A comparison of likelihood methods and their fast randomized versions for stochastic process models with measurement error

Didier A. Girard

A COMPARISON OF LIKELIHOOD METHODS AND THEIR FAST RANDOMIZED VERSIONS FOR STOCHASTIC PROCESS MODELS WITH MEASUREMENT ERROR

Didier A. Girard

CNRS and Université Joseph Fourier

Abstract: To estimate the covariance parameters of a Gaussian spatial process from noisy observations at a finite number, say n, of sites, the methods of solving the likelihood equation or maximizing the likelihood are standard. Evaluating gradients of the log likelihood requires solving linear systems of size n and computing the trace of the associated matrix inverses; and, thus, in many various fields with large n, randomized trace estimates have been used for the second task. The purpose of this article is to quantify what is sacrificed when one uses a single (or the same n_{R}, for the version using an average of n_{R} randomized traces) simulated vector(s) of size n for the all gradient evaluations. We do this mainly in a simple one-dimensional stationary context, with the classical exponential function (thus at most 2 parameters) as covariance of the underlying spatial process, under infill asymptotics, for which consistency and asymptotic distribution results have been shown by Chen, Simpson and Ying (2000). We show that any consistent root of the randomized version of the likelihood equation has the same asymptotic behavior as for the exact version excepted that the asymptotic variances are increased by the factor $1 + 1/n_{\mathrm{R}}$. Moreover, to attack the problem of choosing between multiple, possibly non-consistent, roots, we propose a simple randomized version of the whole likelihood, whose maximizer is proved to be consistent even with $n_{\mathrm{R}} = 1$.

Key words and phrases: asymptotic normality, consistency, Gaussian process, identifiability, infill asymptotics, maximum likelihood estimator, measurement error, nonparametric Bayesian regression, randomized trace.

1. Introduction

We consider the classical problem of building up statistical inferences for the model

$$Y(s) = Z(s) + e(s), \quad s \in R^d \tag{1.1}$$

where Z is a zero mean Gaussian stochastic process whose covariance function is known up to a magnitude factor $b > 0$ and a shape parameter $\theta > 0$, and e is a "measurement error" process independent of Z. These parameters have to be estimated from only $n+1$ observations at sites $s(0), s(1), \ldots, s(n)$ with i.i.d. Gaussian noise $e(s(i)), i = 0, \ldots, n$; and we consider the classical maximum likelihood (sometimes called "marginal ML" or "type II ML" from an empirical Bayesian point of view) principle for this task.

To simplify the presentation we first assume that the noise variance is known, say equal to 1. Denoting by $bK(\theta)$ the assumed model for the covariance matrix of $\mathbf{z} = Z(s(0)), \ldots, Z(s(n))'$, the (marginal) law of $\mathbf{y} = Y(s(0)), \ldots, Y(s(n))'$ is then

$$\mathbf{y} \sim N(0, b_0 K(\theta_0) + I). \tag{1.2}$$

For future references, let us introduce the conditional mean of $Z(s)$ given $\mathbf{y}$, evaluated at $s(i), i = 0, \ldots, n$. This classical optimal prediction of $\mathbf{z}$ is well known to be $A_{\theta_0}(b_0)\mathbf{y}$ where

$$A_\theta(b) = \left(I + b^{-1} K(\theta)^{-1}\right)^{-1} \tag{1.3}$$

is an example of the so-called influence matrices in the spline literature. Now, simple manipulations of the density function (1.2) shows that -2 times the log likelihood function is

$$l(b, \theta) = \mathbf{y}' \left[I - A_\theta(b)\right] \mathbf{y} - \text{logdet}\left[I - A_\theta(b)\right] + (n+1)\log(2\pi) \tag{1.4}$$

and that its derivative w.r.t. the magnitude factor b (a score function) is

$$S_\theta(b) = -\frac{1}{b}\Big(\mathbf{y}' A_\theta(b) \left[I - A_\theta(b)\right] \mathbf{y} - \text{tr} A_\theta(b)\Big). \tag{1.5}$$

Various forms and extensions (e.g. for multidimensional b or θ) of these formulae are almost ubiquitous in the geostatistics field. There, and in many other fields, the size of the data may be very large. So a number of techniques have been developed to implement the likelihood principle.

In recent years, fast randomized versions of the likelihood equation (i.e., here, of solving $S_\theta(b) = 0$, which is called the LE or "scoring" method) has been used in many contexts with large data sets (e.g., Wahba et al. 1994): in its simplest

form, it consists in generating a single vector $\mathbf{w} \sim N(0, I)$, of the same size as $\mathbf{y}$, and replacing the function $S_\theta(b)$ by

$$ {}^{\mathrm{R}}S_\theta(b) = -\frac{1}{b}\Big(\mathbf{y}' A_\theta(b)\left[I - A_\theta(b)\right]\mathbf{y} - \mathbf{w}' A_\theta(b)\mathbf{w}\Big). \tag{1.6}$$

Thus, all we need for evaluating such a randomized score, is a (fast) algorithm for the computation of any conditional mean $A_\theta(b)\mathbf{y}$. We will also denote by ${}^{\mathrm{R}}S_\theta(b)$ the "averaged" randomized score obtained when $\mathbf{w}' A_\theta(b)\mathbf{w}$ is replaced by a simple Monte-Carlo average $(1/n_{\mathrm{R}})\sum_{r=1}^{n_{\mathrm{R}}} \mathbf{w}^{r\prime} A_\theta(b)\mathbf{w}^r$.

