1. Introduction. Gaussian processes are often used to emulate unknown functions from some space \( \mathbb{R}^r \) \((r \geq 1)\) to \( \mathbb{R} \) (Rasmussen and Williams, 2006). Interpreted in Bayesian terms, this analysis sets the distribution of a Gaussian process as the prior distribution of the unknown function. The posterior is still Gaussian, but conditioned on the values of the function observed at specific points of \( \mathbb{R}^r \). In order to show the performance of such models, upper bounds for posterior rates have been obtained by van der Vaart and van Zanten (2008) and lower bounds by Castillo (2008).

The Gaussian prior distribution is usually parametric, with the parameters typically set to the maximum likelihood estimator. Another approach is hierarchical Bayesian modeling, where a prior on the parameters of the prior Gaussian process distribution is defined. This approach is extensively discussed in Chapter 6 of Banerjee, Carlin and Gelfand (2004).

Given prior knowledge about the parameters of the Gaussian prior is often lacking, there is value in objective elicitation of their prior distribution. However, De Oliveira, Kedem and Short (1997) and Stein (1999) noted that commonly used noninformative priors sometimes failed to yield proper posteriors. Berger, De Oliveira and Sansó (2001) were the first to thoroughly investigate the issue. Among several prior distributions – truncated priors,  

MSC 2010 subject classifications: Primary 62F15; secondary 62M30, 60G15  
Keywords and phrases: Gaussian process, Kriging, correlation function, Jeffreys prior, reference prior, integrated likelihood, posterior propriety
vague priors, Jeffreys-rule and independence Jeffreys prior – they showed that the reference prior (Bernardo, 2005) is the most satisfying choice for a default prior distribution. This is due to the fact that, for all isotropic correlation kernels studied by Berger, De Oliveira and Sansó (2001), the reference prior yields a proper posterior distribution. In the following, whenever we mention a “prior”, it is the prior on the parameters of the Gaussian process distribution, not the Gaussian process distribution itself.

Posterior propriety is necessary to Bayesian procedures aiming to quantify parameter uncertainty: it would make no sense to express parameter uncertainty through a posterior distribution that does not integrate to a finite mass. Moreover, in spatial models, one may want to take parameter uncertainty into account when performing prediction. Prediction can be averaged over a proper posterior distribution on parameters, not an improper one.

In this article, we complete the proof by Berger, De Oliveira and Sansó (2001) that the reference posterior is proper. Because of the difficulty involved in obtaining a satisfying default prior distribution which consistently yields a proper posterior, it is important to ascertain that the reference prior actually does. Indeed, a vast literature builds upon this result.

Provided some additional conditions on the design set and the mean function are verified, Paulo (2005) states that the reference posterior is also proper with anisotropic product correlation kernels that have specific properties. As Paulo (2005) notes, these properties are also necessary to make the proof from Berger, De Oliveira and Sansó (2001) work. He then warns the reader that the Squared Exponential kernel (Power Exponential with power $\rho = 2$) does not satisfy one of them.

Ren, Sun and He (2012) investigate the propriety of reference posteriors for isotropic correlation kernels with an additional noise term (nugget effect). They show that the reference prior is not the same for different parameter orders, but that all reference priors lead to a proper posterior. However, the correlation kernel must have the same properties as those required by Berger, De Oliveira and Sansó (2001).

Kazianka and Pilz (2012) consider the same setting and focus on the reference prior with a particular parameter order. They provide an application to the analysis of zinc concentrations in the French river Meuse.

Ren, Sun and Sahu (2013) return to the setting explored by Paulo (2005) – anisotropic product correlation kernels – but relax his conditions on the mean function.

Gu, Wang and Berger (2018) go further still and also relax his conditions on the design set. Moreover, they establish reference posterior propriety with
anisotropic product covariance kernels and an additional noise term (nugget effect). Their work thus generalizes and combines results from Paulo (2005) and Ren, Sun and Sahu (2013) on one side, Ren, Sun and He (2012) and Kazianka and Pilz (2012) on the other.

However, all these papers prove reference posterior propriety under similar assumptions as those made by Berger, De Oliveira and Sansó (2001). They rely on properties that only rough correlation kernels possess.

The main result of this work is Theorem 4.4, which ensures that the reference prior leads to a proper posterior distribution for a large class of smooth isotropic kernels. Secondary results include bounds on the reference prior density and on the function obtained after integrating several parameters out of the likelihood.

The paper is organized as follows. Section 2 describes the Gaussian Process models considered by Berger, De Oliveira and Sansó (2001). Section 3 shows that the proof of the propriety of the reference posterior provided by Berger, De Oliveira and Sansó (2001) only applies to Gaussian Process models with rough correlation kernels – Spherical, Exponential with power \( \rho < 2 \), Matérn with smoothness \( \nu < 1 \). Section 4 provides bounds on reference prior and integrated likelihood which are then used to prove that the reference posterior is also proper for models with smoother correlation kernels, including Exponential kernels with power \( \rho = 2 \), Matérn kernels with smoothness \( \nu \geq 1 \) and Rational Quadratic kernels. While Section 4 provides bounds applying to all cases, tighter bounds applying to specific cases only are obtained in Appendix B.

2. Setting. Berger, De Oliveira and Sansó (2001) consider Gaussian Process models (also known as Universal Kriging models) with isotropic kernels. This article is set in their framework and borrows most of its notations from them. Define \( \| \cdot \| \) as the usual Euclidean norm if applied to a vector and as the Frobenius norm if applied to a matrix. We denote integer intervals by \([ \cdot, \cdot ]:\ [1, 4]\) for instance is the set \{1, 2, 3, 4\}.

In Universal Kriging, an unknown function from a domain \( \mathcal{D} \subset \mathbb{R}^r (r \in \mathbb{Z}_+) \) to \( \mathbb{R} \) is assumed to be a realization of a Gaussian process \( Y \). The mean function \( f \) of the Gaussian process is assumed to belong to some known vector space \( \mathcal{F}_p \) of dimension \( p \in \mathbb{N} \). If \( p \) is non-zero, once a basis \( \{ f_j \}_{j \in [1,p]} \) of \( \mathcal{F}_p \) has been set, \( f \) can be parametrized by \( \beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p \) such that \( f = \sum_{j=1}^p \beta_j f_j \).
$Y - f$ is assumed to be a stationary Gaussian process based on an isotropic correlation kernel. An isotropic correlation kernel is a function $K : [0, +\infty) \to [-1, 1]$ such that for any positive integer $n$ and any collection of $n$ distinct points $(x^{(i)})_{i \in \{1, \ldots, n\}}$ within $D$, the symmetric $n \times n$ matrix $\Sigma$ with $(i, i')$-th element $K(||x^{(i)} - x^{(i')}||)$ is a positive definite correlation matrix. Necessarily, $K(0) = 1$.

The covariance function of the Gaussian process $Y$ is $\sigma^2 K_\theta$, where $K_\theta$ is the correlation kernel parametrized by $\theta \in (0, +\infty)$ and defined by $K_\theta(t) = K(\cdot/\theta)$, making $\sigma^2 \in (0, +\infty)$ the variance of $Y(x)$ for every $x \in D$.

Fix $n \in \mathbb{Z}_+$ and fix a collection of $n$ distinct points $(x^{(i)})_{i \in \{1, \ldots, n\}}$. Let this collection be the design set, i.e. the set of points where $Y$ is observed. $(Y(x^{(1)}), \ldots, Y(x^{(n)}))^\top$ is a Gaussian vector. $(f(x^{(1)}), \ldots, f(x^{(n)}))^\top$ is its mean vector and $\sigma^2 \Sigma_\theta$ its covariance matrix, with $\Sigma_\theta$ being the $n \times n$ matrix with $(i, i')$-th element $K_\theta(||x^{(i)} - x^{(i')}||)$. Table 1 provides the definition of several correlation kernels.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>$K_\theta(t)$ (with $t \in (0, +\infty)$)</th>
<th>Parameter range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spherical</td>
<td>$(1 - \frac{2}{\nu} \left(\frac{t}{\nu}\right) + \frac{1}{\nu} \left(\frac{t}{\nu}\right)^{\nu}) 1_{t \leq \nu}$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>Power Exponential</td>
<td>$\exp{-\frac{1}{\rho} t}$</td>
<td>$\rho \in (0, 2]$</td>
</tr>
<tr>
<td>Rational Quadratic</td>
<td>$\left(1 + \frac{\nu}{\nu(t)}\right)^{-\nu}$</td>
<td>$\nu \in (0, +\infty)$</td>
</tr>
<tr>
<td>Matérn</td>
<td>$\Gamma(\nu)^{-1} 2^{1-\nu} (2\sqrt{\nu} t)^{-\nu} K_\nu(2\sqrt{\nu} t)$</td>
<td>$\nu \in (0, +\infty)$</td>
</tr>
</tbody>
</table>

Table 1 Formulas for several correlation kernel families. The Spherical kernel can only be used if the design space is of dimension 1, 2 or 3, because it is not positive definite for greater dimensions (Stein, 1999). $K_\nu$ is the modified Bessel function of second kind with parameter $\nu$ (Abramowitz and Stegun, 1964, section 9.6). This parametrization of the Matérn family is recommended by Handcock and Wallis (1994). To recover the one used by Berger, De Oliveira and Sansó (2001), simply replace $2\sqrt{\nu} t$ by $t$.

If $p$ is non-zero, let $H$ denote the $n \times p$ matrix with $(i, j)$-th element $f_j(x^{(i)})$. Then $(f(x^{(1)}), \ldots, f(x^{(n)}))^\top = H \beta$. If $p = 0$, then we adopt the convention that any term involving $H$ can be ignored. Let us denote the observed value of the random vector $(Y(x^{(1)}), \ldots, Y(x^{(n)}))^\top$ by $y = (y_1, \ldots, y_n)^\top$. The likelihood function of the parameter triplet $(\beta, \sigma^2, \theta)^\top$ has the following expression:

$$L(y \mid \beta, \sigma^2, \theta) = \left(\frac{1}{2\pi \sigma^2}\right)^{\frac{n}{2}} |\Sigma_\theta|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} (y - H \beta)^\top \Sigma_\theta^{-1} (y - H \beta)\right\}.$$ (2.1)
In order for the model to be identifiable, assume that \( p < n \) and that \( H \) has rank \( p \).

Let us recall the general definition of the reference prior and how it is derived in this setting.

