Improving a surrogate model in uncertainty quantification by real data *

Michael Kohler[1] and Adam Krzyżak [2,†]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: kohler@mathematik.tu-darmstadt.de*

[2] *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8,email: krzyzak@cs.concordia.ca*

March 31, 2017

Abstract

Quantification of uncertainty of a technical system is often based on a surrogate model of a corresponding simulation model. In any application the simulation model will not describe the reality perfectly, and consequently also the surrogate model will be imperfect. In this article we show how observed data of the real technical system can be used to improve such a surrogate model, and we analyze the rate of convergence of density estimates based on the improved surrogate model. The results are illustrated by applying the estimates to simulated and real data.

AMS classification: Primary 62G07; secondary 62P30.

Key words and phrases: Density estimation, imperfect models, L_1 error, surrogate models, uncertainty quantification.

1 Introduction

Any design of complex technical systems by engineers nowadays is based on some sort of mathematical model of the technical system. Such models are never able to describe the reality perfectly, therefore their analysis has to take into account some kind of uncertainty. This uncertainty might occur, e.g., because some of the parameters of the model are not exactly known, because of the use of an imperfect mathematical model of the technical system during the design process which does not really describe all aspects of the underlying technical system, or because of lack of knowledge about future use. A good quantification of the uncertainty of the system is essential in order to avoid oversizing and to conserve resources.

In this article we quantify the uncertainty of a technical system by estimating a density of a real random variable representing the outcome of an experiment with the technical

*Running title: *Improving a surrogate model*

†Corresponding author. Tel: +1-514-848-2424 ext. 3007, Fax:+1-514-848-2830

system. The starting point for our estimation problem is a stochastic model of the technical system. This stochastic model has parameters which are chosen randomly because their exact values are uncertain and consequently not known, and it computes the outcome of the technical system by computing the value of a function depending on concrete values of the parameters. In case that the distribution of the parameters is known (which we will assume from now on) and that the function, which has to be computed, is given, Monte Carlo can be used to estimate either quantiles or the density of the output of the technical system.

Usually, the stochastic model is evaluated using a computer program, and computer experiments can be used to generate values for the Monte Carlo estimates. However, it often happens that generation of the values is rather time consuming, so that standard Monte Carlo estimates cannot be applied. Instead, one has to apply techniques which are able to quantify the uncertainty in the computer experiment using only a few evaluations of the computer program. There is a vast literature on the design and analysis of such computer experiments, cf., e.g., Santner, Williams, and Notz (2003) and the literature cited therein. Often, so–called surrogate models of the computer experiment are used in order to analyze it. Surrogate models have been introduced and investigated with the aid of the simulated and real data in connection with the quadratic response surfaces in Bucher and Burgund (1990), Kim and Na (1997) and Das and Zheng (2000), in context of support vector machines in Hurtado (2004), Deheeger and Lemaire (2010) and Bourinet, Deheeger and Lemaire (2011), in connection with neural networks in Papadrakakis and Lagaros (2002), and in context of kriging in Kaymaz (2005) and Bichon et al. (2008). Consistency and rate of convergence of density estimates based on surrogate models have been studied in Devroye, Felber and Kohler (2013), Bott, Felber and Kohler (2015) and Felber, Kohler and Krzyżak (2015a). A method for the adaptive choice of the smoothing parameter of such estimates has been presented in Felber, Kohler and Krzyżak (2015b).

Of course, in practice a stochastic model will never be able to represent the real technical system perfectly. So it is clear that the mathematical model is imperfect, and consequently also any surrogate model based on the imperfect mathematical model will be imperfect. In Bayesian analysis of computer experiments, Kennedy and O'Hagan (2001), Bayarri et al. (2007), Goh et al. (2013), Han, Santner and Rawlinson (2009), Hidgon et al. (2013) and Wang, Chen and Tsui (2009) model the discrepancy between the computer experiments and the outcome of the technical system by a Gaussian process. Tuo and Wu (2015) pointed out that the this approach might fail in case of an imperfect computer model, for which there exist no values of the parameters which fit the technical system perfectly, and suggested and analyzed non-Bayesian methods for the choice of the parameters of such models.