For example, in the machine learning field, fitting such models is become a popular approach to regression, which was recently widely studied under the name "Gaussian process regression". In the survey by Gibbs and MacKay (1997) the authors advise to use iterative linear solvers and averages of randomized trace estimates for the computation of any gradient of the log likelihood criterion. Gibbs and MacKay (1997) noticed that the number (n_{R} is our notation) of randomized trace estimates needed to obtain sufficiently good estimates is surprisingly small. The cost of such inferences is then reduced from order n^3 to order n^2 in general (that is, for full and not structured matrices). Much greater computational gains can actually be obtained in other contexts; see the final section.

The purpose of this paper is to give theoretical explanations of such behaviors when n_{R} is a fixed number (for example equal to 10); and we do this by an asymptotic analysis in the simple context studied in Chen, Simpson and Ying (2000) (abbreviated as CSY henceforth): in (1.1), Z is a classical Ornstein-Uhlenbeck process, that is, a one-dimensional zero-mean Gaussian process whose covariance function is the exponential function:

$$ E(Z(s)Z(t)) = b\exp\{-\theta|t-s|\}. \tag{1.7}$$

Note that we essentially adopt the same notations as in CSY except that the magnitude of the covariance function is denoted here by b (in place of σ^2 used in CSY). We consider infill (or "fixed domain") asymptotics, that means that the $s(i)$s becomes dense in a compact interval as n increases. Furthermore, as in CSY we consider equally spaced sites $s(i) = i/n, i = 0, \ldots, n$.

Infill asymptotic frameworks may be more useful than the classical time series frameworks to explain some empirical facts when one is faced with strong

correlations in the underlying process. This is emphasized in Zhang (2004), Zhang and Zimmerman (2005). For example, a clear explanation of the practical difficulties of estimating b_0 and θ_0 simultaneously, is provided by the important remark: it is known (e.g. Ying (1991)) that as long as $\theta b = \theta_1 b_1$, the couples (θ, b) and (θ_1, b_1) are not distinguishable from a single sample path $Z(t), t \in [0,1]$ of the Ornstein-Uhlenbeck process with covariance (1.7).

As was also noted by Zhang and Zimmerman (2005) the available results under infill asymptotics are considerably narrower in scope than for increasing domain. This is even more true when the measurements are obtained with errors or there is a so-called nugget effect. In two notable exceptions, analogs of Theorem 2.1 below are proved for the Brownian motion plus white noise model and for its mth order spline generalization, by Stein (1990) and Kou (2003). Note that, of course, in the model of CSY, randomized-traces are not useful for computing ML estimates since, as is well known, for this model, the exact likelihood (and the score) are calculable in $O(n)$ operations by standard procedures akin to Kalman filtering (e.g. Harvey (1994 section 3.4)). Before going on, let us notice, however, that in the context of a simple additive multidimensional extension of this model, these randomized procedures are particularly suitable: this will be discussed in the final section.

In order to simplify the presentation, we assume that the variance of the white noise is known. Note that, for this variance, one could easily construct a lot of estimators which have the same asymptotic behavior as the ML estimate and are easy to compute. As in CSY, we assume that we know lower (> 0) and upper bounds for b_0 and θ_0 so that the (approximate) likelihoods can be maximized over $D = [\underline{b}, \overline{b}] \times [\underline{\theta}, \overline{\theta}]$ with $\underline{b} > 0, \underline{\theta} > 0$.

As in CSY, two cases are distinguished depending on whether θ_0 is known or not. In the second case, in view of the above remark on the non-identifiability of the couple (b_0, θ_0) and the neat result in CSY on the estimation of their product, we focus on the estimation of the single parameter $b_0\theta_0$.

One of the main results of this paper is that any consistent estimate obtained by solving the randomized likelihood equation converges to the true parameter with the same rate as the estimate obtained with the exact likelihood, even with $n_{\mathrm{R}} = 1$, and with an asymptotic variance only inflated by a factor $1 + 1/n_{\mathrm{R}}$.

However, quite often, the likelihood equation, even computed exactly, may have several roots. In such cases, it is indispensable to select a root which globally maximizes the likelihood since consistency is then guaranteed (cf. Theorems 2.2, 2.4 rephrased from CSY and Theorem 2.5).

A second type of results thus concerns a randomized version of the global log likelihood function. From the expression (1.6), it is natural to propose the following definition, in an integral form: having generated $\mathbf{w}$ and chosen a "boundary" point b_1 (although b_1 is not necessarily at the boundary of the search domain D), -2 times this randomized log likelihood function is defined by:

$$ {}^{\mathrm{R}}l_{b_1}(b,\theta) = \mathbf{y}'\left(I - A_\theta(b)\right)\mathbf{y} + \int_{b_1}^{b} \frac{\mathbf{w}' A_\theta(s)\mathbf{w}}{s}\mathrm{d}s - \mathrm{logdet}\left[I - A_\theta(b_1)\right] + (n+1)\log(2\pi). \tag{1.8} $$

As for the score (1.6), this expression is the particular case $n_{\mathrm{R}} = 1$ of an averaged version, also denoted by ${}^{\mathrm{R}}l$ (note that b_1 will often be omitted), directly obtained with $(1/n_{\mathrm{R}})\sum_{r=1}^{n_{\mathrm{R}}} \mathbf{w}^{r\prime} A_\theta(s)\mathbf{w}^r$ in place of $\mathbf{w}' A_\theta(s)\mathbf{w}$. Evaluating this criterion is relatively easy as soon we have at hand an efficient algorithm for the computation of any conditional mean $A_\theta(b)\mathbf{y}$, excepted for the log-determinants at the boundary values (b_1, θ) (see the final section for comments on the approximation of the integral by discrete sum). Of course, in the case θ_0 known, there is no need to compute $\mathrm{logdet}\left[I - A_{\theta_0}(b_1)\right]$ since this is a constant term in the objective function (1.8). This computational simplification is also available in the case where θ is constrained to be a predetermined value θ_1 in the search domain. In this paper, we show that consistency is still guaranteed (for the product θ_1 times the constrained maximizer in the case θ_0 unknown) by maximizing these randomized likelihoods.