For smooth one-dimensional parametric families, the reference prior coincides with the Jeffreys-rule prior (Clarke and Barron, 1994). For smooth finite-dimensional parametric families, the reference prior algorithm requires the user to define groups of dimensions of the parameter and rank them. The reference prior is then defined iteratively (Bernardo, 2005):

1. Compute the Jeffreys-rule prior on the lowest-ranking group of dimensions conditionally on all others.
2. Average the likelihood function over this prior.
3. Compute the Jeffreys-rule prior (based on the integrated likelihood function) on the second-lowest-ranking group of dimensions conditionally on all higher-ranking dimensions.
4. Average the integrated likelihood function over this second prior.
5. Continue the process until the Jeffreys-rule prior on the highest-ranking group of dimensions has been computed.
6. The reference prior is defined as the product of all successively computed priors.

Berger, De Oliveira and Sansó (2001) treat \( \beta \) as the lowest-ranking (possibly multidimensional) parameter and view \((\sigma^2, \theta)\) as a group of 2 equal-ranking parameters when they apply the reference prior algorithm. We denote this ordering by \( \beta \prec (\sigma^2, \theta) \). This choice is not arbitrary; if \( \sigma^2 \) and \( \theta \) were known, then the covariance matrix \( \sigma^2 \Sigma_\theta \) would be known. Since the model is Gaussian, we would then know which linear transformation to apply to the observations to make them mutually independent. In the case of Simple Kriging, \( \beta \) is not used and their reference prior is simply the Jeffreys-rule prior. Note that \( \theta \) is one-dimensional because all correlation kernels considered by Berger, De Oliveira and Sansó (2001) - which are also those covered in the present article - are isotropic.

Ren, Sun and He (2012) show that if we split the parameters further, with \( \beta \) remaining the lowest-ranking parameter, \( \sigma^2 \) being of middle rank and \( \theta \) being the highest-ranking parameter, the reference prior algorithm yields the same reference prior. In other words, the reference prior for the ordering \( \beta \prec \sigma^2 \prec \theta \) is the same as the reference prior for the ordering \( \beta \prec (\sigma^2, \theta) \). It is this prior we consider in the present article.
To express it conveniently, denote $Q_\theta := I_n - H (H^\top \Sigma_\theta^{-1} H)^{-1} H^\top \Sigma_\theta^{-1}$. If $p = 0$, $Q_\theta := I_n$. Also fix $W$, an $n \times (n - p)$ matrix such that $W^\top W = I_{n-p}$ and $H^\top W$ is the $p \times (n - p)$ null matrix. The columns of $W$ form an orthonormal basis of the orthogonal complement of the subspace of $\mathbb{R}^n$ spanned by the columns of $H$. If $p = 0$, fix $W$ as an orthogonal matrix, for example $I_n$.

If $p > 0$, the matrix $W$ can for instance be constructed by computing a Singular Value Decomposition (SVD) of $H$. Let us write this decomposition $H = USV^\top$, $U$ and $V$ are orthogonal matrices of size $n \times n$ and $p \times p$ respectively, and $S$ is an $n \times p$ matrix whose only non-null entries are on the main diagonal. Therefore the last $n - p$ rows of $S$ are filled with zeros. The matrix $W$ can then be defined as the $n \times (n - p)$ matrix formed by the last $n - p$ columns of $U$.

The next two propositions give formulas for the reference prior density and for the integrated likelihood. The latter is obtained by integrating the likelihood against the reference prior on $\sigma^2$ and $\theta$ and against the reference prior on $\sigma^2$ conditionally on $\theta$.

**Proposition 2.1.** The reference prior with ordering $\beta \prec (\sigma^2, \theta)$ or $\beta \prec \sigma^2 \prec \theta$ is $\pi(\beta, \sigma^2, \theta) \propto (\sigma^2)^{-1} \pi(\theta)$, where

\[
\pi(\theta) \propto \sqrt{\text{Tr} \left[ \left\{ \frac{d}{d\theta} \Sigma_\theta \right\} \Sigma_\theta^{-1} Q_\theta \right]^2 - \frac{1}{n-p} \left[ \text{Tr} \left\{ \left( \frac{d}{d\theta} \Sigma_\theta \right) \Sigma_\theta^{-1} Q_\theta \right\} \right]^2}.
\]

Denoting $W^\top \Sigma_\theta W$ by $\Sigma_\theta^W$, $\pi(\theta)$ can also be written as:

\[
\pi(\theta) \propto \sqrt{\text{Tr} \left[ \left\{ \frac{d}{d\theta} \Sigma_\theta^W \right\} (\Sigma_\theta^W)^{-1} \right]^2 - \frac{1}{n-p} \left[ \text{Tr} \left\{ \left( \frac{d}{d\theta} \Sigma_\theta^W \right) (\Sigma_\theta^W)^{-1} \right\} \right]^2}.
\]

The fact that (2.2) is an expression of the reference prior fitting the ordering $\beta \prec (\sigma^2, \theta)$ was first shown in Berger, De Oliveira and Sansó (2001). Ren, Sun and He (2012) then showed that the reference prior fitting the ordering $\beta \prec \sigma^2 \prec \theta$ is the same. The proof that (2.3) is an expression of the same reference prior can be found in Appendix A.1.
Proposition 2.1 shows that \( \pi(\theta) \) is the only untractable factor in the expression of the reference prior \( \pi(\beta, \sigma^2, \theta) \). The parameters \( \beta \) and \( \sigma^2 \) can be marginalized out of \( L(y|\beta, \sigma^2, \theta) \).

**Proposition 2.2.** If \( p \geq 1 \), after marginalizing \( \beta \) and \( \sigma^2 \) out, we have

\[
L_y(\theta) := \int \int L_y(\beta, \sigma^2, \theta) / \sigma^2 \, d\beta \, d\sigma^2
\]

(2.4)

\[
= \left( \frac{2\pi \frac{n-p}{2}}{\Gamma \left( \frac{n-p}{2} \right)} \right)^{-1} \left| \Sigma_\theta^{-1} \right|^{\frac{1}{2}} \left| H^\top \Sigma_\theta^{-1} H \right|^{-\frac{1}{2}} \left( y^\top \Sigma_\theta^{-1} Q_\theta y \right)^{-\frac{n-p}{2}}.
\]

Alternatively, the integrated likelihood with \( p \geq 1 \) can also be written

(2.5)

\[
L_y(\theta) = \left( \frac{2\pi \frac{n-p}{2}}{\Gamma \left( \frac{n-p}{2} \right)} \right)^{-1} \left| H^\top H \right|^{-\frac{1}{2}} \left| W^\top \Sigma_\theta W \right|^{-\frac{1}{2}} \left( y^\top W \left( W^\top \Sigma_\theta W \right)^{-1} W^\top y \right)^{-\frac{n-p}{2}}.
\]

If \( p = 0 \), the integrated likelihood is simply

(2.6)

\[
L_y(\theta) = \int L_y(\sigma^2, \theta) / \sigma^2 \, d\sigma^2 = \left( \frac{2\pi \frac{n}{2}}{\Gamma \left( \frac{n}{2} \right)} \right)^{-1} \left| \Sigma_\theta^{-1} \right|^{\frac{1}{2}} \left( y^\top \Sigma_\theta^{-1} y \right)^{-\frac{n}{2}}.
\]

The proof of this Proposition can be found in Appendix A.2.

Proving that the reference posterior is proper amounts to finding appropriate upper bounds on the tail rates of \( L_y(\theta) \pi(\theta) \) as \( \theta \to 0 \) and as \( \theta \to +\infty \).

3. **Scope of the original proof.** In Berger, De Oliveira and Sansó (2001), Lemmas 1 and 2 require that the correlation kernel and design set should be such that \( \Sigma_\theta = 11^\top + g_0(\theta) D + R_0(\theta) \), where \( 1 \) is the vector with \( n \) entries all equal to 1, \( g_0(\theta) \) is a real-valued function such that \( \lim_{\theta \to +\infty} g_0(\theta) = 0 \), \( D \) is a fixed nonsingular matrix and \( R_0 \) is a function from \((0, +\infty)\) to the set of \( n \times n \) real matrices \( M_n \) such that \( \lim_{\theta \to +\infty} \| \frac{1}{g_0(\theta)} R_0(\theta) \| = 0 \). Berger, De Oliveira and Sansó (2001) use this form to derive an asymptotic expansion of \( \Sigma_\theta^{-1} \) which involves \( D^{-1} \) (see Equations (B.4) and (B.5) of the paper; they are part of the proof of their Lemma 1, which is then used to prove their Theorem 4 which states that the reference posterior is proper).
Remark 1. In fact, Lemma 2 of Berger, De Oliveira and Sansó (2001) has additional hypotheses. The original statement specifies, among other things, that \( \Sigma_\theta = 11^\top + v(\theta)D + w(\theta)D^* + R(\theta) \), where \( D \) (nonsingular) and \( D^* \) are fixed matrices, \( R(\theta) \) is a matrix that depends on \( \theta \), and as \( \theta \to +\infty \), \( w(\theta)/v(\theta) \to 0 \) and \( \|R(\theta)\|/v(\theta) \to 0 \). Additional assumptions on the derivatives of \( v \) and \( w \) are also made. Informally, the goal of all these assumptions is to make sure that when \( \theta \to +\infty \), for the purposes of their proof, \((11^\top + v(\theta)D)^{-1}\) is close enough to \( \Sigma_\theta^{-1} \) and \( v'(\theta)D + w'(\theta)D^* \) is close enough to \( \frac{\partial}{\partial \theta} \Sigma_\theta \). The framework of the present article does not make the assumption that \( D \) is nonsingular, which means that an asymptotic expansion of \( \Sigma_\theta \) at higher orders is required and that neither \( D \) nor \( D^* \) plays any particular role.

The following results make it clear that \( D \) is singular in practically relevant situations.

**Proposition 3.1.** For Spherical, Power Exponential, Rational Quadratic and Matérn correlation kernels, the correlation matrix \( \Sigma_\theta \) can be expressed as

\[
\Sigma_\theta = 11^\top + g_0(\theta)D^{(q)} + R_0(\theta),
\]

where:

- \( D^{(q)} \) is the matrix with \((i, i')\)-th element \( \|x^{(i)} - x^{(i')}\|^{2q} \), with \( q \) given in Table 2;
- \( g_0 \) is a real-valued function given in Table 2;
- \( R_0 \) is a function from \((0, +\infty)\) to the set of \( n \times n \) real matrices \( \mathcal{M}_n \) such that \( \lim_{\theta \to +\infty} \| \frac{1}{g_0(\theta)} R_0(\theta) \| = 0 \).

