Statistica Sinica **8**(1998), 511-526

APPROXIMATING THE MEAN SQUARED PREDICTION ERROR IN LINEAR MODELS UNDER THE FAMILY OF EXPONENTIAL CORRELATIONS

Markus Abt

University of Waterloo

Abstract: We investigate approximations for the mean squared prediction error in a linear regression model with correlated errors. The correlation structure is assumed to be the family of exponential correlations widely used in practical applications of computer experiments. Well known members of this family include the Ornstein-Uhlenbeck process as well as stochastic processes with analytic sample paths. Special emphasis is put on the situation when the true values of the parameters involved in the correlation structure are on the boundary of the parameter space.

Key words and phrases: Fisher information matrix, Gaussian stochastic processes, homogeneous random fields, mean squared prediction error.

1. Introduction

In the design and analysis of computer experiments, the deterministic output $y(t)$ of a computer code for inputs $t \in T$ is commonly considered as a realization of a Gaussian stochastic process Y. A widely accepted model is

$$Y(t) = \beta + Z(t), \ \beta \in \mathbb{R}. \tag{1}$$

We take the input t to be d-dimensional in $T = [0,1]^d$. In (1), the Gaussian process Z is assumed to have zero mean, constant variance σ^2, and a correlation structure given by $\mathrm{Corr}(Z(s), Z(t)) = C_\delta(|s-t|)$ for $s, t \in T$, which depends on some unknown parameter (vector) δ. Therein, by the absolute value of a vector, we mean the vector obtained by taking absolute values of its components. In one dimension, the processes considered will thus be stationary, and in two and more dimensions we are led to homogeneous random fields (see Yaglom (1987), Section 21).

More general models can be considered as well. The constant mean may be replaced by a trend term of the form $\sum_{j=1}^k \beta_j f_j(t)$ with unknown parameters $\beta_1, \ldots, \beta_k$ and given functions $f_1, \ldots, f_k : [0,1]^d \to \mathbb{R}$. However, it has been observed that the model (1) performs sufficiently well in practical applications

(see for example Welch et al. (1992)). We could also incorporate nonstationary covariance structures as those of the Brownian motion or the Brownian sheet. As very often there is no prior knowledge that would justify any of these assumptions, we will not pursue this any further here.

Long running times of the computer code are not unusual. Therefore, only a limited number of runs for inputs $D = \{t^{(1)}, \ldots, t^{(n)}\} \subset T$ can be made, leading to observations $Y_D = (Y(t^{(1)}), \ldots, Y(t^{(n)}))^T$. We use linear predictors of the form $c_\delta(t)^T Y_D$ for predicting $Y(t)$ at untried inputs $t \in T \setminus D$. Following Sacks et al. (1989), the vector $c_\delta(t)$ can be found by minimizing the mean squared prediction error

$$\mathrm{MSE}(c_\delta(t)^T Y_D) = \mathrm{E}\Big[\Big(c_\delta(t)^T Y_D - Y(t)\Big)^2\Big] \tag{2}$$

with respect to $c_\delta(t)$ under the unbiasedness constraint $c_\delta(t)^T \mathbf{1} = 1$, wherein $\mathbf{1} = (1, \ldots, 1)^T$. Let R_δ be the $n \times n$ correlation matrix of Y_D and let $r_\delta(t)$ be the $n \times 1$ vector of correlations between Y_D and $Y(t)$. Then minimization of (2) leads to

$$c_\delta(t) = R_\delta^{-1} r_\delta(t) + R_\delta^{-1} \mathbf{1} \frac{1 - \mathbf{1}^T R_\delta^{-1} r_\delta(t)}{\mathbf{1}^T R_\delta^{-1} \mathbf{1}},$$

from which we find the best linear predictor $\hat{Y}_\delta(t)$ of $Y(t)$ to be

$$\hat{Y}_\delta(t) = c_\delta(t)^T Y_D = \frac{\mathbf{1}^T R_\delta^{-1} Y_D}{\mathbf{1}^T R_\delta^{-1} \mathbf{1}} + r_\delta(t)^T R_\delta^{-1} \Big(Y_D - \frac{\mathbf{1}^T R_\delta^{-1} Y_D}{\mathbf{1}^T R_\delta^{-1} \mathbf{1}} \mathbf{1}\Big). \tag{3}$$

Again, the subscript δ denotes dependence on the parameter δ of the correlation function.

Assuming δ known, the mean squared prediction error of $\hat{Y}_\delta(t)$ is

$$\mathrm{MSE}(\hat{Y}_\delta(t)) = \sigma^2 \Big(1 - r_\delta(t)^T R_\delta^{-1} r_\delta(t)\Big) + \frac{\sigma^2}{\mathbf{1}^T R_\delta^{-1} \mathbf{1}} \Big(1 - r_\delta(t)^T R_\delta^{-1} \mathbf{1}\Big)^2. \tag{4}$$

The first part on the right hand side of (4) is the conditional variance of $Y(t)$ given the observations Y_D. This corresponds to the mean squared error in case of known mean β. The second part considers the additional variability due to estimating β, as, for the maximum likelihood estimate $\hat{\beta}$, we have $\mathrm{Var}(\hat{\beta}) = \sigma^2 / \mathbf{1}^T R_\delta^{-1} \mathbf{1}$.

As a measure of uncertainty of the predictor $\hat{Y}_\delta(t)$, the mean squared error is used in two ways: First, to choose locations $t^{(1)}, \ldots, t^{(n)}$ which provide highest possible accuracy in the predictions, we can look at *optimal design* problems of the form

$$\min_{D \subset T : |D| = n} \max_{t \in T} \mathrm{MSE}\Big(\hat{Y}_\delta(t)\Big) \quad \text{or} \quad \min_{D \subset T : |D| = n} \int_T \mathrm{MSE}\Big(\hat{Y}_\delta(t)\Big)\, dt. \tag{5}$$

The problem of finding optimal robust designs with respect to the integrated mean squared error is considered in Sacks et al. (1989). Second, $(1-\alpha)$-*prediction intervals* for $Y(t)$ are usually found via

$$\hat{Y}_\delta(t) \pm q_{1-\alpha/2}\sqrt{\widehat{\mathrm{MSE}\big(\hat{Y}_\delta(t)\big)}}, \tag{6}$$

wherein $q_{1-\alpha}$ denotes the upper 100α-percentage point of the standard normal distribution. The estimate of $\mathrm{MSE}(\hat{Y}_\delta(t))$ in (6) is generally obtained by replacing δ in (4) by a suitable estimate $\hat{\delta}$.

To compute predictions, the unknown parameter δ in (3) is commonly replaced by the maximum likelihood estimate $\hat{\delta}$, leading to a predictor $\hat{Y}_{\hat{\delta}}(t)$. The usage of (5) and (6) may suffer from this, as, due to the additional variability in $\hat{\delta}$, we expect $\mathrm{MSE}(\hat{Y}_{\hat{\delta}}(t))$ to be larger than $\mathrm{MSE}(\hat{Y}_\delta(t))$.