Our results are stated in Section 3. Before that, in order to make easy the comparisons with the non-randomized versions, we essentially recall in Section 2 the main results of CSY. Section 2.1 and Section 3.1 concern the estimation of b_0 assuming θ_0 to be known. Section 2.2 and Section 3.2 concern the estimation of the product $b_0\theta_0$. Proofs are given in Section 4. In the final section, we connect the proposal (1.8) with two techniques already used in the literature, and we discuss possible applications and extensions of our results.

2. "Asymptotic background" for the exact likelihood principle

First, recall that it is easily seen that, in the case θ_0 known, the standard Fisher information for the magnitude parameter has the following simple expression

$$E\left(\frac{1}{2}\frac{\partial^2 l}{\partial b^2}(b_0,\theta_0)\right) = \mathrm{var}\frac{1}{2}S_{\theta_0}(b_0) = \frac{\mathrm{tr}A^2_{\theta_0}(b_0)}{2{b_0}^2}$$

for the generic model (1.2)-(1.3); and an asymptotic equivalent for this information, as $n \to +\infty$, in the CSY context (1.7) that we study in Sections 2, 3 and 4 under the assumption $s(i) = i/n, i = 0, \ldots, n$, is $c(b_0,\theta_0)n^{1/2}$ with $c(b_0,\theta_0) > 0$. So the occurrences of the power $n^{1/4}$ in the following statements, in place of the usual $n^{1/2}$ of the i.i.d. case (or $\theta = +\infty$), are natural.

Let us recall that we assume that the true parameter is in the interior of the search domain D which is either $[\underline{b},\overline{b}]$ or $[\underline{b},\overline{b}] \times [\underline{\theta},\overline{\theta}]$ with $\underline{b} > 0, \underline{\theta} > 0$.

2.1. Case θ_0 known.

From the work of CSY, we easily deduce than their asymptotic normality results can be stated in the following slightly more general form, since they were, there, restricted to the exact ML estimate:

Theorem 2.1.(CSY) *Let $\hat{b}$ be any candidate estimate obtained by solving the likelihood equation at θ_0, possibly up to $\mathrm{o_p}(n^{1/4})$, viz. satisfying*

$$S_{\theta_0}(\hat{b}) = \mathrm{o_p}(n^{1/4})$$

with $S_{\theta_0}(b)$ defined from (1.5)*. If $\hat{b} \to b_0$ in probability then*

$$n^{1/4}(\hat{b} - b_0) \to_{\mathcal{D}} N(0, 4\sqrt{2}{\theta_0}^{-1/2}b_0^{3/2}).$$

Now we rephrase the first consistency result of CSY. Let $\hat{b}_{\mathrm{ML}}$ denote any minimizer of -2 times the log likelihood function (i.e. of $l(\cdot,\theta_0)$ defined from (1.4)).

Theorem 2.2.(CSY) $\hat{b} := \hat{b}_{\mathrm{ML}}$ *satisfies the conditions of Theorem 2.1.*

2.2. Case θ_0 unknown.

In CSY an asymptotic behavior identical to the behavior of $\hat{b}\theta_0$ of the case θ_0 known above, is proved for the product of LE estimators, precisely:

Theorem 2.3.(CSY) *Let $(\tilde{b}, \tilde{\theta})$ be any couple of candidate estimates obtained by solving the likelihood equation possibly up to $o_p(n^{1/4})$, viz. satisfying*

$$S_{\tilde{\theta}}(\tilde{b}) = o_p(n^{1/4})$$

with $S_\theta(b)$ defined in (1.5)*. If $\tilde{b}\tilde{\theta} \to b_0\theta_0$ in probability then*

$$n^{1/4}(\tilde{b}\tilde{\theta} - b_0\theta_0) \to_{\mathcal{D}} N(0, 4\sqrt{2}(b_0\theta_0)^{3/2}).$$

Now, the following theorem is the important result that the ML principle, which was already justified in the time series framework (e.g. Zhang and Zimmerman (2005)), is also justified here to estimate the identifiable product $b_0\theta_0$. Let $(\tilde{b}_{\mathrm{ML}}, \tilde{\theta}_{\mathrm{ML}})$ denote any couple minimizer, over D, of $l(\cdot,\cdot)$ defined in (1.4).