The proof of this Proposition can be found in Appendix A.3.

**Theorem 4** from Schoenberg (1937) has the following corollary:

**Proposition 3.2.** For \( q \in (0, 1) \), the matrix \( D^{(q)} \) defined in Proposition 3.1 is nonsingular.

The picture is dramatically different when the correlation kernel \( K \) is smooth enough to have \( q = 1 \). This happens as soon as \( K \) is twice continuously differentiable. Theorem 6 from Gower (1985) implies the following result:
For $q = 1$, the matrix $D^{(q)} = D^{(1)}$ defined in Proposition 3.1 has rank lower or equal to $r + 2$.

If $n \leq r + 2$, then Proposition 3.3 becomes trivial. For all practical purposes however, $n$ is much greater than $r$ and $D^{(1)}$ is singular.

Thanks to Propositions 3.2 and 3.3, Proposition 3.1 yields this result:

Proposition 3.4. In the decomposition of $\Sigma_0$ given in Proposition 3.1:

- if the correlation kernel is Spherical, Power Exponential with $\rho < 2$ or Matérn with $\nu < 1$, then $q < 1$ and $D^{(q)}$ is nonsingular;
- if the correlation kernel is Squared Exponential (i.e. Power Exponential with $\rho = 2$), Rational Quadratic or Matérn with $\nu \geq 1$, then $q = 1$;
- if $n > r + 2$ and the correlation kernel is Squared Exponential, Rational Quadratic or Matérn with $\nu \geq 1$, then $D^{(q)} = D^{(1)}$ is singular.

Proposition 3.4 justifies the claim in the abstract that the Squared Exponential kernel, Matérn kernels with smoothness $\nu \geq 1$ and Rational Quadratic kernels require a proof of the reference posterior’s propriety.

4. Propriety of the reference posterior distribution. As shown by Berger, De Oliveira and Sansó (2001), the reference posterior distribution of $\beta$ and $\sigma^2$ conditionally on $\theta$ is proper.

In this section, we prove that the joint reference posterior distribution is proper for Matérn kernels with smoothness $\nu \geq 1$, Rational Quadratic kernels and the Squared Exponential kernel.
**Proposition 4.1.** For Matérn kernels with smoothness $\nu \geq 1$, for Rational Quadratic kernels with parameter $\nu > 0$ and for the Squared Exponential kernel, the “marginal” reference prior distribution $\pi(\theta)$ defined by Proposition 2.1 has the following behavior.

1. When $\theta \to 0$,
   \[
   \pi(\theta) = \begin{cases} 
   o(1) & \text{for Matérn and Squared Exponential kernels;} \\
   O(\theta^{2\nu-1}) & \text{for Rational Quadratic kernels.}
   \end{cases}
   \]

2. When $\theta \to +\infty$,
   \[
   \pi(\theta) = O(\theta^{-1}).
   \]

**Remark 2.** The second assertion of Proposition 4.1 is not strong enough to make the reference prior proper because $\int_{1}^{+\infty} \theta^{-1} d\theta = +\infty$. The first assertion, however, implies that if we truncated the reference prior at some value $T > 0$, only taking $(0, T)$ as its support, the resulting prior would be proper. The choice of $T$ would be very informative, though.

The proof of Proposition 4.1 can be found in Appendix A.4.

Figure 1 shows the tail rate of the reference prior for Matérn kernels with smoothness $\nu = 1.5$ and 2.5 for the following Kriging models:

- Simple Kriging: the mean function is assumed to be null;
- Ordinary Kriging: the mean function is assumed to be an unknown constant: $H$ is $1$, the vector of $\mathbb{R}^n$ filled with ones;
- Affine Kriging: the mean function is assumed to be an unknown affine function: $H$ is the $n \times (r+1)$ matrix whose first column is $1$ and whose last $r$ columns contain the coordinates of the points in the design set.

Figure 1 is consistent with the bound for tail rates given in Proposition 4.1.

For rough kernels (Spherical, Power Exponential with $\rho < 2$, Matérn with $\nu < 1$), Berger, De Oliveira and Sansó (2001) prove that if $1$ belongs to the vector space spanned by the columns of $H$ (as happens in Ordinary and Affine Kriging), then the reference prior is proper.

The upper bound provided by Proposition 4.1, $O(\theta^{-1})$ can be tightened in some cases. For example, in the case of Affine Kriging with a Matérn kernel with smoothness $\nu = 1.5$ (figure 1c), the prior is bounded by $O(\theta^{-2})$. See Proposition B.1 in Appendix B for a proof.
Fig 1: Logarithm of the reference prior density for large values of $\theta$ for varying Kriging models and correlation kernels. In each case, the design space is one-dimensional ($r = 1$) and the design set is the 10-point regular grid on $[0,1]$.

The following lemma deals with the asymptotic behavior of the integrated likelihood when $\theta \to +\infty$.

Let $v_1(\theta) \geq ... \geq v_{n-p}(\theta) > 0$ be the ordered eigenvalues of $W^T \Sigma_{\theta} W$.

**Lemma 4.2.** For Rational Quadratic and Squared Exponential kernels and for Matérn kernels with smoothness $\nu \geq 1$, there exists a hyperplane $H_n$ of $\mathbb{R}^n$ such that for every $y \in \mathbb{R}^n \setminus H_n$, when $\theta \to +\infty$:

$$(y^T W \left(W^T \Sigma_{\theta} W\right)^{-1} W^T y)^{-1} = O(v_{n-p}(\theta)).$$

In particular, in the case of Simple Kriging, $W$ is the identity matrix and $p = 0$:

$$(y^T \Sigma_{\theta}^{-1} y)^{-1} = O(v_n(\theta)).$$

The proof of this lemma can be found in Appendix A.5.
For \( \mathbf{y} \notin \mathcal{H}_n \), Lemma 4.2 shows that the smallest eigenvalue \( \nu_{n-p}(\theta) \) of \( \mathbf{W}^\top \mathbf{\Sigma}_n \mathbf{W} \) is an asymptotic upper bound for one of the factors of the squared integrated likelihood \( L_y(\theta)^2 \) when \( \theta \to +\infty \). We need to take the other factors into account in order to obtain an asymptotic upper bound of the squared integrated likelihood. Combined with Proposition 2.2, Lemma 4.2 implies the following result.

**Proposition 4.3.** If the observation vector \( \mathbf{y} \) belongs to \( \mathbb{R}^n \setminus \mathcal{H}_n \) (\( \mathcal{H}_n \) is defined in Lemma 4.2), then

\[
L_y(\theta)^2 = \prod_{i=1}^{n-p} \frac{O(\nu_{n-p}(\theta))}{\nu_i(\theta)} \quad \text{when} \quad \theta \to +\infty.
\]

Notice that all factors in the product (4.5) are \( O(1) \) when \( \theta \to +\infty \). For \( \mathbf{y} \notin \mathcal{H}_n \), Proposition 4.3 thus implies the existence of a constant asymptotic upper bound for the likelihood \( L_y(\theta) \). It does not provide any tighter explicit upper bound, and this bound can only guarantee posterior propriety if the prior is itself proper (cf. Remark 2). However, it provides an implicit upper bound that will be used in proofs of posterior propriety. For \( \mathbf{y} \notin \mathcal{H}_n \), when \( \theta \to +\infty \), any upper bound of the ratio \( \frac{\nu_{n-p}(\theta)}{\nu_i(\theta)} \) is an upper bound of the squared likelihood \( L_y(\theta)^2 \).

Figure 2 in Appendix B gives an idea of the explicit tail rates of the marginal log-likelihood in several situations.

**Theorem 4.4.** For Rational Quadratic kernels and Squared Exponential kernels and for any Matérn kernel with smoothness \( \nu \geq 1 \), there exists a hyperplane \( \mathcal{H}_n \) of \( \mathbb{R}^n \) such that if \( \mathbf{y} \) belongs to \( \mathbb{R}^n \setminus \mathcal{H}_n \), then the reference posterior distribution \( \pi(\theta|\mathbf{y}) \) is proper.

**Remark 3.** The requirement that \( \mathbf{y} \) should belong to \( \mathbb{R}^n \setminus \mathcal{H}_n \) is not substantially restrictive. If \( \mathbf{y} \) is actually sampled from a nondegenerate Gaussian process, it almost surely does not belong to the hyperplane \( \mathcal{H}_n \). Therefore, if \( \mathbf{y} \) actually does belong to \( \mathcal{H}_n \), it might be better explained by a degenerate Gaussian model. The most compelling example is that of a constant observation vector, for which the Kriging model would be grossly inappropriate. In case there is some doubt about whether a given \( \mathbf{y} \) belongs to \( \mathbb{R}^n \setminus \mathcal{H}_n \), Appendix A.7 provides explicit sufficient conditions instead.

The proof of the Theorem can be found in Appendix A.6. Depending on the matrix \( \mathbf{H} \) (or its absence in the case of Simple Kriging), the bounds of
the likelihood function may differ. In some cases, it is even possible to derive a tighter bound for the prior than the general one given in Proposition 4.1. Examples are given in Appendix B: Affine Kriging with a Matérn kernel with smoothness $\nu \in [1, 2)$ (Propositions B.1 and B.2) and Ordinary Kriging with a Matérn kernel with smoothness $\nu = 1$ (Proposition B.3).

5. Conclusion. In this work, we proved that for a large class of smooth kernels, the reference prior leads to a proper posterior distribution. This class contains the Squared Exponential correlation kernel as well as the important Matérn family (Stein, 1999) with smoothness parameter $\nu \geq 1$. The seldom used Rational Quadratic kernels are also included within this class.

Berger, De Oliveira and Sansó (2001) proved this result for a class of rough correlation kernels. This class includes the complementary set of the Matérn family – kernels with smoothness parameter $\nu < 1$ – as well as all other Power Exponential kernels. Spherical kernels, which are mostly used in the field of geostatistics also belong to this class.

The results from Berger, De Oliveira and Sansó (2001), together with Theorem 4.4, show how polyvalent the reference prior is, insofar as it is able to adapt to very different correlation kernels and always leads to a proper posterior. The key to this flexibility is the way the tail rate of the reference prior adapts to the tail rate of the integrated likelihood, which depends on the correlation matrix $\Sigma_\theta$ and the trend matrix $H$. For rough correlation kernels, Berger, De Oliveira and Sansó (2001) were able to express likelihood tail rates as an explicit function of $\theta$. More research is needed to do the same for smoother correlation kernels, even though Appendix B provides explicit tail rates in a few specific cases.