In the present paper we do not investigate the problem of deriving accurate prediction intervals. Instead, we concentrate on an approximation for $\mathrm{MSE}(\hat{Y}_{\hat{\delta}}(t))$ and evaluate it by simulation. The emphasis is on the family of exponential correlation functions. In one dimension, for $s,t \in [0,1]$, it is given by

$$C_\delta(|s-t|) = \exp(-\theta|s-t|^{2-\alpha}),\ (\theta,\alpha) \in (0,\infty)\times[0,2).$$

Special cases included are the correlation of the Ornstein-Uhlenbeck process for $\alpha = 1$ and the Gaussian correlation for $\alpha = 0$. The first can be constructed by an exponential transformation of a Brownian motion (see Ross (1983), page 218), and the sample paths are therefore nowhere differentiable (as are those of stochastic processes obtained for any other value $\alpha \in (0,2)$). On the other hand, the sample paths in the case $\alpha = 0$ are infinitely differentiable (see Loève (1978), page 186). As prior information one might know at most whether the sample paths are differentiable or not. We therefore restrict attention to the cases $\alpha = 0$ and $\alpha = 1$. According to whether α is assumed to be known or unknown, the vector δ of unknown parameters is either $\delta = \theta$ or $\delta = (\theta,\alpha)$.

In $d \geq 2$ dimensions, for $s,t \in [0,1]^d$, we use the 'product correlation rule'

$$C_\delta(|s-t|) = \prod_{k=1}^{d} \exp\left(-\theta_k|s_k-t_k|^{2-\alpha_k}\right),\ (\theta_1,\ldots,\theta_d,\alpha_1,\ldots,\alpha_d) \in (0,\infty)^d\times[0,2)^d,$$

to define a correlation on $[0,1]^d$. Again, we may have $\delta = (\theta_1,\ldots,\theta_d)$ or $\delta = (\theta_1,\ldots,\theta_d,\alpha_1,\ldots,\alpha_d)$, that is, we only look at the cases where all α_k's are either known or unknown, whereas the θ_k's are always assumed to be unknown.

In Section 2 we review an approximation for the mean squared prediction error of $\hat{Y}_{\hat{\delta}}(t)$. It in turn requires an approximation of the mean squared error

matrix of $\hat{\delta}$, for which we suggest using the inverse of the Fisher information matrix. This was seen to be appropriate before in Abt and Welch (1998). Some computational aspects in deriving it are discussed in Section 3. Section 4 finally evaluates the approximations by simulation. We focus only on the two cases where all the α_k's are either zero or one, as these correspond to two important families of stochastic processes. The first case is of particular interest, as zero is the boundary value for the parameters $\alpha_1, \ldots, \alpha_d$. Appropriate modifications of the information matrix are considered.

2. Approximating the Mean Squared Prediction Error

Approximations of $\mathrm{MSE}(\hat{Y}_{\hat{\delta}}(t))$ were derived earlier in the literature (see Harville and Jeske (1992), Kackar and Harville (1984), and Zimmerman and Cressie (1992)). We will briefly follow the outline in Zimmerman and Cressie (1992) and then concentrate on the peculiarities of the product correlation structure in this context. We note, as the multivariate normal distribution is symmetric about its mean, $\hat{Y}_{\hat{\delta}}(t)$ is unbiased for $Y(t)$ (see Zimmerman and Cressie (1992)). We can write

$$\mathrm{MSE}\Big(\hat{Y}_{\hat{\delta}}(t)\Big) = \mathrm{E}\Big[\Big((c_\delta(t)^T Y_D - Y(t)) + (c_{\hat{\delta}}(t)^T Y_D - c_\delta(t)^T Y_D)\Big)^2\Big]. \tag{7}$$

Again, due to multivariate normality, the two components herein are shown to be independent in Zimmerman and Cressie (1992), which leads to

$$\mathrm{MSE}\Big(\hat{Y}_{\hat{\delta}}(t)\Big) = \mathrm{MSE}\Big(\hat{Y}_{\delta}(t)\Big) + \mathrm{E}\Big[\Big(c_{\hat{\delta}}(t)^T Y_D - c_\delta(t)^T Y_D\Big)^2\Big]. \tag{8}$$

By a second order Taylor series expansion of $h(\hat{\delta}; t) = (c_{\hat{\delta}}(t)^T Y_D - c_\delta(t)^T Y_D)^2$ around the true parameter δ, we find the approximation

$$\mathrm{AMSE}\Big(\hat{Y}_{\hat{\delta}}(t)\Big) = \mathrm{MSE}\Big(\hat{Y}_{\delta}(t)\Big) + \sigma^2 \mathrm{trace}\Big(\Gamma_\delta G_\delta(t)\Big) \tag{9}$$

for $\mathrm{MSE}(\hat{Y}_{\hat{\delta}}(t))$. The second term considers the extra variability due to estimating the correlation parameters δ. Therein, Γ_δ denotes a matrix that equals or approximates the mean squared error matrix $\mathrm{E}[(\hat{\delta} - \delta)(\hat{\delta} - \delta)^T]$. The matrix $G_\delta(t)$ is given by $(G_\delta(t))_{k,\ell} = c_\delta^{(k)}(t)^T R_\delta c_\delta^{(\ell)}(t)$, wherein the vector $c_\delta^{(k)}(t)$ is the componentwise partial derivative of $c_\delta(t)$ with respect to δ_k. The matrix $G_\delta(t)$ and the vector $c_\delta^{(k)}(t)$ are in d dimensions if the α_k's are known and in $2d$ dimensions if the α_k's are unknown. With $B_\delta = R_\delta^{-1}\mathbf{1}/\mathbf{1}^T R_\delta^{-1}\mathbf{1}$ and $A_\delta = R_\delta^{-1} - B_\delta \mathbf{1}^T R_\delta^{-1}$, we have

$$c_\delta^{(k)}(t) = A_\delta\Big(r_\delta^{(k)}(t) - R_\delta^{(k)}(B_\delta + A_\delta r_\delta(t))\Big),$$

again denoting elementwise partial derivatives with respect to δ_k by a superscript (k). For $t^{(i)} \in D$, we have $c_\delta^{(k)}(t^{(i)}) = 0$ for all k and thus also $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t^{(i)})) = 0$.

We note that the derivation of (9) essentially assumes Y_D and $\hat{\delta}$ to be uncorrelated, or, as stated in Kackar and Harville (1984), that $\hat{\delta}$ was obtained from previous data. The first does not hold for the maximum likelihood estimate $\hat{\delta}$ and the latter is either rarely true or impossible in practical applications. A first order Taylor series expansion of $\hat{\delta}$ around the data y_D can be used to overcome this drawback. This leads to an additional term in (9). A detailed investigation of this approach, which might be particularly useful for estimating the prediction errors, has not been carried out yet but is the topic of future work.