Theorem 2.4.(CSY) $(\tilde{b}, \tilde{\theta}) := (\tilde{b}_{\mathrm{ML}}, \tilde{\theta}_{\mathrm{ML}})$ *satisfies the conditions of Theorem 2.3.*

We now claim a result which was not stated in CSY (however see Ying (1991) for the case of no measurement error) which is useful here for the computational simplification outlined in the Introduction (penultimate paragraph). Let $\theta_1 > 0$ be an arbitrarily fixed constant. Let $\hat{b}_{\mathrm{ML}}(\theta_1)$ denote any minimizer of l defined in (1.4) restricted to $\theta = \theta_1$, i.e.

$$l(\hat{b}_{\mathrm{ML}}(\theta_1), \theta_1) = \inf_{b \in [\underline{b}, \overline{b}]} l(b, \theta_1).$$

Theorem 2.5. *Assume that $b_0\theta_0/\theta_1$ is in the interior of $[\underline{b}, \overline{b}]$. Then $(\tilde{b}, \tilde{\theta}) := (\hat{b}_{\mathrm{ML}}(\theta_1), \theta_1)$ satisfies the conditions of Theorem 2.3.*

3. Results

The context is the same as the one of Section 2. Of course the simulated vector(s) $\mathbf{w}$(s) is (are) assumed independent of the observation $\mathbf{y}$. We first state a result on the global accuracy of the randomized likelihood function (1.8) as an

approximation of the exact one on the whole domain. Since $\mathrm{logdet}\,[I - A_\theta(b)] - \mathrm{logdet}\,[I - A_\theta(b_1)] = -\int_{b_1}^{b} s^{-1}\mathrm{tr}A_\theta(s)\mathrm{d}s$, it is stated in the following form:

Lemma 3.1. *For any given $b_1 > 0$, for all $\alpha > 0$, we have*

$$\int_{b_1}^{b} \left(\frac{\mathbf{w}' A_\theta(s)\mathbf{w}}{s} - \frac{\mathrm{tr}A_\theta(s)}{s} \right) \mathrm{d}s = \mathrm{o_p}(n^{1/4+\alpha})$$

uniformly in $(b, \theta) \in D$.

Remark 3.1. The resulting order of convergence for the global difference ${}^{\mathrm{R}}l_{b_1} - l$ will be seen sufficient. Of course an extension of this result to the case $b_1 = 0$ would be very interesting for the practice (since the boundary log-determinants in ${}^{\mathrm{R}}l_{b_1}$ would then be trivially 0) but we think that the assumption $b_1 > 0$ is important; this can be seen by analyzing the behavior of the continuous integrand at the limit value $b = 0$, noting that $b^{-1}A_\theta(b) = \left(b + K(\theta)^{-1}\right)^{-1}$ approaches $K(\theta)$; indeed the minimal order of $\mathbf{w}'K(\theta)\mathbf{w} - \mathrm{tr}K(\theta)$ is easily seen to be the much larger $\mathrm{o_p}(n^{1+\alpha})$ instead of $\mathrm{o_p}(n^{1/4+\alpha})$.

3.1. Case θ_0 known.

We first state an analog of Theorem 2.1 for consistent roots of the randomized score. Note that the same n_{R} vectors $\mathbf{w}$s are used for all (b, θ), in (1.6) or (1.8).

Theorem 3.1. *Let ${}^{\mathrm{R}}\hat{b}$ be any candidate estimate obtained by solving the randomized likelihood equation at θ_0, possibly up to $\mathrm{o_p}(n^{1/4})$, viz. satisfying*

$${}^{\mathrm{R}}S_{\theta_0}({}^{\mathrm{R}}\hat{b}) = \mathrm{o_p}(n^{1/4})$$

with ${}^{\mathrm{R}}S_{\theta_0}(b)$ defined, in the case $n_{\mathrm{R}} = 1$, from (1.6). If ${}^{\mathrm{R}}\hat{b} \to b_0$ in probability then

$$n^{1/4}({}^{\mathrm{R}}\hat{b} - b_0) \to_{\mathcal{D}} N\left(0, \left(1 + n_{\mathrm{R}}^{-1}\right) 4\sqrt{2}\theta_0{}^{-1/2} b_0^{3/2}\right).$$

Remark 3.2. In the quite different context of nonparametric estimation of deterministic functions by kernel methods, results which have a roughly similar appearance are given in Girard (1998) for randomized GCV, except that, there, the obtained relative increase in variance was strictly lower than $1 + 1/n_{\mathrm{R}}$.

Now we claim that the consistency property is intact with randomized traces. Let $\hat{b}_{\mathrm{RML}}$ denote a minimizer of -2 times the randomized log likelihood function (i.e. of ${}^{\mathrm{R}}l_{b_1}(\cdot, \theta_0)$ defined from (1.8) with b_1 any fixed "boundary" value > 0).

Theorem 3.2. *For any fixed $n_{\mathrm{R}} \geq 1$, ${}^{\mathrm{R}}\hat{b} := \hat{b}_{\mathrm{RML}}$ satisfies the conditions of Theorem 3.1.*

3.2. Case θ_0 unknown.

The analog of Theorem 2.3 for any consistent couple of roots for the scalar randomized score is the following:

Theorem 3.3. *Let $({}^{\mathrm{R}}\tilde{b}, {}^{\mathrm{R}}\tilde{\theta})$ be any couple of candidate estimates obtained by solving the randomized likelihood equation possibly up to $\mathrm{o_p}(n^{1/4})$, viz. satisfying*

$$ {}^{\mathrm{R}}S_{{}^{\mathrm{R}}\tilde{\theta}}({}^{\mathrm{R}}\tilde{b}) = \mathrm{o_p}(n^{1/4}) $$

with ${}^{\mathrm{R}}S_\theta(b)$ defined, in the case $n_{\mathrm{R}} = 1$, in (1.6). If ${}^{\mathrm{R}}\tilde{b}{}^{\mathrm{R}}\tilde{\theta} \to b_0\theta_0$ in probability then

$$ n^{1/4}({}^{\mathrm{R}}\tilde{b}{}^{\mathrm{R}}\tilde{\theta} - b_0\theta_0) \to_{\mathcal{D}} N\left(0, \left(1 + n_{\mathrm{R}}^{-1}\right) 4\sqrt{2}(b_0\theta_0)^{3/2}\right). $$

We now claim two consistency results.