The reference prior’s flexibility means no ad-hoc technique is required to derive useable inference. This makes the approach appealing from a Bayesian point of view when no explicit prior information is available. Even when explicit prior information is available, following Druiijet and Marin (2007), the reference prior can be used to derive maximum a posteriori (MAP) estimates or High Probability Density (HPD) sets that are invariant under reparametrization.

In Section 3, we recalled that the original proof from Berger, De Oliveira and Sansó (2001) relies on the matrix $D^{(q)}$ from the Taylor expansion of the correlation matrix being nonsingular (cf. Proposition 3.1). Provided there are enough observation points, this implies that $q < 1$. Stein (1999) shows
that the correlation kernel cannot then be twice continously differentiable at 0 (p. 28, section “Principal irregular term”).

Since they make the same assumption, this restriction to rough correlation kernels also applies to all works generalizing the original result from Berger, De Oliveira and Sansó (2001). Gu, Wang and Berger (2018) provide the largest generalization to date: their setting admits anisotropic kernels defined as products of one-dimensional kernels and a possible additional noise term (nugget effect). It does not admit anisotropic geometric kernels, however (see Table 3 for a definition). They prove that one of the reference priors leads to a proper posterior: the prior derived from the reference prior algorithm where \( \beta \) is the lower-ranking parameter and \( \sigma^2 \), \( \theta = (\theta_1, \ldots, \theta_r) \) and possibly the parameter controlling the nugget effect are collectively the higher-ranking group of parameters. Like Berger, De Oliveira and Sansó (2001), they assume that the matrix \( D^{(q)} \) from Proposition 3.1 is nonsingular regardless of \( q \), so their proof applies to products of rough correlation kernels (Spherical, Power Exponential with \( \rho < 2 \) and Matérn with \( \nu < 1 \)). Unfortunately, the proof used in the present article to deal with smoother kernels (Rational Quadratic, Squared Exponential and Matérn with \( \nu \geq 1 \)) cannot easily be adjusted to their setting. The corresponding reference prior is indeed much more complex as it is proportional to the square root of the Fisher information matrix of either \( r + 1 \) parameters \( (\sigma, \theta_1, \ldots, \theta_r) \) or \( r + 2 \) parameters if there is a nugget effect.

\[
\begin{array}{|c|c|c|}
\hline
\text{Correlation lengths} & \text{Product} & \text{Anisotropic geometric} \\
\theta = (\theta_1, \ldots, \theta_r) & \prod_{i=1}^{r} K_{\theta_i}(t_i) & K_1 \left( \sum_{i=1}^{r} \frac{t_i^2}{\sigma^2} \right)^{1/2} \\
\hline
\end{array}
\]

Table 3

Anisotropic correlation kernels. \( K_\theta \) is a 1-dimensional kernel with correlation length \( \theta \).

Whenever the reference posterior is known to be proper – whether in the case of isotropic correlation kernels as shown in the present article or in the case of a product of rough correlation kernels as shown in Gu, Wang and Berger (2018) – it is theoretically possible to propagate parameter uncertainty to the predictions of the Gaussian process model. This can be done through Markov Chain Monte-Carlo (MCMC) sampling of the marginal reference posterior distribution on \( \theta \). The spread of the predictions obtained using the different values of \( \theta \) account for parameter uncertainty: not only the uncertainty on \( \theta \), but also on \( \beta \) and \( \sigma^2 \) since the latter can be marginalized out of the model (cf. Proposition 2.2). In practice though, Gu, Wang and Berger (2018) do not advocate this method because of the computational cost of MCMC in this
setting. They use the maximum a posteriori (MAP) estimate for $\theta$ instead. This effectively means that the density of the reference posterior distribution acts as a penalization factor on the likelihood. An alternative proposal is to sample the $\theta_i$ ($1 \leq i \leq r$) from one-dimensional reference posterior distributions in order to make MCMC tractable (Muré, 2019).

Acknowledgements. The author would like to thank his PhD advisor Professor Josselin Garnier (École Polytechnique, Centre de Mathématiques Appliquées) for his guidance, Loic Le Gratiet (EDF R&D, Chatou) and Anne Dutfoy (EDF R&D, Saclay) for their advice and helpful suggestions. He also thanks the editor, the associate editor and the referees for their comments which substantially improved this article. The author acknowledges the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-13-MONU-0005 (project CHORUS).
APPENDIX A: PROOFS

The proofs presented in this Appendix rely on auxiliary facts from Appendix C. Appendix C can be found in supplementary material.

A.1. Proof of Proposition 2.1.

Proof. Both assertions follow from results proved by Ren, Sun and He (2012). The first follows from their Proposition 5 (a) (taken in the particular case with no nugget effect). The second is a consequence of Lemma C.1, which restates their Lemma 6: it implies that

(A.1) \[ W (W^\top \Sigma \theta W)^{-1} W^\top = \Sigma_\theta^{-1} \left( I_n - H \left( H^\top \Sigma_\theta^{-1} H \right)^{-1} H^\top \Sigma_\theta^{-1} \right) . \]

\[ \square \]

A.2. Proof of Proposition 2.2.

Proof. The result for \( p = 0 \) and the first result for \( p \geq 1 \) are from Berger, De Oliveira and Sansó (2001).

From (A.1), all that remains to be proved is the determinant equality \( |\Sigma_\theta| = |W^\top \Sigma_\theta W| |H^\top H| |H^\top \Sigma_\theta^{-1} H|^{-1} \). Choose an \( n \times p \) matrix \( P \) with columns forming an orthonormal basis of the \( p \)-dimensional subspace of \( \mathbb{R}^n \) spanned by the columns of \( H \). Let \( (WP) \) denote the matrix whose left \( n \times (n-p) \) block is \( W \) and whose right \( n \times p \) block is \( P \). \( (WP) \) is an \( n \times n \) orthogonal matrix, so we have \( |\Sigma_\theta| = |(WP)^\top \Sigma_\theta (WP)| \). Using Schur’s complement (see for example Serre (2002) p. 139),

(A.2) \[ |\Sigma_\theta| = \left| W^\top \Sigma_\theta W \right| \left| P^\top \Sigma_\theta \left( I_n - W \left( W^\top \Sigma_\theta W \right)^{-1} W^\top \Sigma_\theta \right) P \right| . \]

Equation (A.1) is equivalent to:

(A.3) \[ \Sigma_\theta \left( I_n - W \left( W^\top \Sigma_\theta W \right)^{-1} W^\top \Sigma_\theta \right) = H \left( H^\top \Sigma_\theta^{-1} H \right)^{-1} H^\top . \]

Plugging this in Equation (A.2), we obtain:
\[ |\Sigma_\theta| = \left| W^T \Sigma_\theta W \right| \left| P^T H \right|^2 \left| H^T \Sigma_\theta^{-1} H \right| \]

\[ = \left| W^T \Sigma_\theta W \right| \left| H^T \frac{PP^T}{I_p} H \right| \left| H^T \Sigma_\theta^{-1} H \right|. \]

(A.4)


**Proof.** Let us consider the \((i,i')\)-th element of the matrix \(\Sigma_\theta\). Letting \(K_\theta\) denote the correlation kernel, it is given by \(K_\theta \left( x^{(i)} - x^{(i')} \right)\).

1. With a spherical kernel,

\[ K_\theta \left( x^{(i)} - x^{(i')} \right) := 1 - \frac{3}{2} \left| \frac{x^{(i)} - x^{(i')}}{\theta} \right|^3 + 1 \left| \frac{x^{(i)} - x^{(i')}}{\theta^3} \right|^3. \]

We can identify \(g_0(\theta)\) as \(-\frac{3}{2} \theta^{-1}\), \(q\) as \(1/2\) and \(R_0(\theta)\) as the matrix whose \((i,j)\)-th element is \(\frac{1}{2} \left| \frac{x^{(i)} - x^{(j')}}{\theta^3} \right|^3\).

2. With a Power Exponential kernel,

\[ K_\theta \left( x^{(i)} - x^{(i')} \right) := \exp \left( -\frac{\|x^{(i)} - x^{(i')}\|^\rho}{\theta^\rho} \right). \]

When \(\theta \to +\infty\),

\[ K_\theta \left( x^{(i)} - x^{(i')} \right) = 1 - \frac{\|x^{(i)} - x^{(i')}\|^\rho}{\theta^\rho} + O \left( \theta^{-2\rho} \right). \]

We can identify \(g_0(\theta)\) as \(-\theta^{-\rho}\) and \(q\) as \(\rho/2\). Then \(\|R_0(\theta)\| = O(\theta^{-2\rho})\) when \(\theta \to \infty\).

3. With a Rational Quadratic kernel,
\[ K_{\theta}(x^{(i)} - x^{(i)'}) := \left( 1 + \frac{\|x^{(i)} - x^{(i)'}\|^2}{\theta^2} \right)^{-\nu}. \]

When \( \theta \to +\infty \),

\[ K_{\theta}(x^{(i)} - x^{(i)'}) = 1 - \nu \frac{\|x^{(i)} - x^{(i)'}\|^2}{\theta^2} + O(\theta^{-4}). \]

We can identify \( g_0(\theta) \) as \(-\nu\theta^{-2}\) and \( q \) as 1. Then \( \|R_0(\theta)\| = O(\theta^{-4}) \) when \( \theta \to \infty \).

4. With a Matérn kernel, we only need to refer to the appropriate decomposition of \( \Sigma_{\theta} \) in Appendix C.4 (Lemma C.15 if \( \nu \) is no integer and Lemma C.17 if \( \nu \) is an integer) to identify \( g_0(\theta) \), \( D^{(q)} \) and obtain the relevant properties for \( R_0(\theta) = \Sigma_{\theta} - \mathbf{1} \mathbf{1}^\top - g_0(\theta) D^{(q)} \).

- If \( \nu < 1 \):
  - \( g_0(\theta) = \Gamma(-\nu)\nu^\nu/(\Gamma(\nu)\theta^{2\nu}) \);
  - \( q = \nu \);
  - \( \|R(\theta)\| = O(\theta^{-2}) = o(g_0(\theta)) \) when \( \theta \to \infty \).