We suggest using the inverse of the Fisher information matrix as an approximation for the mean squared error matrix of $\hat{\delta}$. Justifications for this approach can be given in an asymptotic sense. Consider a sequence of equispaced designs D_n of size n in $[0,1]$ and the corresponding sequence D_n^d of product designs of size n^d in $[0,1]^d$. Let $\theta = (\theta_1, \ldots, \theta_d)^T$. Then, for $d \geq 2$ and $\alpha_1 = \cdots = \alpha_d = 1$, Ying (1993), Theorems 2 and 4 shows the strong consistency of $(\hat{\sigma}^2, \hat{\theta}^T)^T$ and

$$n^{(d-1)/2} \begin{pmatrix} \hat{\sigma}^2 - \sigma^2 \\ \hat{\theta} - \theta \end{pmatrix} \rightarrow_{\mathcal{D}} \mathcal{N}(0, \Sigma) \tag{10}$$

for $n \to \infty$. The covariance matrix Σ is specified in more detail in Ying (1993). The proof of (10) relies heavily on the fact that the inverse covariance matrix corresponding to the finite dimensional distribution of an Ornstein-Uhlenbeck process in one dimension is tridiagonal. This property is lost as soon as α_k's different from one are considered, which makes any proof of similar results for these cases impossible. It is shown in Abt and Welch (1998) that Σ in (10) satisfies $\Sigma = \lim_{n\to\infty} n^{d-1} \mathcal{I}^{-1}_{\sigma^2,\theta}$. Therein $\mathcal{I}_{\sigma^2,\theta}$ is the Fisher information matrix of $(\sigma^2, \theta^T)^T$. It is worth noting that the same relation between the covariance matrix of the asymptotic distribution of $(\hat{\sigma}^2, \hat{\theta}^T)^T$ and the inverse of the Fisher information matrix can be shown for the product triangular correlation $\mathrm{Corr}(Z(s), Z(t)) = \prod_{k=1}^d (1 - \theta_k |s_k - t_k|)$. This however exhausts the possibilities for theoretical considerations. For the product Gaussian correlation, the simulation results given in Abt and Welch (1998) indicate that $\mathcal{I}^{-1}_{\sigma^2,\theta}$ provides a good approximation for the mean squared error matrix of $(\hat{\sigma}^2, \hat{\theta}^T)^T$.

3. Computing the Information Matrix

We only consider product designs of the form $D_n^d = \{(i-1)/(n-1) : i = 1, \ldots, n\}^d$ of size n^d in $[0,1]^d$. Taking partial derivatives of the log-likelihood

function

$$L(\beta,\sigma^2,\delta) = -\frac{1}{2}\Big(n^d\log(2\pi)+n^d\log\sigma^2+\log\det(R_\delta)+\frac{1}{\sigma^2}(y_D-\beta\mathbf{1})^T R_\delta^{-1}(y_D-\beta\mathbf{1})\Big)$$

with respect to β,σ^2, and δ, we find the information matrix of β,σ^2, and δ to be

$$\mathcal{I}_{\beta,\sigma^2,\delta} = \begin{pmatrix} \mathbf{1}^T R_\delta^{-1}\mathbf{1}/\sigma^2 & 0 & 0 \\ 0 & n^d/(2\sigma^4) & i_{\sigma^2,\delta}^T \\ 0 & i_{\sigma^2,\delta} & \mathcal{I}_\delta \end{pmatrix},$$

wherein $(\mathcal{I}_\delta)_{k,\ell} = \operatorname{tr}(R_\delta^{-1}R_\delta^{(k)}R_\delta^{-1}R_\delta^{(\ell)})/2$ and $(i_{\sigma^2,\delta})_k = \operatorname{tr}(R_\delta^{-1}R_\delta^{(k)})/(2\sigma^2)$ for $k,\ell = 1,\dots,d$ if the α_k's are known and $k,\ell = 1,\dots,2d$ if the α_k's are unknown. From this, the information matrix $\mathcal{I}_{\sigma^2,\delta}$ of σ^2 and δ is given by

$$\mathcal{I}_{\sigma^2,\delta} = \begin{pmatrix} n^d/(2\sigma^4) & i_{\sigma^2,\delta}^T \\ i_{\sigma^2,\delta} & \mathcal{I}_\delta \end{pmatrix}.$$

Denoting by $C_{\bar{1},\bar{1}}$ the matrix obtained from a matrix C after removing the 1-st row and the 1-st column, we suggest using $\Gamma_\delta = (\mathcal{I}_{\sigma^2,\delta}^{-1})_{\bar{1},\bar{1}}$ as an approximation of the mean squared error matrix of $\hat{\delta}$.

In the setup of product designs, some further simplifications are possible. The correlation matrix R_δ can be written as a Kronecker product, $R_\delta = \otimes_{k=1}^d R_k$, wherein the matrices R_k are given by $(R_k)_{i,j} = \exp(-\theta_k|(i-j)/(n-1)|^{2-\alpha_k})$ for $i,j = 1,\dots,n$. From this it can be verified that, if all α_k's are unknown,

$$\Big(\mathcal{I}_{\sigma^2,\delta}^{-1}\Big)_{\bar{1},\bar{1}} = \begin{pmatrix} A_{\theta\theta} & A_{\theta\alpha} \\ A_{\theta\alpha} & A_{\alpha\alpha} \end{pmatrix}^{-1},$$

wherein $A_{\theta\theta}, A_{\theta\alpha}$, and $A_{\alpha\alpha}$ are $d\times d$ diagonal matrices given by

$$\begin{aligned} A_{\theta\theta} &= \operatorname{diag}\Big(n^{d-2}(ng_{\theta\theta}^{(k)} - h_\theta^{(k)}h_\theta^{(k)})/2, k = 1,\dots,d\Big), \\ A_{\theta\alpha} &= \operatorname{diag}\Big(n^{d-2}(ng_{\theta\alpha}^{(k)} - h_\theta^{(k)}h_\alpha^{(k)})/2, k = 1,\dots,d\Big), \\ A_{\alpha\alpha} &= \operatorname{diag}\Big(n^{d-2}(ng_{\alpha\alpha}^{(k)} - h_\alpha^{(k)}h_\alpha^{(k)})/2, k = 1,\dots,d\Big), \end{aligned}$$

with

$$\begin{aligned} g_{\theta\theta}^{(k)} &= \operatorname{tr}\Big(R_k^{-1}\Big(\frac{\partial}{\partial\theta_k}R_k\Big)R_k^{-1}\Big(\frac{\partial}{\partial\theta_k}R_k\Big)\Big), h_\theta^{(k)} = \operatorname{tr}\Big(R_k^{-1}\Big(\frac{\partial}{\partial\theta_k}R_k\Big)\Big), \\ g_{\alpha\alpha}^{(k)} &= \operatorname{tr}\Big(R_k^{-1}\Big(\frac{\partial}{\partial\alpha_k}R_k\Big)R_k^{-1}\Big(\frac{\partial}{\partial\alpha_k}R_k\Big)\Big),\ h_\alpha^{(k)} = \operatorname{tr}\Big(R_k^{-1}\Big(\frac{\partial}{\partial\alpha_k}R_k\Big)\Big), \\ g_{\theta\alpha}^{(k)} &= \operatorname{tr}\Big(R_k^{-1}\Big(\frac{\partial}{\partial\theta_k}R_k\Big)R_k^{-1}\Big(\frac{\partial}{\partial\alpha_k}R_k\Big)\Big). \end{aligned}$$