Let $b_1 > 0$ be an arbitrarily fixed value for the boundary point used in (1.8). Let $(\tilde{b}_{\mathrm{RML}}, \tilde{\theta}_{\mathrm{RML}})$ denote any couple minimizer, over D, of ${}^{\mathrm{R}}l := {}^{\mathrm{R}}l_{b_1}$ defined in (1.8).

Theorem 3.4. *For any fixed $n_{\mathrm{R}} \geq 1$, $({}^{\mathrm{R}}\tilde{b}, {}^{\mathrm{R}}\tilde{\theta}) := (\tilde{b}_{\mathrm{RML}}, \tilde{\theta}_{\mathrm{RML}})$ satisfies the conditions of Theorem 3.3.*

The randomized analog of Theorem 2.5, which is a consistency result allowing a misspecified model, also holds. Let $\theta_1 > 0$ be an arbitrarily fixed constant. Let $\hat{b}_{\mathrm{RML}}(\theta_1)$ denote any minimizer of ${}^{\mathrm{R}}l$ defined in (1.8) restricted to $\theta = \theta_1$

$$ {}^{\mathrm{R}}l(\hat{b}_{\mathrm{RML}}(\theta_1), \theta_1) = \inf_{b \in [\underline{b}, \overline{b}]} {}^{\mathrm{R}}l(b, \theta_1). $$

Theorem 3.5. *Assume that $b_0\theta_0/\theta_1$ is in the interior of $[\underline{b}, \overline{b}]$. Then, for any fixed $n_{\mathrm{R}} \geq 1$, $({}^{\mathrm{R}}\tilde{b}, {}^{\mathrm{R}}\tilde{\theta}) := (\hat{b}_{\mathrm{RML}}(\theta_1), \theta_1)$ satisfies the conditions of Theorem 3.3.*

4. Proofs

Let us first recall basic results of CSY. These authors exhibit approximations for the exact likelihood and for its derivative w.r.t. b which are much more manageable. Define $\rho_\theta := \exp\{-\theta/n\}$, $\lambda_{b,\theta} := b(1-\rho_\theta^2) + (1-\rho_\theta)^2$. We also define $\rho_0 := \rho_{\theta_0}$ and $\lambda_0 := \lambda_{b_0,\theta_0}$ the true parameters. Note that CSY uses the notation λ_θ for $\lambda_{b,\theta}$, λ for λ_{b,θ_0} and ρ for ρ_0.

CSY gives, in their Lemma 1, a quasi-diagonalized formula for the exact log likelihood. Note that this was stated in the case θ fixed at θ_0 but it is clear in CSY that this formula holds globally of course with $\rho_\theta, \lambda_{b,\theta}$ in place of ρ, λ in their Lemma 1 (note that one must also use T_θ in place of T_{θ_0} where T_θ denotes the $n \times (n+1)$ bidiagonal matrix satisfying $(T_\theta \mathbf{y})_k = y_{k+1} - \rho_\theta y_k$ for any vector $\mathbf{y}$). To exploit this as in CSY, we need to recall the definition of

$$\gamma_k := 2\left(1 - \cos\frac{\pi k}{n+1}\right),\ k = 1, \ldots, n, \tag{4.1}$$

which are the eigenvalues of the $n \times n$ tridiagonal Toeplitz matrix with 2 on the diagonal and -1 on the two neighboring off-diagonals. Let $\mathbf{u}_k$, $k = 1, \ldots, n$ denote the corresponding eigenvectors.

4.1 Proofs of preliminary results

To prove Lemma 3.1, we first state a diagonalized approximation which holds with a largely sufficient accuracy: there exist i.i.d. ${}^{\mathrm{R}}W_k \sim N(0,1)$, $k = 1, \ldots, n$, such that, for all $\alpha > 0$, for all b, θ,

$$\frac{\mathbf{w}' A_\theta(b) \mathbf{w}}{b} - \frac{\mathrm{tr} A_\theta(b)}{b} = (1-\rho_\theta^2)\left(\sum_{k=1}^n \frac{{}^{\mathrm{R}}W_k{}^2}{\rho_\theta \gamma_k + \lambda_{b,\theta}} - \sum_{k=1}^n \frac{1}{\rho_\theta \gamma_k + \lambda_{b,\theta}}\right) + \mathrm{o_p}(n^\alpha) \tag{4.2}$$

Here and in the following the occurrences of $\mathrm{o}(\cdot), \mathrm{O}(\cdot)$ or $\mathrm{o_p}(\cdot)$ terms which are function of b or (b, θ) will always denote terms uniformly convergent over, respectively, fixed $[\underline{b}, \overline{b}]$ or $D = [\underline{b}, \overline{b}] \times [\underline{\theta}, \overline{\theta}]$ with $\underline{b} > 0, \underline{\theta} > 0$. Starting from their Lemma 1 and an analysis of the term of index $k = 0$ in the exact quasi-diagonalized formula which has $n+1$ terms, CSY actually gave a more general result than (4.2), concerning the score and stated in their (3.37), (3.41).

