $\sup_\theta \|B_{1,\theta}\|, \dots, \sup_\theta \|B_{k+1,\theta}\|$ are bounded (w.r.t. n and X). Then*

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

is bounded w.r.t. n.

Lemma A.26. *Consider a fixed $\theta \in \Theta$. With the notation of Definition A.14, we have, when $n_2 = o(n)$,*

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

Lemma A.27. *Let $C(t)$ be as in Definition A.14. Define, for $T \geq 0$, $f(T) = \int_{\mathbb{R}^d \backslash [-T,T]^d} 1/(1 + |t|^{d+1})\,dt$. Define, for $x \in [0, n^{1/d}]^d$, $D_\Delta(x) = \inf_{t \in \mathbb{R}^d \backslash C(x)} |x - t|$. Define $D_\Delta(x_1, \dots, x_m) = \min_{i=1,\dots,m} D_\Delta(x_i)$. Then, there exists a finite constant $C_{\sup}$ so that, for any n, for any $x_1, x_2 \in [0, n^{1/d}]^d$,*

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

Lemma A.28. *Use the notation n_2, Δ, $C(t)$, $f(T)$ and $D_\Delta(x_1,x_2)$ of Definition A.14 and Lemma A.27. Then, when $n_2 = o(n)$,*

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

Lemma A.29. *Use the notation n_2, Δ and $C_1, \dots, C_{n_2}$ of Definition A.14. Let, for $i = 1, \dots, n_2$, $X_1^i, \dots, X_{N_i}^i$ be the N_i components of X that are in C_i (so that the order of their indices in X is preserved). Then*

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

(ii) *Conditionally to $N_i = k_i$, $X_1^i, \dots, X_{k_i}^i$ 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 $(X_1^i, \dots, X_{k_i}^i)$ and $(X_1^j, \dots, X_{k_j}^j)$ are independent, and their components are independent and uniformly distributed on C_i and C_j, respectively.*

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

(iv) *The variables $f_1(X), \dots, f_{n_2}(X)$ 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.30. *Use the notation of Lemma A.29, and consider n_2 functions $f_1, \dots, f_{n_2}$ that satisfy the conditions of Lemma A.29. Assume that there exist fixed even natural numbers q, l and a finite constant $C_{\sup}$ (independent of n and X) so that $\mathbb{E}(f_i^2(X)|N_i = k) \leq C_{\sup}(1 + k^q + k^{q+l}/\Delta^l)$. Then, if $\Delta \to_{n \to \infty} +\infty$ and $\Delta = O(n^{1/(2q+5)})$,*

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

Lemma A.31. *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.32. *Let n_2, Δ and $C_1, \dots, C_{n_2}$ be as in Definition A.14. Assume that Δ is lower bounded, as a function of n. Then, there exists a finite constant $C_{\sup}$ so that for any n, $i \in \{1, \dots, n_2\}$,*

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

Lemma A.33. *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 research leading to this paper was partly carried out when the author was affiliated to the University of Vienna. The author thanks Josselin Garnier, Benedikt Pötscher and Luc Pronzato for constructive discussions. In addition, the author is grateful to two anonymous referees for their comments, that led to an improvement of the manuscript.

Supplementary Material

Figures and proof of the technical results (DOI: 10.3150/16-BEJ906SUPP; .pdf). In the supplementary material [8], we provide Figures 1 and 2, complementing the one-dimensional illustrative Monte Carlo simulation. We also give the proof of the lemmas stated in Section A.6.

References

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

[2] Anderes, E. (2010). On the consistent separation of scale and variance for Gaussian random fields. *Ann. Statist.* **38** 870–893. MR2604700

[3] Andrianakis, I. and Challenor, P.G. (2012). The effect of the nugget on Gaussian process emulators of computer models. *Comput. Statist. Data Anal.* **56** 4215–4228. MR2957866

[4] Azencott, R. and Dacunha-Castelle, D. (1986). *Series of Irregular Observations*: *Forecasting and Model Building*. *Applied Probability. A Series of the Applied Probability Trust*. New York: Springer. MR0848355

[5] Bachoc, F. (2013). Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Comput. Statist. Data Anal.* **66** 55–69. MR3064023

[6] Bachoc, F. (2013). Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments. Ph.D. thesis, Université Paris-Diderot – Paris VII. Available at https://tel.archives-ouvertes.fr/tel-00881002/document.

[7] Bachoc, F. (2014). Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *J. Multivariate Anal.* **125** 1–35. MR3163828

[8] Bachoc, F. (2016). Supplement to "Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case." DOI:10.3150/16-BEJ906SUPP.

[9] Bachoc, F., Bois, G., Garnier, J. and Martinez, J.M. (2014). Calibration and improved prediction of computer models by universal Kriging. *Nucl. Sci. Eng.* **176** 81–97.