This makes the numerical evaluation of Γ_δ rather simple for product designs. If $\alpha_1, \ldots, \alpha_d$ are all known, we are left with $(\mathcal{I}^{-1}_{\sigma^2,\theta})_{\bar{1},\bar{1}} = A^{-1}_{\theta\theta}$.

4. Simulation Results

In the following we report simulation results on the quality of the mean squared error approximation given in (9) for known as well as unknown parameters $\alpha_1, \ldots, \alpha_d$. For dimensions $d = 1, 2$, a regular grid is laid out over the region $T = [0,1]^d$ of interest and a design D is chosen as a subset of T. On the grid, N sets of observations from the finite dimensional marginal distribution of a Gaussian stochastic process are generated for each considered combination of parameters. We chose:

(a) in one dimension: $D = \{0, 1, (2i+1)/14 : i = 1, \ldots, 5\}$ and $T = \{i/14 : i = 0, \ldots, 14\}$ with $N = 2500$;
(b) in two dimensions: $D = \{i/4 : i = 0, \ldots, 4\}^2$ and $T = \{1/4, 3/4, i/10 : i = 0, \ldots, 10\}^2$ with $N = 1500$.

Simulations were also carried out in three dimensions. As the results were not found to be qualitatively different from what we report below, they are omitted here.

All data were obtained taking $\sigma^2 = 1$ (which is always assumed to be unknown), as we found that changing σ^2 leads merely to rescaling of the empirical mean squared errors and the approximations. The simulations are based on $\beta = 0$, which is considered to be unknown as well. The observations corresponding to locations in D are used to estimate the parameters $\beta, \sigma^2, \theta_1, \ldots, \theta_d$, and $\alpha_1, \ldots, \alpha_d$. With $m = |T|$ and denoting the locations in T by $s^{(1)}, \ldots, s^{(m)}$, we look at the maximum empirical mean squared error

$$\text{MEMSE} = \max\Big\{\frac{1}{N}\sum_{j=1}^{N}\Big(\hat{Y}^{(j)}_{\hat{\delta}}(s^{(i)}) - Y^{(j)}(s^{(i)})\Big)^2 : i = 1, \ldots, m\Big\} \tag{11}$$

and the integrated empirical mean squared error

$$\text{IEMSE} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N}\sum_{j=1}^{N}\Big(\hat{Y}^{(j)}_{\hat{\delta}}(s^{(i)}) - Y^{(j)}(s^{(i)})\Big)^2. \tag{12}$$

Therein, for $j = 1, \ldots, N$, the quantity $Y^{(j)}(s)$ is the jth simulation of the Gaussian process Y at location $s \in T$ and $\hat{Y}^{(j)}_{\hat{\delta}}(s)$ is the predictor (3) obtained for $s \in T$ using the data $\{Y^{(j)}(t) : t \in D\}$ and the parameters estimated from these.

The maximum likelihood estimates were obtained using the software package GaSP developed by Welch (1995). As a function of the correlation parameters δ, explicit formulas can be obtained for $\hat{\beta}$ and $\hat{\sigma}^2$. Numerical maximization

of $L(\hat{\beta}, \hat{\sigma}^2, \delta)$ with respect to δ then uses Powell's conjugate directions method (see Press et al. (1988), Section 10.5). The constraints on the range of $\theta_1, \ldots, \theta_k$ and $\alpha_1, \ldots, \alpha_k$ are dealt with by suitable transformations that then allow an unconstrained optimization.

In the tables that follow, the first line gives the MEMSE and the IEMSE. A raised number x after the decimal point indicates multiplication by 10^{-x}. The second line gives

$$\max_{i=1,\ldots,m} \mathrm{MSE}\Big(\hat{Y}_\delta\Big(s^{(i)}\Big)\Big)/\mathrm{MEMSE} \quad \text{and} \quad \frac{1}{m}\sum_{i=1}^{m} \mathrm{MSE}\Big(\hat{Y}_\delta\Big(s^{(i)}\Big)\Big)/\mathrm{IEMSE}, \quad (13)$$

whereas in each table the third line considers the full approximation as given in (9) and shows

$$\max_{i=1,\ldots,m} \mathrm{AMSE}\Big(\hat{Y}_{\hat{\delta}}\Big(s^{(i)}\Big)\Big)/\mathrm{MEMSE} \text{ and } \frac{1}{m}\sum_{i=1}^{m} \mathrm{AMSE}\Big(\hat{Y}_{\hat{\delta}}\Big(s^{(i)}\Big)\Big)/\mathrm{IEMSE}. \quad (14)$$

Values for these ratios that are less than one indicate that the corresponding approximations underestimate the true mean squared error, whereas values greater than one would indicate approximations that are too conservative.

4.1. Simulation results for known parameters $\alpha_1, \ldots, \alpha_d$

In one dimension, it is well known that for $\alpha = 1$ the parameters θ and σ^2 cannot be identified, even if an entire path of the stochastic process Y would be available (see Jørsboe (1968), page 58). Numerical computation of the maximum likelihood estimates is nevertheless possible, but θ is overestimated severely and the approximate Γ_θ for the mean squared error of $\hat{\theta}$ as provided by the inverse information matrix is very poor (see Abt and Welch (1998)). We therefore do not investigate this scenario any further.

There are no identifiability problems when we assume $\alpha = 0$ (see Ibragimov and Rozanov (1978), page 95). For $\theta = 3, 5, 7.5, 25$, and 50, Table 1 summarizes the results. The smallest value of θ was chosen in a way such that the Choleski decomposition used for the simulations was numerically well conditioned. The results indicate that the approximation provided by $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$ in (9) performs rather well, whereas, as the ratios in the second line are much smaller than one, $\mathrm{MSE}(\hat{Y}_\delta(t))$ performs poorly. Thus there is considerable loss by ignoring the fact that the correlation parameters are estimated.

In two and more dimensions, all parameters are identifiable. For $\alpha_1 = \alpha_2 = 1$, the results are depicted in Table 2. For small values of θ_1 and θ_2, the differences between $\mathrm{MSE}(\hat{Y}_\delta(t))$ and $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$ are negligible. For increasing values of θ_1 or θ_2, the usage of $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$ proves to be slightly superior.