One of the steps in the proofs of CSY is the claim that, if $W_k, k = 1, \ldots, n$, are i.i.d. $N(0,1)$, then, for any $\alpha > 0$,

$$\sum_{k=1}^{n} \frac{{W_k}^2 - 1}{\rho_\theta \gamma_k + \lambda_{b,\theta}} = \mathrm{o_p}(n^{5/4+\alpha}). \tag{4.3}$$

(generalizing their Lemma 3(i)). So Lemma 3.1 is obtained by applying the mean value theorem and noting that $1 - {\rho_\theta}^2 = 2\theta/n + \mathrm{O}(n^{-2})$. The proof of (4.3) is not detailed in CSY. We think that the following comments might be useful for the reader : the fact that it holds pointwise is easy to see, by Chebyshev inequality, since the Lemma 3 of SCY (concerning the behavior of $\mathrm{tr}A_\theta(b)$ and $\mathrm{tr}A_\theta^2(b)$ and which is stated there for the case θ fixed at θ_0) also holds, of course, with ρ_0 (the β in SCY) replaced by ρ_θ and λ_{b,θ_0} (the λ in CSY) replaced by $\lambda_{b,\theta}$ provided appropriate simple changes are made for the constants. Next, one way to prove the required uniformity is to use a particular Lipschitz property of the weights of the $({W_k}^2 - 1)$'s, as functions of b and θ, precisely

$$\sup_{(b,\theta)\neq(b',\theta')} \frac{|g_{n,k}(b,\theta) - g_{n,k}(b',\theta')|}{\|(b,\theta) - (b',\theta')\|} \leq M \sup_{(b,\theta)} |g_{n,k}(b,\theta)| \tag{4.4}$$

where $g_{n,k}$ denotes $(\rho_\theta \gamma_k + \lambda_{b,\theta})^{-1}$, $\|\cdot\|$ is the Euclidean distance, the sup are over D, and M is independent of k, because it is known that this property is sufficient to translate a pointwise convergence in probability to a uniform one (see e.g. Strook and Varadhan (1979) or Corollary A of Wu (1981)). Indeed the property (4.4) can be checked via appropriate bounds over D for the partial derivatives of $g_{n,k}$ w.r.t. to b and θ; the details are tedious and thus omitted.

Starting from the observations

$$\rho_\theta = 1 - \frac{\theta}{n} + \mathrm{O}(n^{-2}), \;\; \lambda_{b,\theta} = \frac{2b\theta}{n} + \mathrm{O}(n^{-2}), \tag{4.5}$$

the following comments explain how to easily deduce Theorem 2.5 from the results of SCY. A key intermediate result in the proof of their Theorem 3 is the following:

$$S_\theta(b) = -\left(1 - \rho_\theta^2\right) \sum_{k=1}^{n} \frac{(\rho_0 \gamma_k + \lambda_0){W_k}^2}{(\rho_\theta \gamma_k + \lambda_{b,\theta})^2} + \left(1 - \rho_\theta^2\right) \sum_{k=1}^{n} \frac{1}{\rho_\theta \gamma_k + \lambda_{b,\theta}} + \mathrm{o_p}(n^{1/4}) \tag{4.6}$$

where $W_k = {\mathbf{u}_k}' T_{\theta_0} \mathbf{y}$ does not depend on θ or b. Now with (4.5) and appropriate

uniform convergence results like (4.3), it can be checked that

$$S_\theta(b) = -\frac{2\theta}{n}\sum_{k=1}^{n}\frac{\left(\gamma_k + \frac{2b_0\theta_0}{n}\right){W_k}^2}{\left(\gamma_k + \frac{2b\theta}{n}\right)^2} + \frac{2\theta}{n}\sum_{k=1}^{n}\frac{1}{\gamma_k + \frac{2b\theta}{n}} + \mathrm{o_p}(n^{1/4}) \tag{4.7}$$

(because the score will be seen to be of order $n^{1/4}$ at the true parameter, this uniform approximation indicates in passing that the product $b\theta$ is a natural choice for reparametrization of the numerical search). Let us denote

$$\tilde{l}(b,\theta) = \sum_{k=1}^{n}\frac{\left(\gamma_k + \frac{2b_0\theta_0}{n}\right){W_k}^2}{\gamma_k + \frac{2b\theta}{n}} + \sum_{k=1}^{n}\log\left(\gamma_k + \frac{2b\theta}{n}\right) \tag{4.8}$$

the ideal (not exactly observed) likelihood. The consistency toward b_0 of the minimizer of $\tilde{l}(b,\theta_0)$ could be obtained exactly as in the detailed proof of CSY for $l(b,\theta_0)$. A key intermediary result is that

$$\tilde{l}(b,\theta_0) - \tilde{l}(b_0,\theta_0) \geq \sum_{k=1}^{n^{1/3}}\left(\frac{\gamma_k + \frac{2b_0\theta_0}{n}}{\gamma_k + \frac{2b\theta}{n}} - 1 - \log\frac{\gamma_k + \frac{2b_0\theta_0}{n}}{\gamma_k + \frac{2b\theta}{n}}\right) + \mathrm{o_p}(n^{1/4+\alpha}), \tag{4.9}$$

and an analysis of the deterministic sum allows CSY to conclude (cf. their proof of Theorem 1). Now by observing that, denoting $\tilde{S}_\theta(b) := (\partial\tilde{l}/\partial b)(b,\theta)$, that is, the sums term at the right side of (4.7)

$$l(s,\theta_1) - l(b_0\frac{\theta_0}{\theta_1},\theta_1) = \tilde{l}(s,\theta_1) - \tilde{l}(b_0\frac{\theta_0}{\theta_1},\theta_1) + \int_{b_0\frac{\theta_0}{\theta_1}}^{s}\left(S_{\theta_1}(t) - \tilde{S}_{\theta_1}(t)\right)\mathrm{d}t \tag{4.10}$$

$$= \tilde{l}(s\frac{\theta_1}{\theta_0},\theta_0) - \tilde{l}(b_0,\theta_0) + \mathrm{o_p}(n^{1/4+\alpha}). \tag{4.11}$$

Thus from (4.9) applied to $b := s\theta_1/\theta_0$, we have $\hat{b}_{\mathrm{ML}}(\theta_1) \to b_0\theta_0/\theta_1$ in probability. The second condition (stationarity of the likelihood $l(\cdot,\theta_1)$ at $\hat{b}_{\mathrm{ML}}(\theta_1)$) is satisfied with a probability tending to 1 since $b_0\theta_0/\theta_1$ is assumed in the interior of $[\underline{b},\bar{b}]$.