- If \( \nu = 1 \):
  - \( g_0(\theta) = -2 \log(\theta)/\theta^2 \);
  - \( q = 1 \);
  - \( \|R(\theta)\| = O(\theta^{-2}) = o(g_0(\theta)) \) when \( \theta \to \infty \).

- If \( \nu > 1 \):
  - \( g_0(\theta) = -\nu(\nu - 1)^{-1}/\theta^2 \);
  - \( q = 1 \);
  - \( \|R(\theta)\| = O(\theta^{-2\min(2, \nu)}) = o(g_0(\theta)) \) when \( \theta \to \infty \).

Proof. When $\theta \to 0$, $\Sigma_\theta$ converges to $I_n$, so its inverse does too. Therefore, in order to prove the first assertion, it is enough to prove that $\| \frac{d}{d\theta} \Sigma_\theta \| = o(1)$ for Matérn ($\nu \geq 1$) and Squared Exponential kernels and that $\| \frac{d}{d\theta} \Sigma_\theta \| = O(\theta^{2\nu-1})$ for Rational Quadratic kernels. To do this, we prove that these bounds hold for every element of the matrix $\frac{d}{d\theta} \Sigma_\theta$.

Letting $K_\theta$ be one of the considered kernels, the $(i, i')$-th element of $\frac{d}{d\theta} \Sigma_\theta$ is given by $\frac{d}{d\theta} K_\theta \left( \| x^{(i)} - x^{(i')} \| \right)$.

If $K_\theta$ is Squared Exponential, $\lim_{\theta \to 0} \frac{d}{d\theta} K_\theta \left( \| x^{(i)} - x^{(i')} \| \right) = 0$. This also holds if $K_\theta$ is a Matérn kernel with smoothness $\nu \geq 1$ (see Abramowitz and Stegun (1964) 9.6.28. and 9.7.2.).

If $K_\theta$ is a Rational Quadratic kernel with parameter $\nu > 0$, then when $\theta \to 0$, $\frac{d}{d\theta} K_\theta \left( \| x^{(i)} - x^{(i')} \| \right) / \theta^{2\nu-1}$ admits a finite limit.

Let us prove the second assertion.

Lemma C.20 shows that for Matérn kernels with smoothness $\nu$, for all $\theta > 0$ and all $\xi \in \mathbb{R}^n$:

\begin{equation}
0 \leq \xi^\top \left( r\theta^{-1} \Sigma_\theta - \frac{d}{d\theta} \Sigma_\theta \right) \xi \leq (2\nu + r)\theta^{-1} \xi^\top \Sigma_\theta \xi. \tag{A.10}
\end{equation}

Because of this, Lemma C.3 yields an upper bound on the reference prior density:

\begin{equation}
\sqrt{\operatorname{Tr} \left[ \left\{ \left( \frac{d}{d\theta} \Sigma_\theta \right) \Sigma_\theta^{-1} Q_\theta \right\}^2 \right]} - \frac{1}{n - p} \left[ \operatorname{Tr} \left\{ \left( \frac{d}{d\theta} \Sigma_\theta \right) \Sigma_\theta^{-1} Q_\theta \right\} \right]^2 \leq (n - p)(2\nu + r)\theta^{-1}. \tag{A.11}
\end{equation}

For Squared Exponential and Rational Quadratic kernels, a similar proof is possible. Lemma C.19 implies that there exists a positive constant $C$ such that for large enough $\theta$, for all $\xi \in \mathbb{R}^n$,

\begin{equation}
0 \leq \xi^\top \left( \theta^{-1} \Sigma_\theta - \frac{d}{d\theta} \Sigma_\theta \right) \xi \leq C\theta^{-1} \xi^\top \Sigma_\theta \xi. \tag{A.12}
\end{equation}
Like in the Matérn case, Lemma C.3 shows that this implies an upper bound on the reference prior density:

\[ \sqrt{\text{Tr} \left( \left\{ \left( \frac{d}{d\theta} \Sigma_{\theta} \right) \Sigma^{-1}_{\theta} Q_{\theta} \right\}^2 \right)} \leq \frac{1}{n-p} \text{Tr} \left( \left\{ \left( \frac{d}{d\theta} \Sigma_{\theta} \right) \Sigma^{-1}_{\theta} Q_{\theta} \right\}^2 \right) \leq (n-p) C \theta^{-1}. \]

(A.13)

A.5. Proof of Lemma 4.2.

Proof. The kernel of a matrix \( M \) is denoted by \( \text{Ker}(M) \).

For Rational Quadratic and Squared Exponential kernels, Lemma C.13 provides an asymptotic expansion of \( \Sigma_{\theta} \). When \( \theta \) is large enough,

\[ \Sigma_{\theta} = \sum_{k=0}^{\infty} \frac{a_k}{\theta^{2k}} D^{(k)}. \]

(A.14)

In Equation (A.14), for every \( k \), \( D^{(k)} \) is the \( n \times n \) matrix with \((i, i')\)-th element \( \| x^{(i)} - x^{(i')} \|^{2k} \) and \( a_k \) is a non-null real number that depends on the kernel.

Because \( \Sigma_{\theta} \) is nonsingular, the intersection \( \cap_{k=0}^{\infty} \text{Ker} D^{(k)} \) is the trivial vector space, i.e. the vector space containing only the null vector. This means there must exist (cf. Lemma C.4) a nonnegative integer \( k' \) such that the vector space \( \cap_{k=0}^{k'} \text{Ker} \left( W^\top D^{(k)} W \right) \) is trivial and such that the vector space \( \cap_{0 \leq k < k'} \text{Ker} \left( W^\top D^{(k)} W \right) \) is non-trivial (if \( k' = 0 \), the intersection is done over an empty index set, so we take it to be \( \mathbb{R}^{n-p} \) by convention).

This implies (cf. Lemma C.6) that for any \( y' \in \mathbb{R}^{n-p} \) that does not belong to the vector subspace \( A_{k'-1} \) spanned by the columns of the matrices \( W^\top D^{(k)} W \) \( (0 \leq k \leq k' - 1) \), there exists \( c_{y'} > 0 \) such that for large enough \( \theta \),

\[ (y')^\top \left( W^\top \Sigma_{\theta} W \right)^{-1} y' \geq c_{y'} \left\| \left( W^\top \Sigma_{\theta} W \right)^{-1} \right\|. \]

(A.15)
As a consequence, for every $y \in \mathbb{R}^n$ such that $W^\top y \notin A_{k'}$, there exists $c_y > 0$ such that for large enough $\theta$,

$$y^\top W \left( W^\top \Sigma_\theta W \right)^{-1} W^\top y \geq c_y \left\| \left( W^\top \Sigma_\theta W \right)^{-1} \right\|.$$  

(A.16)

Let $W A_{k'}$ denote the vector subspace of $\mathbb{R}^n$ of all vectors $v \in \mathbb{R}^n$ such that $W^\top v$ does belong to $A_{k'}$. Because the matrix $W^\top$ has full row rank, $W A_{k'}$ is included within a hyperspace $H_n$ of $\mathbb{R}^n$. Therefore, for every $y \in \mathbb{R}^n \setminus H_n$, there exists $c_y > 0$ such that for large $\theta$ the equation above holds.

For Matérn kernels with noninteger smoothness $\nu > 0$ (resp. with integer smoothness $\nu > 0$), Lemma C.15 (resp. Lemma C.17) allows a similar argument.

$$\Sigma_\theta = \sum_{k=0}^{[\nu]} \frac{a_k}{\theta^{2k}} D^{(k)} + \frac{a_\nu}{\theta^{2\nu}} D^{(\nu)} + R_\nu(\theta) \quad \text{if } \nu \text{ is noninteger.}$$  

(A.17)

$$\Sigma_\theta = \sum_{k=0}^{\nu-1} \frac{a_k}{\theta^{2k}} D^{(k)} + \tilde{a}_\nu \left( \log(\theta) D^{(\nu)} + \frac{1}{\theta^{2\nu}} \tilde{D}^{(\nu)} \right) + \tilde{R}_\nu(\theta) \quad \text{if } \nu \text{ is an integer.}$$  

(A.18)

In these expressions the $a_k$, $a_\nu$ and $\tilde{a}_\nu$ are non-null real numbers, for every $k$, $D^{(k)}$ is the $n \times n$ matrix with $(i,i')$-th element $\|x^{(i)} - x^{(i')}\|^{2k}$, $D^{(\nu)}$ is the $n \times n$ matrix with $(i,i')$-th element $\|x^{(i)} - x^{(i')}\|^{2\nu}$, $\tilde{D}^{(\nu)}$ is another non-null symmetric $n \times n$ matrix, and $R_\nu$ (resp. $\tilde{R}_\nu$) is a function from $(0, +\infty)$ to the space of $n \times n$ matrices such that $\|R(\nu)\| = o(\theta^{2\nu})$ (resp. $\|\tilde{R}(\nu)\| = O(\theta^{2\nu})$) when $\theta \to +\infty$.

With Matérn kernels, when $\theta \to +\infty$, $\|\Sigma_\theta^{-1}\| = O(\theta^{2\nu})$ (cf. Lemma C.11), so in the decomposition of $\Sigma_\theta$ given by Equation (A.17) (resp. Equation (A.18)), the intersection $\bigcap_{k=0}^{[\nu]} \ker \left( W^\top D^{(k)} W \right) \cap \ker \left( W^\top D^{(\nu)} W \right)$ (resp. the intersection $\bigcap_{k=0}^{\nu} \ker \left( W^\top D^{(k)} W \right) \cap \ker \left( W^\top \tilde{D}^{(\nu)} W \right)$) is necessarily the trivial vector space.

The rest of the proof is the same as in the case of Rational Quadratic and Squared Exponential kernels.

**Proof.** The first assertion of Proposition 4.1 implies the reference prior \( \pi(\theta) \) is integrable in the neighborhood of 0. Furthermore, when \( \theta \to 0 \), \( \Sigma_\theta \to I_n \) so the reference posterior \( \pi(\theta|y) \propto L(y|\theta)\pi(\theta) \) is integrable in the neighborhood of 0 as well.

All that remains to be proved is therefore that the reference posterior is integrable in the neighborhood of \(+\infty\). In the following \( \theta \to +\infty \), so we rely on the asymptotic expansion of \( \Sigma_\theta \) which is detailed in Appendix C.4.

Let \( \mathcal{H}_n \) be the hyperplane of \( \mathbb{R}^n \) defined by Lemma 4.2. Let us fix the observation vector \( y \in \mathbb{R}^n \setminus \mathcal{H}_n \).