Table 1. Simulation results for the Gaussian correlation in one dimension for known α. A raised number x after the decimal point indicates that the empirical mean squared error has to be multiplied by 10^{-x}.

	$\theta = 3.0$		$\theta = 5.0$		$\theta = 7.5$		$\theta = 25.0$		$\theta = 50.0$	
	Max	Int	Max	Int	Max	Int	Max	Int	Max	Int
EMSE	$0.^{5}693$	$0.^{5}118$	$0.^{3}133$	$0.^{4}236$	$0.^{2}103$	$0.^{3}194$	$0.^{1}925$	$0.^{1}235$	0.374	0.133
MSE($\hat{Y}_{\hat{\delta}}(t)$)	0.715	0.722	0.743	0.748	0.813	0.815	0.865	0.886	0.825	0.778
AMSE($\hat{Y}_{\hat{\delta}}(t)$)	0.944	0.952	0.971	0.976	1.047	1.045	1.027	1.029	1.010	0.953

Table 2. Simulation results for the product Ornstein-Uhlenbeck process in two dimensions for known α_1 and α_2.

	$\theta_1 = \theta_2 = 0.01$		$\theta_1 = \theta_2 = 0.1$		$\theta_1 = \theta_2 = 1.0$		$\theta_1 = \theta_2 = 4.0$		$\theta_1 = \theta_2 = 10.0$	
	Max	Int	Max	Int	Max	Int	Max	Int	Max	Int
EMSE	$0.^{2}261$	$0.^{2}124$	$0.^{1}258$	$0.^{1}124$	0.239	0.119	0.791	0.445	1.085	0.778
MSE($\hat{Y}_{\hat{\delta}}(t)$)	0.918	0.996	0.925	0.989	0.942	0.997	0.881	0.921	0.918	0.907
AMSE($\hat{Y}_{\hat{\delta}}(t)$)	0.918	0.996	0.925	0.989	0.942	0.997	0.890	0.931	0.971	1.026

	$\theta_1 = 0.01, \theta_2 = 0.1$		$\theta_1 = 0.01, \theta_2 = 1.0$		$\theta_1 = 0.01, \theta_2 = 4.0$		$\theta_1 = 0.01, \theta_2 = 10.0$	
	Max	Int	Max	Int	Max	Int	Max	Int
EMSE	$0.^{1}138$	$0.^{2}679$	0.123	$0.^{1}606$	0.479	0.246	0.971	0.529
MSE($\hat{Y}_{\hat{\delta}}(t)$)	0.953	0.997	0.983	1.020	0.939	0.948	0.900	0.886
AMSE($\hat{Y}_{\hat{\delta}}(t)$)	0.953	0.997	0.983	1.020	0.948	0.955	1.021	0.989

Table 3. Simulation results for the product Gaussian correlation in two dimensions for known α_1 and α_2.

	$\theta_1 = \theta_2 = 2.5$		$\theta_1 = \theta_2 = 5.0$		$\theta_1 = \theta_2 = 7.5$		$\theta_1 = \theta_2 = 10.0$	
	Max	Int	Max	Int	Max	Int	Max	Int
EMSE	$0.^{3}387$	$0.^{4}880$	$0.^{2}652$	$0.^{2}166$	$0.^{1}260$	$0.^{2}757$	$0.^{1}632$	$0.^{1}206$
MSE($\hat{Y}_{\hat{\delta}}(t)$)	0.868	0.887	0.871	0.903	0.899	0.917	0.876	0.885
AMSE($\hat{Y}_{\hat{\delta}}(t)$)	0.945	0.962	0.936	0.960	0.954	0.956	0.921	0.911

Table 3 shows the results in two dimensions for $\alpha_1 = \alpha_2 = 0$. The approximation (9) leads to a clear improvement over (4). Again, the choice of $\theta_1 = \theta_2 = 2.5$ as the smallest values considered was made to avoid numerical singularity problems of the Choleski factor. The quality of the results is the same when θ_1 and θ_2 are chosen to be different and thus omitted.

Large values of θ_1 or θ_2, like $\theta_1 = \theta_2 = 20$ or more lead to weakly correlated observations. In these cases and for the given design used here, which is equispaced in each dimension, the algorithm has difficulties in identifying the true values of θ_1 and θ_2 and also the variability of their maximum likelihood

estimates tends to be overestimated by the inverse Fisher information matrix. As a consequence, $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$ becomes too large. Larger designs that include nearby observations are required in these situations. We tried a design of size 57 that includes locations at various distances and found that then also large θ_k's and their standard errors are estimated well. Due to this, the approximation (9) again turned out to be in good agreement with the true mean squared error. Generally, it was observed that the abovementioned effect large θ_k's have on $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$ is less severe if the sample path of the stochastic process Y is smooth, i.e., $\alpha_1 = \alpha_2 = 0$.

4.2. Simulation results for unknown parameters $\alpha_1, \ldots, \alpha_d$

In this section, we have $\delta = (\theta_1, \ldots, \theta_d, \alpha_1, \ldots, \alpha_d)$. We first look at the situation $\alpha_1 = \cdots = \alpha_d = 0$, when these parameters take values on the boundary of their range.

Using $(\mathcal{I}^{-1}_{\sigma^2,\delta})_{\bar{1},\bar{1}}$ as an approximation for the mean squared error matrix of $\hat{\delta}$ requires that the true parameter δ is in the interior of the parameter space. If the true parameters are on the boundary, the variability in the maximum likelihood estimates is expected to be smaller than the approximation provided by $(\mathcal{I}^{-1}_{\sigma^2,\delta})_{\bar{1},\bar{1}}$. Modifications of $\mathcal{I}_{\sigma^2,\delta}$ to adjust for this problem are suggested in Moran (1971). Before we briefly outline these for $d = 1$ and $d = 2$, we need some results on the moments of mixtures of distributions as well as the moments of truncated normal distributions.