4.2 Proofs of Theorems 3.2, 3.4 and 3.5

We only consider the general case of global minimization of ${}^{\mathrm{R}}l$, that is, θ_0 unknown. The proofs for the other cases are similar. To prove Theorem 3.4 the first essential argument is the following: by examining the detailed proof in CSY,

we observe that the proof of consistency for RML estimates can be reduced to the proof of

$$ {}^{\mathrm{R}}l(b,\theta) - {}^{\mathrm{R}}l(b_0,\theta_0) \geq \sum_{k=1}^{n^{1/3}} \left(\frac{\rho_0\gamma_k + \lambda_0}{\rho_\theta\gamma_k + \lambda_{b,\theta}} - 1 - \log\frac{\rho_0\gamma_k + \lambda_0}{\rho_\theta\gamma_k + \lambda_{b,\theta}} \right) + \mathrm{o_p}(n^{1/4+\alpha}). \tag{4.12} $$

The second essential argument is then Lemma 3.1 whose application gives enough accuracy to deduce (4.12) from its non-randomized analog proved in CSY.

4.3 Proofs of Theorems 3.1 and 3.3

As usual the asymptotic laws of the (approximate) roots are obtained via Taylor approximation. A simplifying feature of the CSY model is that the third derivative can be uniformly bounded in probability. First a simpler approximate form for the randomized score, similar as the approximate form (that we denote by $\tilde{S}_\theta(b)$) stated for $S_\theta(b)$ in (4.7), can be obtained with exactly the same proof: there exist i.i.d. ${}^{\mathrm{R}}W_k \sim N(0,1), k = 1,\ldots,n$, independent of the W_k's such that

$$ {}^{\mathrm{R}}\tilde{S}_\theta(b) = -\frac{2\theta}{n}\sum_{k=1}^{n} \frac{\left(\gamma_k + \frac{2b_0\theta_0}{n}\right) {W_k}^2}{\left(\gamma_k + \frac{2b\theta}{n}\right)^2} + \frac{2\theta}{n}\sum_{k=1}^{n} \frac{{}^{\mathrm{R}}{W_k}^2}{\gamma_k + \frac{2b\theta}{n}} \tag{4.13} $$

satisfies ${}^{\mathrm{R}}S_\theta(b) - {}^{\mathrm{R}}\tilde{S}_\theta(b) = \mathrm{o_p}(n^{1/4})$. We consider, to simplify the notation, the case θ_0 known (extension for the product $b\theta$, when θ_0 is unknown will be seen to be immediate) and we drop the index θ_0 in the following. The first condition of Theorem 3.1 is thus also equivalent to ${}^{\mathrm{R}}\tilde{S}({}^{\mathrm{R}}\hat{b}) = \mathrm{o_p}(n^{1/4})$. By Taylor expansion of ${}^{\mathrm{R}}\tilde{S}(\cdot)$ at b_0 and the second condition of Theorem 3.1, there exists $b^\star$ which tends to b_0 in probability, such that

$$ \mathrm{o_p}(n^{1/4}) = {}^{\mathrm{R}}\tilde{S}(b_0) + \left({}^{\mathrm{R}}\hat{b} - b_0\right) {}^{\mathrm{R}}\tilde{S}'(b_0) + \frac{1}{2}\left({}^{\mathrm{R}}\hat{b} - b_0\right)^2 {}^{\mathrm{R}}\tilde{S}''(b^\star) $$

Theorem 3.1 can be classically proved by combining the three following properties: First, ${}^{\mathrm{R}}\tilde{S}(b_0)$ satisfies a CLT theorem similar as $\tilde{S}(b_0)$ excepted that the variance is increased by a factor $1 + 1/n_{\mathrm{R}}$. Second, ${}^{\mathrm{R}}\tilde{S}'(b_0)$ is uniformly equivalent to the same $c(b_0,\theta_0)n^{1/2}$ as $\tilde{S}'(b_0)$. Third, $\left({}^{\mathrm{R}}\hat{b} - b_0\right) {}^{\mathrm{R}}\tilde{S}''(b^\star) = \mathrm{o_p}(n^{1/2})$. The first and the second properties are easily shown to hold true, similarly as for

their nonrandomized analogs. The third one can be derived from

$$\sup_{b\in[\underline{b},\bar{b}]} |{}^{\mathrm{R}}\tilde{S}''(b)| \leq 6\left(\frac{2\theta_0}{n}\right)^3 \sup_{b\in[\underline{b},\bar{b}]}\left(\sum_{k=1}^{n}\frac{\left(\gamma_k+\frac{2b_0\theta_0}{n}\right){W_k}^2}{\left(\gamma_k+\frac{2b\theta_0}{n}\right)^4}+\sum_{k=1}^{n}\frac{{}^{\mathrm{R}}{W_k}^2}{\left(\gamma_k+\frac{2b\theta_0}{n}\right)^3}\right)$$

$$\leq \frac{M}{n^3}\sum_{k=1}^{n}\frac{{W_k}^2+{}^{\mathrm{R}}{W_k}^2}{\left(\gamma_k+\frac{2b_0\theta_0}{n}\right)^3} = \mathrm{O_p}(n^{1/2})$$

where the second inequality, with M independent of n, results from Lemma 3(iii) of CSY and the last equality results from $\sum_{k=1}^{n}(\gamma_k+\frac{2b\theta}{n})^{-3} \sim \mathrm{const}\, n^{7/2}$ and Markov inequality.