The proof is somewhat trickier for Matérn kernels with integer smoothness, so we tackle this case at the end. Until further notice, assume the kernel is Rational Quadratic, Squared Exponential or Matérn with noninteger smoothness \( \nu > 1 \).

A.6.1. Rational Quadratic, Squared Exponential and Matérn kernels with noninteger smoothness \( \nu > 1 \). For Rational Quadratic and Squared Exponential (resp. Matérn with noninteger smoothness parameter \( \nu > 1 \)) kernels, Lemma C.14 (resp. Lemma C.16) shows how \( W^\top \Sigma_\theta W \) can be decomposed as

\[
W^\top \Sigma_\theta W = g(\theta) \left( W^\top Z W + g^*(\theta) W^\top Z^* W + R_g(\theta) \right),
\]

where:

- \( g \) is a positive differentiable function on \((0, +\infty)\);
- \( g^*(\theta) = \theta^{-2l} \) with \( l \in (0, +\infty) \) (actually, if the kernel is Rational Quadratic or Squared Exponential, \( l \in \mathbb{Z}_+ \));
- \( R_g \) is a differentiable function from \((0, +\infty)\) to \( \mathcal{M}_n \) such that \( \|R_g(\theta)\| = o(g^*(\theta)) \) and \( \frac{d}{d\theta} R_g(\theta) = o(g^*(\theta)) \);
- \( Z \) and \( Z^* \) are both fixed symmetric matrices;
- \( W^\top Z W \) is non-null.

Remark 4. Readers familiar with Berger, De Oliveira and Sansó (2001) may recognize similarities with the assumptions in Lemma 2 of the paper.
This is not coincidental. The matrices $Z$ and $Z^*$ essentially play the roles of the matrices $D$ and $D^*$ respectively. However, $Z$ (resp. $Z^*$) is not necessarily equal to $D$ (resp. $D^*$). In fact, in the Simple Kriging case where $W$ is the identity matrix, $Z = 11^\top$.

Let us differentiate $W^\top \Sigma_\theta W$:

(A.20) \[
\frac{d}{d\theta} W^\top \Sigma_\theta W = g'(\theta) W^\top \Sigma_\theta W + g(\theta) \left( g''(\theta) W^\top Z^* W + \frac{d}{d\theta} R_g(\theta) \right).
\]

This decomposition of the matrix $\frac{d}{d\theta} W^\top \Sigma_\theta W$ implies (cf. Lemma C.2) that it can be replaced in Equation (2.3) by $g(\theta) \left( g''(\theta) W^\top Z^* W + \frac{d}{d\theta} R_g(\theta) \right)$:

(A.21) \[
\text{Tr} \left[ \left\{ g'(\theta) W^\top \Sigma_\theta W + g(\theta) \left( g''(\theta) W^\top Z^* W + \frac{d}{d\theta} R_g(\theta) \right) \right\} \left( W^\top \Sigma_\theta W \right)^{-1} \right]^2 - \frac{1}{n - p} \left[ \text{Tr} \left\{ g''(\theta) W^\top Z^* W + \frac{d}{d\theta} R_g(\theta) \right\} \left( W^\top \Sigma_\theta W \right)^{-1} \right]^2
\]

So $\pi(\theta) \propto w(\theta)$, where

(A.22) \[
w(\theta)^2 := \text{Tr} \left[ \left\{ g(\theta) \left( g''(\theta) W^\top Z^* W + \frac{d}{d\theta} R_g(\theta) \right) \right\} \left( W^\top \Sigma_\theta W \right)^{-1} \right]^2 - \frac{1}{n - p} \left[ \text{Tr} \left\{ g(\theta) \left( g''(\theta) W^\top Z^* W + \frac{d}{d\theta} R_g(\theta) \right) \right\} \left( W^\top \Sigma_\theta W \right)^{-1} \right]^2.
\]

We have $w(\theta) \leq \tilde{w}(\theta)$, where

(A.23) \[
\tilde{w}(\theta) := \sqrt{\text{Tr} \left[ \left\{ g(\theta) \left( g''(\theta) W^\top Z^* W + \frac{d}{d\theta} R_g(\theta) \right) \right\} \left( W^\top \Sigma_\theta W \right)^{-1} \right]^2}.
\]
If \( W^\top ZW \) is nonsingular, then \( \tilde{w}(\theta) = O(g^*(\theta)) \). This implies \( \pi(\theta) = O(g^*(\theta)) = O(\theta^{-2l-1}) \), so the reference prior is proper. The likelihood function is bounded due to Proposition 4.3, so the reference posterior is proper.

**Remark 5.** Recall the decomposition of \( \Sigma_\theta \) from either Proposition 3.1 or Appendix A.5. If the vector 1 is one of the columns of \( H \), then we have \( W^\top D^{(0)} W = W^\top 11^\top W = 0 \). Berger, De Oliveira and Sansó (2001) assume that the matrix \( D^{(1)} \) is necessarily nonsingular, which implies that \( W^\top D^{(1)} W \) is non-null (and thus equal to \( W^\top ZW \)) and even nonsingular, so the paragraph above is applicable. This is why they reach the conclusion that the reference prior is proper as soon as 1 is one of the columns of \( H \) (denoted by \( X \) in their article). Because the underlying assumption that \( D^{(1)} \) is nonsingular does not generally hold (cf. Proposition 3.4), there is reason to doubt the conclusion. Indeed, Figures 1b, 1e and 1f do not seem to support the claim.

If \( W^\top ZW \) is singular, Proposition 4.1 still ensures that \( \pi(\theta) = O(\theta^{-1}) \). Moreover, for any non-null vector \( \xi \in \mathbb{R}^{n-p} \) that belongs to the kernel of \( W^\top ZW \), \( \xi^\top W^\top \Sigma_\theta W \xi = O(g(\theta)g^*(\theta)) \). A fortiori, \( v_{n-p}(\theta) = O(g(\theta)g^*(\theta)) \). As the rank of \( W^\top ZW \) is at least one, \( v_1(\theta)^{-1} = O(g(\theta)^{-1}) \). Gathering this, \( v_{n-p}(\theta)/v_1(\theta) = O(g^*(\theta)) \). **Proposition 4.3 implies that** \( L_y(\theta) = O(g^*(\theta))^{1/2} = O(\theta^{-1/2}) \). The reference posterior is proportional to \( L_y(\theta)\pi(\theta) = O(\theta^{-l-1}) \) and is proper.

**A.6.2. Matérn kernels with integer smoothness \( \nu \).** We now address the case where the correlation kernel is Matérn with integer smoothness \( \nu \). The proof strategy remains the same as for the other kernels, but the execution is a little trickier.

\( D^{(\nu)} \) is the \( n \times n \) matrix with \( (i, i') \)-th element \( \|x^{(i)} - x^{(i')}\|^{2\nu} \). Let \( \tilde{D}^{(\nu)} \) denote the \( n \times n \) matrix with null diagonal and \( (i, i') \)-th element \( (i \neq i') \) given by
\[
\|x^{(i)} - x^{(i')}\|^{2\nu} \left\{ -\log \left( \|x^{(i)} - x^{(i')}\| \right) - \frac{\log(\nu)}{2} - \gamma + \sum_{l=1}^{\nu} \frac{1}{2l} \right\},
\]
where \( \gamma \) is Euler’s constant.

Both \( D^{(\nu)} \) and \( \tilde{D}^{(\nu)} \) can appear in the decomposition of \( W^\top \Sigma_\theta W \) provided by Lemma C.18:

(A.24) \[
W^\top \Sigma_\theta W = g(\theta) \left( W^\top ZW + g^*(\theta)W^\top Z^*W + R_g(\theta) \right),
\]
where:

- $g$ is a positive differentiable function on $(1, +\infty)$;
- $Z$ and $Z^*$ are both fixed symmetric matrices;
- $W^T Z W$ is non-null;
- $g^*(\theta) = \log(\theta)^{-1}$ if there exist non-null real numbers $\lambda, \lambda^*$ such that $Z = \lambda D^{(\nu)}$ and $Z^* = \lambda^* \tilde{D}^{(\nu)}$;
- $g^*(\theta) = \theta^{-2l}$ or $g^*(\theta) = \log(\theta)\theta^{-2l}$ with $l \in (0, +\infty)$ otherwise;
- $R_\theta$ is a differentiable function from $(0, +\infty)$ to $\mathcal{M}_n$ such that $\|R_\theta(\theta)\| = o(g^*(\theta))$ and $\|\frac{d}{d\theta} R_\theta(\theta)\| = o(g^*(\theta))$ when $\theta \to +\infty$.

First, assume either that for all $\lambda \neq 0$, $Z \neq \lambda D^{(\nu)}$ or that for all $\lambda^* \neq 0$, $Z^* \neq \lambda^* \tilde{D}^{(\nu)}$.

In Equation (A.19), according to Lemma C.18, $g^*(\theta)$ may be $\theta^{-2l}\log(\theta)$ instead of $\theta^{-2l}$. If $g^*(\theta) = \theta^{-2l}$ for some $l \in (0, +\infty)$, then the proof is the same as for Rational Quadratic, Squared Exponential and Matérn kernels with noninteger $\nu$. Assume therefore that $g^*(\theta) = \theta^{-2l}\log(\theta)$ for some $l \in (0, +\infty)$. Then its derivative is $g^*(\theta) = \theta^{-2l-1}(1 - 2l\log(\theta))$.

If $W^T Z W$ is nonsingular, then the reference prior distribution is proper since $\pi(\theta) = O(g^*(\theta)) = O(\theta^{-2l-1}\log(\theta))$. Proposition 4.3 guarantees that the likelihood function is bounded and therefore that the reference posterior is proper.

If $W^T Z W$ is singular, Proposition 4.1 still ensures that the reference prior is $O(\theta^{-1})$. Given the rank of $W^T Z W$ is at least one, $v_{n-1}(\theta)/v_1(\theta) = O(g^*(\theta))$. Proposition 4.3 implies that $L_y(\theta) = O(g^*(\theta)^{1/2}) = O(\theta^{-l}\log(\theta)^{1/2})$, so the reference posterior is proportional to $L_y(\theta)\pi(\theta) = O(\theta^{-l-1}\log(\theta)^{1/2})$ and thus proper.

Now, assume there exist non-null real numbers $\lambda, \lambda^*$ such that $Z = \lambda D^{(\nu)}$ and $Z^* = \lambda^* \tilde{D}^{(\nu)}$.