Let $P_1, \ldots, P_s$ be probability distributions on $\mathbb{R}^n$ and $\gamma_1, \ldots, \gamma_s \in (0, 1)$ constants with $\sum_{i=1}^s \gamma_i = 1$. Then $P = \sum_{i=1}^s \gamma_i P_i$ is called the mixture of $P_1, \ldots, P_s$ and is a probability distribution on $\mathbb{R}^n$ as well. Denoting by $\mathrm{E}(P)$ and $\mathrm{Var}(P)$ the expectation vector and the covariance matrix of an n-dimensional random vector with distribution P, we have

$$\mathrm{E}(P) = \sum_{i=1}^{s} \gamma_i \mathrm{E}(P_i) \tag{15}$$

and it can be shown that

$$\mathrm{Var}(P) = \sum_{i=1}^{s} \gamma_i \Big(\mathrm{Var}(P_i) + \mathrm{E}(P_i)\mathrm{E}(P_i)^T \Big) - \sum_{i,j=1}^{s} \gamma_i \gamma_j \mathrm{E}(P_i)\mathrm{E}(P_j)^T. \tag{16}$$

As a next step, we need to find the first and second moment of a truncated normal distribution. Let R be a positive definite $n \times n$ matrix and let $(R)_{i,j} = \rho_{i,j}$ for $i, j = 1, \ldots, n$ and $\rho_{i,i} = 1$ for $i = 1, \ldots, n$. For index sets $I, J \subseteq \{1, \ldots, n\}$ we mean by $R_{I,J}$ the matrix obtained from R by selecting the rows with subscripts in I and the columns with subscripts in J. For an index set I let $\bar{I} = \{1, \ldots, n\} \setminus I$.

For a set consisting of one point only, say $I = \{k\}$, we use the abbreviation $\bar{k}$ for $\overline{\{k\}}$. With this, the $2 \times (n-2)$ matrix obtained from R by selecting the kth and ℓth rows and all but the kth and ℓth columns is denoted by $R_{k\ell,\bar{k}\bar{\ell}}$. Similarly, $x_{k\ell}$ would be the vector $(x_k, x_\ell)^T$, whereas $x_{\bar{k}\bar{\ell}}$ is the vector x with components x_k and x_ℓ removed. Furthermore, $\phi_n(x; R)$ denotes the n-variate normal density with mean zero, variance one and correlation matrix R. For $b \in \mathbb{R}^n$ we define

$$\Phi_n(b; R) = \int_{b_1}^{\infty} \cdots \int_{b_n}^{\infty} \phi_n(x; R)\, dx_n \cdots dx_1 = \int_{[b,\infty)} \phi_n(x; R)\, dx.$$

With this notation, we have the following Lemma, the proof of which can be found in Tallis (1961).

Lemma. *Let $Z_1, \ldots, Z_n$ be real valued random variables defined on some probability space $(\Omega, \mathcal{A}, Q)$, whose joint distribution has density $\phi_n(z; R)$ as defined above. By $Q[Z_i]$ we mean the distribution of Z_i on $\mathbb{R}$ induced by Q. For an index set $K \subseteq \{1, \ldots, n\}$ and $k, \ell \in K$ let $K_k = K \setminus \{k\}$ and $K_{k\ell} = K \setminus \{k, \ell\}$. If $Y_1, \ldots, Y_n$ are random variables on $(\Omega, \mathcal{A}, Q)$ with*

$$Q[Y_1, \ldots, Y_n] = Q[Z_1, \ldots, Z_n | Z_K \geq 0],$$

we have

$$\mathrm{E}(Y_i) = \frac{1}{\alpha_K \sqrt{2\pi}} \sum_{k \in K} \rho_{i,k} \Phi_{|K|-1}\Big(0; R_{K_k,K_k} - R_{K_k,k} R_{K_k,k}^T\Big), \tag{17}$$

and

$$\mathrm{E}(Y_i Y_j) = \rho_{i,j} + \frac{1}{\alpha_K 2\pi} \sum_{k \in K} \Big\{ \rho_{k,i} \sum_{\ell \in K_k} \frac{\rho_{\ell,j} - \rho_{k,\ell}\rho_{k,j}}{\sqrt{1 - \rho_{k,\ell}^2}} \Phi_{|K|-2}(0; \tilde{R}_{k,\ell}) \Big\}, \tag{18}$$

with $\tilde{R}_{k,\ell} = R_{K_{k\ell},K_{k\ell}} - R_{K_{k\ell},k\ell} R_{k\ell,k\ell}^{-1} R_{K_{k\ell},k\ell}^T$. Therein $\Phi_0(0; \cdot) = 1, \Phi_n(0; \cdot) = 0$ for $n < 0$, and $\alpha_K = \Phi_{|K|}(0; R_{K,K})$.

As it is seen, tail probabilities of the form $\Phi_n(0; R)$ are required for various values of n. Results are available for $n = 1, 2, 3$ in Tong (1990), Section 8.2.2. Our implementation is so far limited to these situations. A recursive formula reducing the n-dimensional integration to one-dimensional integrations is given in Plackett (1954).

We are now ready to outline the modifications of the inverse Fisher information matrix in order to use it as an approximation for the mean squared error matrix of $\hat{\delta}$ if the parameters $\alpha_1, \ldots, \alpha_d$ take their boundary value zero. From Section 3 we know that, as σ^2 is unknown, the matrix

$$A = \begin{pmatrix} A_{\theta\theta} & A_{\theta\alpha} \\ A_{\theta\alpha} & A_{\alpha\alpha} \end{pmatrix} \in \mathbb{R}^{2d,2d}$$

is the information matrix of δ. Define the random vector X as having a $2d$-dimensional normal distribution with mean zero and covariance matrix A, and let $Z = A^{-1}X$. Then the distribution of Z is multivariate normal with mean zero and covariance matrix A^{-1}. Think of Z as the parameter vector δ. Again we assume all random variables to be defined on a probability space $(\Omega, \mathcal{A}, Q)$.

In the one-dimensional case, $d = 1$, we have $\delta = (\theta, \alpha)$. Suppose the true (but unknown) value of the parameter α is zero. Then, applying the results in Moran (1971) to this situation, an approximation for the covariance matrix of $\hat{\delta}$ is found as the second moment of a distribution P on $\mathbb{R}^2$ given by the mixture $P = (P_0 + P_1)/2$, with P_0 and P_1 both being probability distributions on $\mathbb{R}^2$. Therein, $P_0 = Q[Z|Z_2 \geq 0]$ is the conditional distribution of Z given $Z_2 \geq 0$, and P_1 is concentrated on $\mathbb{R} \times \{0\}$, the nonsingular part being $Q[\tilde{Z}_1|Z_2 \leq 0]$ with $\tilde{Z}_1 = (A_{\bar{2},\bar{2}})^{-1}X_1$. Using the previous results given in (15) to (18), the second moment of P is readily derived.