5. Discussion

Concerning the randomized likelihood function defined in (1.8), a connection with two other works is in order.

This approximation of the log-determinant has some similarities with techniques, today called path sampling, introduced by Ogata (1990) for implementing ML hyperparameter estimation in Gaussian (or Gaussian approximation of) Bayesian models of large size, like those encountered in image analysis. There the integral expression for a log-determinant is approximated by a weighted sum (from e.g. the trapezoidal rule of numerical integration) of simulated averages ($(1/n_{\mathrm{R}})\sum_{r=1}^{n_{\mathrm{R}}} s^{-1}\mathbf{w}^{r\prime}A_\theta(s)\mathbf{w}^r$ in our notation) on a grid of s-values, where the n_{R} inner products may be generated by a MCMC technique.

From the symmetry of $A_\theta(s)$, the expression of its eigenvalues (and those of its powers $A_\theta^k(s)$) as functions of s, and the identity $-\int_0^b(s+c)^{-1}\mathrm{d}s = -\sum_{k=0}^{\infty}(1+b^{-1}c)^{-k}/k$ which holds for any $b>0, c>0$, it is direct to deduce that $\int_0^b s^{-1}\mathbf{w}'A_\theta(s)\mathbf{w}\mathrm{d}s = \sum_{k=0}^{\infty}\mathbf{w}'A_\theta^k(b)\mathbf{w}/k$. This integral can thus be computed by truncating this series, with the attractive property that computing the first k terms requires only k matrix-vector products invoking repeatedly the same $A_\theta(b)$. Such a technique has already been introduced (without passing by the integral form) by Barry and Pace (1999) in the context of fitting conditional autoregressive (or CAR) models. Bounds for the bias incurred by truncating with few (e.g. 10) terms, and very encouraging numerical experiments analyzing the variability of these log-determinant estimators, can be found in that paper.

Concerning the choice of an algorithm to approximate the integrals by discrete sums or the choice of a truncation index in the power series approach, it seems quite hard to a priori control their impact on the statistical accuracy of the parameter estimates. To circumvent this difficulty, one may look after other methods (e.g. speudo-likelihoods) which would be able, at a reasonable cost, to produce a consistent, if not efficient, estimate, and this initial estimate would then be a starting point for a local search of an approximate root of the randomized likelihood equation (analogs of Theorems 3.1 and 3.3 might then be useful). This deserves further study.

A simple example of multi-dimensional process is the additive model

$$Y(s_1, s_2) = Z_1(s_1) + Z_2(s_2) + e(s_1, s_2)$$

where Z_1 and Z_2 are two independent Gaussian processes and e represents the measurement error. CSY alluded to this example and wrote that, even if we assume the same Ornstein-Uhlenbeck probability measures (parametered by b and θ) for the two one-dimensional components, the maximum likelihood estimators for this model appear to be more elusive. On the other hand, it is known that any conditional mean $A_\theta(b)\mathbf{y}$ can be efficiently computed, even with very large n, by iterative techniques called backfitting. Thus the proposed randomized likelihood techniques are particularly suitable. It would be thus interesting to establish the asymptotic behaviors of the produced estimates of b and θ in these settings. In view of the arguments used for the proofs in Section 4.3, it is expected that, at least in the case θ_0 known, the increase of variance by the factor $1 + 1/n_{\mathrm{R}}$ should also hold.

References

Barry R. and Pace R.K. (1999) A Monte Carlo estimator of the log determinant of large sparse matrices. *Linear Algebra and its Applications* **289**, 41-54

Chen H.S., Simpson D.G. and Ying Z. (2000). Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica* **10**, 141-156.

Gibbs M.N. and MacKay D.J.C. (1997). Efficient implementation of Gaussian processes. Technical report, Department of Physics, Cambridge University.

Girard D.A. (1998). Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression. *Ann. Statist.* **26**, 315-334

Kou S.C. (2003). On the efficiency of selection criteria in spline regression. *Probab. Theory Relat. Fields* **127**, 153-176.

Ogata Y. (1990) A Monte Carlo method for an objective Bayesian procedure. *Ann. Inst. Stat. Math.* **42**, 403-433.

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.

Strook D.W. and Varadhan S.R.S. (1979). Multidimensional Diffusion Processes. Springer, Berlin.

Wahba G., Johnson D.R., Gao F. and Gong J. (1994). Adaptive tuning of numerical weather prediction models: Part I: randomized GCV and related methods in three and four dimensional data assimilation. Technical report 920, Department of Statistics, University of Wisconsin.

Wu C.F. (1981) Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9**, 501-513.

Ying Z. (1991) Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate analysis* **26**, 280-296

Zhang H. (2004) Inconsistent estimation and asymptotically equivalent interpolation in model-based geostatistics. *J. Amer. Statist. Assoc.* **99**, 250-261.

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

CNRS, LMC - IMAG, Tour des Math., B.P. 53, F 38041 GRENOBLE cedex 9, France

E-mail: didier.girard@imag.fr