In that case, according to Lemma C.18, $g^*(\theta) = \log(\theta)^{-1}$. Its derivative is $g^*(\theta) = -\theta^{-1}\log(\theta)^{-2}$.

If $W^T Z W$ is nonsingular, the reference prior is proper since $\pi(\theta) = O(g^*(\theta)) = O(\theta^{-1}\log(\theta)^{-2})$. Proposition 4.3 implies that the likelihood function is bounded and therefore that the reference posterior is proper.

If $W^T Z W$ is singular, it nevertheless turns out that for large enough $\theta$, $W^T Z W + g^*(\theta)W^T Z^* W$ is nonsingular. This is due to Lemma C.11, which asserts that $\|\Sigma^{-1}\| = O(\theta^{2\nu})$. The reference prior is then $O(g^*(\theta)g^*(\theta)^{-1}) = O(\theta^{-1}\log(\theta)^{-1})$. Besides, as the rank of $W^T Z W$ is at least one, $1/v_1(\theta) = O(g(\theta)^{-1})$ and therefore $v_{n-p}(\theta)/v_1(\theta) = O(g^*(\theta))$. Proposition 4.3 implies
that $L_y(\theta) = O(g^*(\theta)^{1/2}) = O(\log(\theta)^{-1/2})$. The reference posterior is then proportional to $L_y(\theta)\pi(\theta) = O(\theta^{-1} \log(\theta)^{-3/2})$ and is proper. 

\[\Box\]

**A.7. A more precise formulation of Lemma 4.2, Proposition 4.3 and Theorem 4.4.** The proof of Lemma 4.2 actually proves a slightly stronger result, which we provide in this section. This stronger result in turn leads to slightly stronger versions of Proposition 4.3 and Theorem 4.4.

In order to be able to state this result, we must use the notations of the proof of Lemma 4.2, together with additional definitions:

**Definition A.1.** For any vector $y \in \mathbb{R}^n$ and for any nonnegative real number $t$, let $A_k(y)$ be the following statement:

$W^T y$ does not belong to the vector subspace of $\mathbb{R}^{n-p}$ spanned by the columns of the matrices $W^T D^{(k)} W$ with nonnegative integer $k$ strictly smaller than $t$ (resp. $W^T y \neq 0$ if $t = 0$). However, the vector subspace of $\mathbb{R}^{n-p}$ spanned by these columns and the columns of $W^T D^{(t)} W$ (resp. by the columns of $W^T D^{(0)} W$ if $t = 0$) is $\mathbb{R}^{n-p}$ itself.

**Definition A.2.** For any vector $y \in \mathbb{R}^n$ and for any positive integer $\nu$, let $\tilde{A}_\nu(y)$ be the following statement:

$W^T y$ does not belong to the vector subspace of $\mathbb{R}^{n-p}$ spanned by the columns of the matrices $W^T D^{(k)} W$ with nonnegative integer $k$ smaller or equal to $\nu$. However, the vector subspace of $\mathbb{R}^{n-p}$ spanned by these columns and the columns of $W^T \tilde{D}^{(\nu)} W$ is $\mathbb{R}^{n-p}$ itself.

**Remark 6.** In the case of Simple Kriging, in both definitions, $p = 0$ and the matrix $W$ is the identity $n \times n$ matrix.

The more precise version of Lemma 4.2 is:

**Lemma A.3.** Depending on the correlation kernel, the following condition on $y \in \mathbb{R}^n$ is sufficient for Equation (4.3) when $\theta \to +\infty$:

- Rational Quadratic and Squared Exponential kernels: there exists a nonnegative integer $k'$ such that Assumption $A_{k'}(y)$ holds;
- Matérn kernels with noninteger smoothness $\nu > 1$: either there exists a nonnegative integer $k' < \nu$ such that Assumption $A_{k'}(y)$ holds or Assumption $\tilde{A}_{\nu}(y)$ holds;
- Matérn kernels with integer smoothness $\nu \geq 1$: either there exists a nonnegative integer $k' \leq \nu$ such that Assumption $A_{k'}(y)$ holds or Assumption $\tilde{A}_{\nu}(y)$ holds.
Regardless of whether the kernel is Rational Quadratic, Squared Exponential, or Matérn (with integer or noninteger smoothness $\nu \geq 1$), the set of all $y \in \mathbb{R}^n$ that do not satisfy this sufficient condition is a vector subspace of $\mathbb{R}^n$ of dimension smaller or equal to $n - 1$.

**Proof.** The proof of Lemma 4.2 also proves this result. \hfill \Box

The condition on $y \in \mathbb{R}^n$ stated in Lemma A.3 replaces the condition from Lemma 4.2 about $y$ not belonging to $H_n$.

Proposition 4.3 can thus be replaced by the following proposition.

**Proposition A.4.** The condition on $y \in \mathbb{R}^n$ stated in Lemma A.3 is sufficient for Equation (4.5) to hold when $\theta \to +\infty$.

This more precise version of Proposition 4.3 leads to a more precise version of Theorem 4.4:

**Theorem A.5.** The condition on $y \in \mathbb{R}^n$ stated in Lemma A.3 is sufficient for the reference posterior distribution $\pi(\theta | y)$ to be proper. The set of all $y \in \mathbb{R}^n$ that do not satisfy this sufficient condition is a vector subspace of $\mathbb{R}^n$ of dimension smaller or equal to $n - 1$.

**Proof.** The proof of Theorem 4.4 can be used to prove this result, provided Proposition A.4 is used instead of Proposition 4.3. \hfill \Box

**APPENDIX B: SOME TAIL RATES OF LIKELIHOOD AND PRIOR**

The purpose of this appendix is twofold. First, to list examples that show how the tail rate of the reference prior density varies to accommodate the various tail rates of the likelihood function while making sure the reference posterior is always proper. Second, to show that addressing the various cases considered in the proof of Theorem 4.4, Appendix A.6, is not merely necessary for the sake of mathematical rigor, but because these cases do occur in practice. The proofs of the results presented in this Appendix rely on auxiliary facts from Appendix C. Appendix C can be found in supplementary material.

Figure 2 gives a sample of the wide variety of tail rates for the likelihood function depending on the Kriging model and the smoothness of the correlation kernel. Among the cases considered in Figure 2, the most remarkable is 2c. Affine Kriging with a Matérn kernel with smoothness $\nu = 3/2$ leads to a likelihood function that does not vanish when $\theta \to +\infty$. 

Fig 2: Logarithm of the likelihood function for large values of $\theta$ for varying Kriging models and correlation kernels. In each case, the design space is one-dimensional ($r = 1$) and the design set is the 10-point regular grid on $[0, 1]$. At every design point $x$, the observed value is $\sin(\pi x)$.

This behavior of the likelihood is a good reason to investigate the behavior of the tail rate of the reference prior density more closely than we did in Proposition 4.1.

All propositions in this appendix are valid under the assumption that some $n$-point design set has been fixed and that $n > p$.

PROPOSITION B.1. In the case of Affine Kriging, with a Matérn kernel with smoothness $\nu \in (1, 2)$, the reference prior $\pi(\theta)$ is $O(\theta^{-2(2-\nu)-1})$ when $\theta \to +\infty$. It is a proper prior distribution. Furthermore, for every $y \in \mathbb{R}^n$ such that $W^\top y$ is non-null, the likelihood function converges to a non-null constant when $\theta \to +\infty$.

REMARK 7. While this Proposition only provides an upper bound for the tail rate of the reference prior, the bound $O(\theta^{-2(2-\nu)-1})$ seems to be tight, judging by Figure 1c which was obtained with $\nu = 3/2$. 
PROPRIETY OF THE REFERENCE POSTERIOR DISTRIBUTION

PROOF. According to Lemma C.15, \( \Sigma_\theta \) can be written as:

\[
\Sigma_\theta = D^{(0)} + \frac{a_1}{\theta^2} D^{(1)} + \frac{a_\nu}{\theta^{2\nu}} D^{(\nu)} + \frac{a_2}{\theta^4} D^{(2)} + R(\theta).
\]

In the expression above,

- \( a_1 = -\Gamma(\nu - 1)\nu/\Gamma(\nu) \),
- \( a_2 = \Gamma(\nu - 2)\nu^2/\Gamma(\nu) \),
- \( a_\nu = \Gamma(-\nu)\nu^\nu/\Gamma(\nu) \),
- \( D^{(0)} \) is the \( n \times n \) matrix filled with ones,
- \( D^{(1)} \) is the \( n \times n \) matrix with \((i, i')\)-th element \( \| x^{(i)} - x^{(i')} \|_2^2 \),
- \( D^{(2)} \) is the \( n \times n \) matrix with \((i, i')\)-th element \( \| x^{(i)} - x^{(i')} \|_4^4 \),
- \( D^{(\nu)} \) is the \( n \times n \) matrix with \((i, i')\)-th element \( \| x^{(i)} - x^{(i')} \|_{2\nu}^{2\nu} \),
- \( R \) is a differentiable function from \((0, +\infty)\) to the space of real \( n \times n \) matrices that satisfies \( \| R(\theta) \| = o(\theta^{-4}) \) and \( \| \frac{d}{d\theta} R(\theta) \| = o(\theta^{-5}) \) when \( \theta \to +\infty \).

Under Affine Kriging, \( H \) is the \( n \times (r + 1) \) matrix whose first column is \( 1 \) (the vector of \( \mathbb{R}^n \) filled with ones) and whose last \( r \) columns contain the coordinates of the design set.

Because \( W \) is orthogonal to \( H \), Schoenberg (1937) (for example) implies that \( W^T D^{(0)} W \) and \( W^T D^{(1)} W \) are both null. We have therefore

\[
W^T \Sigma_\theta W = \frac{a_\nu}{\theta^{2\nu}} W^T D^{(\nu)} W + \frac{a_2}{\theta^4} W^T D^{(2)} W + W^T R(\theta) W.
\]

Using the notations from Appendix A.6, we can identify \( Z := a_\nu D^{(\nu)} \) and \( Z^* := a_2 D^{(2)} \). Thus \( g(\theta) := \theta^{-2\nu} \) and \( g(\theta)^* := \theta^{-2l} \) with \( l = 2 - \nu \).