For two dimensions, $d = 2$, let $\delta = (\theta_1, \theta_2, \alpha_1, \alpha_2)$ and assume the true underlying values for α_1 and α_2 are zero. Then the approximate covariance matrix of $\hat{\delta}$ can be found as the second moment of the mixture $P = \gamma_0 P_0 + \gamma_1 P_1 + \gamma_2 P_2 + \gamma_{12} P_{12}$ of probability distributions P_0, P_1, P_2, P_{12} on $\mathbb{R}^4$. Therein $\gamma_0, \gamma_1, \gamma_2, \gamma_{12} \geq 0$ and they sum up to one and $P_0 = Q[Z|Z_3 \geq 0, Z_4 \geq 0]$ with $\gamma_0 = Q(Z_3 \geq 0, Z_4 \geq 0)$. With $\tilde{Z}^{(3)} = (A_{\bar{3},\bar{3}})^{-1}X_{\bar{3}}$, the measure P_1 is a probability distribution on $\mathbb{R}^2 \times \{0\} \times (0, \infty)$, the nonsingular part of which is given by $Q[\tilde{Z}^{(3)}|Z_3 \leq 0, \tilde{Z}_3^{(3)} \geq 0]$ and $\gamma_1 = Q(Z_3 \leq 0, \tilde{Z}_3^{(3)} \geq 0)$. Similarly, the nonsingular part of the distribution P_2 on $\mathbb{R}^2 \times (0, \infty) \times \{0\}$ is found as $Q[\tilde{Z}^{(4)}|\tilde{Z}_3^{(4)} \geq 0, Z_4 \leq 0]$ and $\gamma_2 = Q(\tilde{Z}_3^{(4)} \geq 0, Z_4 \leq 0)$, with $\tilde{Z}^{(4)} = (A_{\bar{4},\bar{4}})^{-1}X_{\bar{4}}$. The probability distribution P_{12} finally is defined on $\mathbb{R}^2 \times \{0\}^2$, the part on $\mathbb{R}^2$ being $Q[(A_{\bar{3}\bar{4},\bar{3}\bar{4}})^{-1}X_{\bar{3}\bar{4}}|\tilde{Z}_3^{(3)} \leq 0, \tilde{Z}_3^{(4)} \leq 0]$ and $\gamma_{12} = Q(\tilde{Z}_3^{(3)} \leq 0, \tilde{Z}_3^{(4)} \leq 0)$. As we are looking at product designs only, it can be shown from equations (56) to (58) in Moran (1971) that $\gamma_0 = \gamma_1 = \gamma_2 = \gamma_{12} = 1/4$. Appropriate modifications of the proof of Theorems I and II in Moran (1971) allow immediate generalization to three and more dimensions. In general, mixtures of 2^d distributions will arise.

Denoting the covariance matrix of the distribution P obtained in this way by $\Gamma_{\delta,b}$, where we use the subscript b to indicate the boundary situation, we suggest using the approximation (9) with Γ_δ replaced by $\Gamma_{\delta,b}$ in cases where the true α_k's are assumed to lie on the boundary of their range. It will be denoted by $\mathrm{AMSE}_b(\hat{Y}_{\hat{\delta}}(t))$. We note that the diagonal elements of $\Gamma_{\delta,b}$ are found to be smaller than those in Γ_δ, thus indicating smaller variability in the $\hat{\alpha}_k$'s if true values are on the boundary.

The usage of $\mathrm{AMSE}_b(\hat{Y}_{\hat{\delta}}(t))$ instead of $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$ is investigated by simulation for the Gaussian correlation function. Tables 4 and 5 summarize the results for one and two dimensions.

Table 4. Simulation results for the Gaussian correlation in one dimension for unknown α.

	$\theta = 3.0$		$\theta = 5.0$		$\theta = 7.5$		$\theta = 25.0$		$\theta = 50.0$	
	Max	Int	Max	Int	Max	Int	Max	Int	Max	Int
EMSE	$0.^{4}475$	$0.^{5}863$	$0.^{3}531$	$0.^{4}975$	$0.^{2}276$	$0.^{3}583$	0.146	0.051	0.514	0.202
$\mathrm{MSE}(\hat{Y}_{\hat{\delta}}(t))$	0.138	0.130	0.218	0.194	0.304	0.271	0.547	0.408	0.601	0.512
$\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$	2.486	2.606	2.482	2.503	2.420	2.498	2.166	1.848	4.307	4.209
$\mathrm{AMSE}_b(\hat{Y}_{\hat{\delta}}(t))$	0.967	1.001	1.034	1.020	1.083	1.081	1.033	0.942	1.936	1.849

Table 5. Simulation results for the product Gaussian correlation in two dimensions for unknown α_1 and α_2.

	$\theta_1 = \theta_2 = 2.5$		$\theta_1 = \theta_2 = 5.0$		$\theta_1 = \theta_2 = 7.5$		$\theta_1 = \theta_2 = 10.0$	
	Max	Int	Max	Int	Max	Int	Max	Int
EMSE	$0.^{3}680$	$0.^{3}142$	$0.^{2}872$	$0.^{2}229$	$0.^{1}336$	$0.^{1}111$	$0.^{1}913$	$0.^{1}360$
$\mathrm{MSE}(\hat{Y}_{\hat{\delta}}(t))$	0.494	0.549	0.651	0.652	0.694	0.626	0.607	0.547
$\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$	2.031	2.059	1.820	1.654	1.788	1.514	1.701	1.563
$\mathrm{AMSE}_b(\hat{Y}_{\hat{\delta}}(t))$	1.047	1.094	1.081	1.021	1.095	0.946	1.000	0.904

It is clearly seen, that $\mathrm{MSE}(\hat{Y}_{\hat{\delta}}(t))$ seriously underapproximates the maximum as well as the integrated mean squared error. Using $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$, and thus ignoring the fact that the true parameters are on the boundary of the parameter space, leads to a severe overapproximation of the true mean squared errors. However, the approximation $\mathrm{AMSE}_b(\hat{Y}_{\hat{\delta}}(t))$ based on $\Gamma_{\delta,b}$ captures the maximum and the integrated mean squared error rather well.

Similarly to the discussion at the end of Subsection 4.1, Table 4 indicates that large values of at least one θ_k may inflate the approximation $\mathrm{AMSE}_b(\hat{Y}_{\hat{\delta}}(t))$. However, $\mathrm{MSE}(\hat{Y}_{\hat{\delta}}(t))$ underestimates the true mean squared error. Similar behavior, but not shown in the tables, can be observed for $\theta_1 = \theta_2 = 20$ or $\theta_1 = \theta_2 = 30$ in two dimensions as well. The reason again is in the overestimation of the variance of $\hat{\theta}_k$ and $\hat{\alpha}_k$ by the information matrix. For weak correlations, the 25 point design we used is not suitable for estimating the correlation structure properly. Increasing the number of design points and including nearby locations shows that the inverse of the Fisher information matrix again provides a very accurate approximation of the standard errors of $\hat{\theta}_k$ and $\hat{\alpha}_k$ and, as a consequence, the mean squared errors are accurately approximated by $\mathrm{AMSE}_b(\hat{Y}_{\hat{\delta}}(t))$ in case of $\alpha_1 = \alpha_2 = 0$ and large values for θ_1 or θ_2.