[10] Chevalier, C., Ginsbourger, D., Bect, J., 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* **56** 455–465. MR3290615

[11] Conti, S. and O'Hagan, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *J. Statist. Plann. Inference* **140** 640–651. MR2558393

[12] Cressie, N. and Lahiri, S.N. (1993). The asymptotic distribution of REML estimators. *J. Multivariate Anal.* **45** 217–233. MR1221918

[13] Cressie, N. and Lahiri, S.N. (1996). Asymptotics for REML estimation of spatial covariance parameters. *J. Statist. Plann. Inference* **50** 327–341. MR1394135

[14] Dubrule, O. (1983). Cross validation of Kriging in a unique neighborhood. *J. Int. Assoc. Math. Geol.* **15** 687–699. MR0720633

[15] Furrer, R., Bachoc, F. and Du, J. (2016). Asymptotic properties of multivariate tapering for estimation and prediction. *J. Multivariate Anal.* **149** 177–191. MR3507322

[16] Furrer, R., Genton, M.G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. MR2291261

[17] Gneiting, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988

[18] Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

[19] Gramacy, R.B. and Apley, D.W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. MR3357395

[20] Gray, R.M. (2006). Toeplitz and circulant matrices: A review. *Found. Trends Commun. Inf. Theory* **2** 155–239.

[21] Hallin, M., Lu, Z. and Yu, K. (2009). Local linear spatial quantile regression. *Bernoulli* **15** 659–686. MR2555194

[22] Handcock, M.S. and Wallis, J.R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *J. Amer. Statist. Assoc.* **89** 368–390. MR1294070

[23] 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.

[24] Jones, D.R., Schonlau, M. and Welch, W.J. (1998). Efficient global optimization of expensive blackbox functions. *J. Global Optim.* **13** 455–492. MR1673460

[25] Kou, S.C. (2003). On the efficiency of selection criteria in spline regression. *Probab. Theory Related Fields* **127** 153–176. MR2013979

[26] Lahiri, S.N. (2003). Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs. *Sankhyā* **65** 356–388. MR2028905

[27] 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. *Ann. Inst. Statist. Math.* **56** 225–250. MR2067154

[28] Lahiri, S.N. and Robinson, P.M. (2016). Central limit theorems for long range dependent spatial linear processes. *Bernoulli* **22** 345–375. MR3449786

[29] Lahiri, S.N. and Zhu, J. (2006). Resampling methods for spatial regression models under a class of stochastic designs. *Ann. Statist.* **34** 1774–1813. MR2283717

[30] Le Gratiet, L. and Garnier, J. (2014). Asymptotic analysis of the learning curve for Gaussian process regression. *Mach. Learn.* 1–27.

[31] Loh, W.-L. (2005). Fixed-domain asymptotics for a subclass of Matérn-type Gaussian random fields. *Ann. Statist.* **33** 2344–2394. MR2211089

[32] Mardia, K.V. and Marshall, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135–146. MR0738334

[33] Marrel, A., Iooss, B., Van Dorpe, F. and Volkova, E. (2008). An efficient methodology for modeling complex computer codes with Gaussian processes. *Comput. Statist. Data Anal.* **52** 4731–4744. MR2521618

[34] Martin, J.D. and Simpson, T.W. (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.

[35] Paulo, R., García-Donato, G. and Palomo, J. (2012). Calibration of computer models with multivariate output. *Comput. Statist. Data Anal.* **56** 3959–3974. MR2957846

[36] Putter, H. and Young, G.A. (2001). On the effect of covariance function estimation on the accuracy of Kriging predictors. *Bernoulli* **7** 421–438. MR1836738

[37] Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. Cambridge, MA: MIT Press. MR2514435

[38] Ripley, B.D. (1981). *Spatial Statistics*. New York: Wiley. MR0624436

[39] Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–423. MR1041765

[40] Santner, T.J., Williams, B.J. and Notz, W.I. (2003). *The Design and Analysis of Computer Experiments. Springer Series in Statistics*. New York: Springer. MR2160708

[41] Scheuerer, M. (2010). Regularity of the sample paths of a general second order random field. *Stochastic Process. Appl.* **120** 1879–1897. MR2673978

[42] Shi, T., Belkin, M. and Yu, B. (2009). Data spectroscopy: Eigenspaces of convolution operators and clustering. *Ann. Statist.* **37** 3960–3984. MR2572449

[43] Stein, M. (1990). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *Ann. Statist.* **18** 850–872. MR1056340

[44] Stein, M.L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Ann. Statist.* **16** 55–63. MR0924856

[45] Stein, M.L. (1990). Bounds on the efficiency of linear predictions using an incorrect covariance function. *Ann. Statist.* **18** 1116–1138. MR1062701

[46] Stein, M.L. (1990). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Statist.* **18** 1139–1157. MR1062702

[47] Stein, M.L. (1993). Spline smoothing with an estimated order parameter. *Ann. Statist.* **21** 1522–1544. MR1241277

[48] Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging. Springer Series in Statistics*. New York: Springer. MR1697409

[49] Sundararajan, S. and Keerthi, S.S. (2001). Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural Comput.* **13** 1103–1118.

[50] Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4** 389–396. MR1366510

[51] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

[52] Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *J. Multivariate Anal.* **36** 280–296. MR1096671

[53] Ying, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *Ann. Statist.* **21** 1567–1590. MR1241279

[54] Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303

[55] Zhang, H. and Wang, Y. (2010). Kriging and cross-validation for massive spatial data. *Environmetrics* **21** 290–304. MR2842244

[56] Zhang, H. and Zimmerman, D.L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* **92** 921–936. MR2234195

Received November 2015 and revised June 2016