Besides, Lemma C.11 asserts that \( \| \Sigma_\theta^{-1} \| = O(\theta^{2\nu}) \) when \( \theta \to +\infty \). Applying Lemma C.1,

\[
\Sigma_\theta^{-1} = W \left( W^T \Sigma_\theta W \right)^{-1} W^T + \Sigma_\theta^{-1} H \left( H^T \Sigma_\theta^{-1} H \right)^{-1} H^T \Sigma_\theta^{-1}.
\]

As both matrices in the right term are positive semi-definite, we have \( \left\| W \left( W^T \Sigma_\theta W \right)^{-1} W^T \right\| = O(\theta^{2\nu}) \). And since \( W^T W = I_{n-p} \), this implies that \( \left\| \left( W^T \Sigma_\theta W \right)^{-1} \right\| = O(\theta^{2\nu}) \).

\[
(B.1) \quad \Sigma_\theta = D^{(0)} + \frac{a_1}{\theta^2} D^{(1)} + \frac{a_\nu}{\theta^{2\nu}} D^{(\nu)} + \frac{a_2}{\theta^4} D^{(2)} + R(\theta).
\]
Therefore \( \mathbf{W}^{\top} \mathbf{D}^{(\nu)} \mathbf{W} \) is nonsingular. The proof of Theorem 4.4 given in Appendix A.6 yields the first assertion of this Proposition: the paragraph of Appendix A.6 below Equation (A.23) establishes in this situation that 
\( \pi(\theta) = O(\theta^{-2l-1}) = O(\theta^{-2(2-\nu)-1}) \) and that the reference prior \( \pi(\theta) \) is proper.

Moreover, given that \( \mathbf{W}^{\top} \mathbf{D}^{(\nu)} \mathbf{W} \) is nonsingular, Equation (B.2) implies that, for every \( y \in \mathbb{R}^n \) such that \( \mathbf{W}^{\top} y \) is non-null,

\[
\lim_{\theta \to +\infty} \left| \mathbf{W}^{\top} \Sigma_{\theta} \mathbf{W} \right|^{-\frac{1}{2}} \left( \mathbf{y}^{\top} \mathbf{W} \left( \mathbf{W}^{\top} \Sigma_{\theta} \mathbf{W} \right)^{-1} \mathbf{W}^{\top} \mathbf{y} \right)^{-\frac{n-p}{2}} = \left| \mathbf{W}^{\top} \mathbf{D}^{(\nu)} \mathbf{W} \right|^{-\frac{1}{2}} \left( \mathbf{y}^{\top} \mathbf{W} \left( \mathbf{W}^{\top} \mathbf{D}^{(\nu)} \mathbf{W} \right)^{-1} \mathbf{W}^{\top} \mathbf{y} \right)^{-\frac{n-p}{2}}.
\]

(B.4) This implies that the likelihood function converges to a non-null constant. \( \square \)

As mentioned in Appendix A.6, the case of Matérn kernels with integer smoothness is particularly tricky. Figure 3 focuses on the case where \( \nu = 1 \). Once again, the most striking subfigure is the one showing the likelihood function under Affine Kriging: 3f. The behavior of the likelihood function and reference prior in this case is given in the following Proposition.

**Proposition B.2.** In the case of Affine Kriging, with a Matérn kernel with smoothness \( \nu = 1 \), the reference prior \( \pi(\theta) \) is \( O(\theta^{-3} \log(\theta)) \) when \( \theta \to +\infty \). It is a proper prior distribution. Furthermore, for every \( y \in \mathbb{R}^n \) such that \( \mathbf{W}^{\top} y \) is non-null, the likelihood function converges to a non-null constant when \( \theta \to +\infty \).

**Remark 8.** While this Proposition is only able to provide an upper bound for the tail rate of the reference prior, the bound \( O(\theta^{-3} \log(\theta)) \) seems to be tight, judging by Figure 3c.

**Proof.** We use Lemma C.17 to obtain:

\[
\Sigma_{\theta} = D^{(0)} - 2 \left( \frac{\log(\theta)}{\theta^2} D^{(1)} + \frac{1}{\theta^2} \tilde{D}^{(1)} \right) + \frac{\log(\theta)}{\theta^4} D^{(2)} + R(\theta).
\]

(B.5) In the expression above,

- \( D^{(0)} \) is the \( n \times n \) matrix filled with ones,
Fig 3: Logarithm of the prior (top) and likelihood function (bottom) for large values of $\theta$ for varying Kriging models using the Matérn kernel with smoothness $\nu = 1$. In each case, the design space is one-dimensional ($r = 1$) and the design set is the 10-point regular grid on $[0, 1]$. At every design point $x$, the observed value is $\sin(\pi x)$.

- $D^{(1)}$ is the $n \times n$ matrix with $(i, i')$-th element $\|x^{(i)} - x^{(i')}\|^2$,
- $D^{(2)}$ is the $n \times n$ matrix with $(i, i')$-th element $\|x^{(i)} - x^{(i')}\|^4$,
- $\tilde{D}^{(1)}$ is the $n \times n$ matrix with null diagonal and $(i, i')$-th element $(i \neq i')$ given by
  $$\|x^{(i)} - x^{(i')}\|^2 \left\{ -\log \left( \|x^{(i)} - x^{(i')}\| \right) - \gamma + \frac{1}{2} \right\},$$
  where $\gamma$ is Euler’s constant,
- $R$ is a differentiable function from $(0, +\infty)$ to the space of real $n \times n$ matrices that satisfies $\|R(\theta)\| = O(\theta^{-4})$ and $\|\frac{d}{d\theta} R(\theta)\| = O(\theta^{-5})$ when $\theta \to +\infty$.

The rest is similar to the proof of Proposition B.1. Using the notations of Appendix A.6, we have $Z := -2\tilde{D}^{(1)}, Z^* := D^{(2)}, g(\theta) := \theta^{-2}, g^*(\theta) := \ldots$
\[ \theta^{-2l} \log(\theta) \text{ with } l = 1. \] 

The relevant part of Appendix A.6 is the part concerning Matérn kernels with integer smoothness: Appendix A.6.2. As \( \nu = 1 \), we are in the case where either for all \( \lambda \neq 0 \), \( Z \neq \lambda D^{(\nu)} \) or for all \( \lambda^* \neq 0 \), \( Z^* \neq \lambda^* \tilde{D}^{(\nu)} \). Here, both checks hold. Further, for the same reason as in the proof of Proposition B.1, we are in the subcase where \( W^\top Z W \) is nonsingular. In this subcase, Appendix A.6.2 states that the reference prior is \( O(\theta^{-2l-1} \log(\theta)) = O(\theta^{-3} \log(\theta)) \) and that it is proper. Finally, \( W^\top Z W \) being nonsingular also leads to the conclusion that, for every \( y \in \mathbb{R}^n \) such that \( W^\top y \) is non-null, the likelihood function converges to a non-null constant when \( \theta \) goes to infinity.

The most striking behavior of the reference prior density is given in Figure 3b (Ordinary Kriging), where the tail rate seems to be \( \theta^{-1} \log(\theta)^{-1} \) multiplied by some constant factor. It decreases a little faster than the upper bound given in Proposition 4.1, \( O(\theta^{-1}) \), but still not fast enough to make the reference prior proper. Proposition B.3 below shows that \( O(\theta^{-1} \log(\theta)^{-1}) \) is indeed an upper bound for the reference prior.

**Proposition B.3.** In the case of Ordinary Kriging, with a Matérn kernel with smoothness \( \nu = 1 \), if \( n > r + 3 \), then:

- the reference prior \( \pi(\theta) \) is \( O(\theta^{-1} \log(\theta)^{-1}) \) when \( \theta \to +\infty \);
- the matrix \( W^\top D^{(1)} W \) is singular;
- for every vector \( y \in \mathbb{R}^n \) such that \( W^\top y \) does not belong to the vector subspace of \( \mathbb{R}^{n-1} \) spanned by the columns of \( W^\top D^{(1)} W \), the reference posterior distribution \( \pi(\theta | y) \) is \( O(\theta^{-1} \log(\theta)^{-3/2}) \) when \( \theta \to +\infty \).

**Remark 9.** The upper bound given for the reference posterior tail rate just barely makes it proper. Yet the tail rate suggested by Figure 3e for the likelihood function is consistent with it: \( \log(\theta)^{-1/2} \) multiplied by some constant factor.

**Proof.** The second assertion follows from Corollary 3.3. Only the first and third assertions need to be proved.

Let us use Lemma C.17 to obtain Equation (B.5).

With Ordinary Kriging, the matrix \( H \) is the vector \( 1 \in \mathbb{R}^n \) whose entries are all equal to 1. Therefore \( W^\top D^{(0)} W \) is the null \((n-1) \times (n-1)\) matrix and
\[ W^\top \Sigma W = -2 \left( \frac{\log(\theta)}{\theta^2} W^\top D^{(1)} W + \frac{1}{\theta^2} W^\top \tilde{D}^{(1)} W \right) + \frac{\log(\theta)}{\theta^4} W^\top D^{(2)} W + W^\top R(\theta) W. \]

Using the notations of Appendix A.6, we have \( Z := -2 D^{(1)} \) and \( Z^* := -2 \tilde{D}^{(1)} \), \( g(\theta) = \theta^{-2} \log(\theta) \), \( g^*(\theta) := \log(\theta)^{-1} \).

Corollary 3.3 states that the rank of \( Z \) is lower or equal to \( r + 2 \), so this is a fortiori true for the rank of \( W^\top Z W \). Since \( r + 2 < n - 1 \), this implies that \( W^\top Z W \) is singular.

The proof of Theorem 4.4, Appendix A.6, then yields both results of this Proposition. The relevant part of Appendix A.6 is the part concerning Matérn kernels with integer smoothness: Appendix A.6.2. As \( \nu = 1 \), we are in the case where there exists \( \lambda \neq 0 \) such that \( Z = \lambda D^{(\nu)} \) (\( \lambda = -2 \)) and there exists \( \lambda^* \neq 0 \) such that \( Z^* = \lambda^* \tilde{D}^{(\nu)} \) (\( \lambda^* = -2 \)). Further, we are in the subcase where \( W^\top Z W \) is singular. This subcase is dealt with in the last paragraph of Appendix A.6.2: it yields the first assertion.

Any vector \( y \in \mathbb{R}^n \) such that \( W^\top y \) does not belong to the subspace of \( \mathbb{R}^{n-1} \) spanned by the columns of \( W^\top D^{(1)} W \) verifies Assumption \( \hat{A}_1(y) \) from Appendix A.7. Therefore Proposition A.4 can be applied instead of Proposition 4.3 and the last paragraph of Appendix A.6.2 yields the third assertion. \( \square \)

REFERENCES