Table 6 shows the simulation results for unknown $\alpha_1 = \alpha_2 = 1$. As the true values of α_1 and α_2 are inside their range $[0, 2)$, we can use $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$. For small values of θ_1 and θ_2, the difference in the approximations seems to be negligible, with $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$ performing better. For larger values of one or both

of θ_1 and θ_2, slight superiority of $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$ becomes clear. Also here, a similar discussion as above applies for large values of θ_1 or θ_2. As opposed to Table 2, the overapproximation of $\mathrm{MSE}(\hat{Y}_{\delta}(t))$ is already observed for $\theta_1 = \theta_2 = 10$. This is because based on the same number of observations α_1 and α_2 need also to be estimated here.

Table 6. Simulation results for the product Ornstein-Uhlenbeck process in two dimensions for unknown α_1 and α_2.

	$\theta_1 = \theta_2 = 0.01$		$\theta_1 = \theta_2 = 0.1$		$\theta_1 = \theta_2 = 1.0$		$\theta_1 = \theta_2 = 4.0$		$\theta_1 = \theta_2 = 10.0$	
	Max	Int	Max	Int	Max	Int	Max	Int	Max	Int
EMSE	$0.^{2}269$	$0.^{2}130$	$0.^{1}269$	$0.^{1}132$	0.256	0.129	0.812	0.474	1.157	0.822
$\mathrm{MSE}(\hat{Y}_{\delta}(t))$	0.890	0.946	0.886	0.928	0.876	0.921	0.859	0.865	0.861	0.858
$\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$	0.906	0.961	0.903	0.944	0.913	0.956	1.064	1.124	25.003	18.948

	$\theta_1 = 0.01, \theta_2 = 0.1$		$\theta_1 = 0.01, \theta_2 = 1.0$		$\theta_1 = 0.01, \theta_2 = 4.0$		$\theta_1 = 0.01, \theta_2 = 10.0$	
	Max	Int	Max	Int	Max	Int	Max	Int
EMSE	$0.^{1}143$	$0.^{2}717$	0.134	$0.^{1}662$	0.503	0.266	1.005	0.555
$\mathrm{MSE}(\hat{Y}_{\delta}(t))$	0.924	0.943	0.899	0.935	0.894	0.876	0.870	0.845
$\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$	0.942	0.959	0.939	0.971	1.184	1.157	24.920	21.025

5. Conclusions

The approximation $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$ for the mean squared prediction error that accounts for the fact that the correlation parameters are estimated turns out to be superior to the usage of $\mathrm{MSE}(\hat{Y}_{\delta}(t))$. However, in all cases, large values of the parameters θ_k lead to overapproximation of the mean squared error by $\mathrm{AMSE}(\hat{Y}_{\hat{\delta}}(t))$. This is due to the fact that the observations are weakly correlated in this case, which, in turn, may lead to overestimation of the variability of $\hat{\theta}_k$ and $\hat{\alpha}_k$. Increasing the design size and the inclusion of nearby observations for model fitting overcomes this problem and is therefore necessary. In general, using different designs for model fitting and prediction can be beneficial. Special care is also needed when the true values of some of the parameters take values on the boundary of the parameter space. As suggested, a modification of the inverse Fisher information matrix can be used as an approximation for the mean squared error matrix of the maximum likelihood estimates in (9).

We are aware of the limitations of the approach when it comes to the usage of the approximation in the optimal design criteria (5). Exact formulas for the tail probabilities are available for up to three dimensions only. Therefore the investigation in this paper is limited to these cases. Especially, the computation of $\Gamma_{\delta,b}$ becomes very time intensive for dimensions larger than three, as numerical integration routines are needed to evaluate the multivariate normal tail probabilities. The moments of a mixture of 2^d truncated normal distributions would be

required in each iteration of an optimal design algorithm. Using the approximation for optimal design purposes is thus prohibitive. As we found no qualitative difference in the results for dimensions ranging from one to three, we expect the performance of the approximation not to be affected by the dimensionality of the problem.

More relevant for practical applications seems to be the accurate estimation of $\mathrm{MSE}(\hat{Y}_{\hat{\delta}}(t))$ to find prediction intervals as indicated in (6). The design would be fixed, and the approximation of the mean squared error then needs to be evaluated once only for each site t. We suggest using the approximation (9) with δ replaced by the maximum likelihood estimate $\hat{\delta}$. Prior knowledge of the response Y as well as the parameter estimates $\hat{\alpha}_k$ obtained from fitting the model might help to decide whether to use Γ_δ or $\Gamma_{\delta,b}$ as an approximation for the mean squared error matrix of $\hat{\delta}$. The suggested methodology can be readily adapted to models similar to (1) that also include an additional random error term. Many practical situations, as they arise in environmental statistics or the geostatistical literature are thus covered. Further work will be undertaken in this direction.

Acknowledgements

I acknowledge financial support under the postdoctoral program of the Deutsche Forschungsgemeinschaft in Germany, as well as from the University of Waterloo in Canada. I am very grateful to Professor William J. Welch for his valuable suggestions and for making his software package GaSP available to me. I also like to express my thanks to two anonymous referees for helpful comments on an earlier version of the paper.

References

Abt, M. and Welch, W. J. (1998). Fisher information and maximum likelihood estimation of covariance parameters in Gaussian stochastic processes. *Canad. J. Statist.*, to appear.

Harville, D. A. and Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *J. Amer. Statist. Assoc.* **87**, 724-731.

Ibragimov, I. A. and Rozanov, Y. A. (1978). *Gaussian Random Processes.* Springer, New York.

Jørsboe, O. G. (1968). *Equivalence or Singularity of Gaussian Measures on Function Spaces.* Various Publications Series No. 4, Matematisk Institut, Aarhus Universitet.

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Amer. Statist. Assoc.* **79**, 853-862.

Loève, M. (1978). *Probability Theory II.* Springer, New York.

Moran, P. A. P. (1971). Maximum-likelihood estimation in non-standard conditions. *Proc. Camb. Phil. Soc.* **70**, 441-450.

Plackett, R. L. (1954). A reduction formula for normal multivariate integrals. *Biometrika* **41**, 351-360.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1988). *Numerical Recipes in C.* Cambridge University Press, Cambridge.

Ross, S. M. (1983). *Stochastic Processes.* Wiley, New York.

Sacks, J., Schiller, S. B. and Welch, W. J. (1989). Designs for computer experiments. *Technometrics* **31**, 41-47.

Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution. *J. Roy. Statist. Soc. Ser. B* **23**, 223-229.

Tong, Y. L. (1990). *The Multivariate Normal Distribution.* Springer, New York.

Welch, W. J. (1995). *GaSP* (*Gaussian Stochastic Process*) Version 2.0 Beta, Unpublished Software.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J. and Morris, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics* **34**, 15-25.

Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions I, Basic Results.* Springer, New York.

Ying, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *Ann. Statist.* **21**, 1567-1590.

Zimmerman, D. L. and Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math.* **44**, 27-43.

Department of Statistics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.

(On leave from: Institut für Mathematik, Universität Augsburg, D-86135 Augsburg, Germany.)

E-mail: abt@uni-augsburg.de

(Received August 1996; accepted June 1997)