In this article we quantify uncertainty outside the framework of Bayesian analysis. We assume that we have available an additional (small) sample of the real technical system, and we consider the problem of estimation from this sample together with the imperfect simulation model an improved surrogate model.

The mathematical setting which we consider is as follows: Let (X, Y), (X_1, Y_1), (X_2, Y_2), ... be independent and identically distributed random variables with values in $\mathbb{R}^d \times \mathbb{R}$, and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. Here Y describes the outcome

of an experiment with our technical system, and our aim is to predict the density g of Y (w.r.t. the Lebesgue measure), which we assume to exist. The random vector X and the measurable function m describe our stochastic model of the technical system, and in this model we use $m(X)$ as an approximation of Y. Given the data

$$\begin{aligned}&(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, m(X_{n+1})), \ldots, (X_{n+L_n}, m(X_{n+L_n})),\\ &X_{n+L_n+1}, \ldots, X_{n+L_n+N_n}\end{aligned} \tag{1}$$

(where $L_n, N_n \in \mathbb{N}$) our goal is to construct an estimate for g.

The simplest way of doing this is to ignore X and m and to use only the data

$$Y_1, \ldots, Y_n \tag{2}$$

to define a kernel density estimate

$$\hat{g}_{Y,n}(y) = \frac{1}{n \cdot h_n} \cdot \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right). \tag{3}$$

Here $K : \mathbb{R}^d \to \mathbb{R}$ (so-called kernel, which is assumed to be a density) and $h_n > 0$ (so-called bandwidth) are parameters of the estimate.

In the sequel we assume that for $m^* : \mathbb{R}^d \to \mathbb{R}$ defined by $m^*(x) = \mathbf{E}\{Y|X = x\}$ the expected squared error occurring in approximating Y by $m^*(X)$,

$$\mathbf{E}\left\{|Y - m^*(X)|^2\right\},$$

is small. In this case an alternative way to estimate g is to use the data

$$(X_1, Y_1), \ldots, (X_n, Y_n), X_{n+L_n+1}, \ldots, X_{n+L_n+N_n} \tag{4}$$

in order to construct an estimate

$$m_{(X,Y),n}(\cdot) = m_{(X,Y),n}(\cdot, (X_1, Y_1), \ldots, (X_n, Y_n)) : \mathbb{R}^d \to \mathbb{R} \tag{5}$$

of m^* and to define the corresponding surrogate density estimate

$$\hat{g}_{(X,Y),n} = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K\left(\frac{y - m_{(X,Y),n}(X_{n+L_n+i})}{h_n}\right). \tag{6}$$

In this article we are interested in situations, where the sample size n is rather small (in our application in Section 4 we will have $n = 10$), since the collection of the real data (2) is rather expensive. Consequently, it might also be useful to use data from our model to estimate g. One possibility of doing this is to define an estimate of g on the basis of the data

$$(X_{n+1}, m(X_{n+1})), \ldots, (X_{n+L_n}, m(X_{n+L_n})), X_{n+L_n+1}, \ldots, X_{n+L_n+N_n} \tag{7}$$

by estimating in a first step a surrogate

$$m_{(X,m(X)),L_n}(\cdot) = \tag{8}$$
$$m_{(X,m(X)),L_n}(\cdot,(X_{n+1},m(X_{n+1})),\dots,(X_{n+L_n},m(X_{n+L_n}))) : \mathbb{R}^d \to \mathbb{R}$$

of m and by defining in the second step the corresponding surrogate density estimate via

$$\hat{g}_{(X,m(X)),L_n} = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K\left(\frac{y - m_{(X,m(X)),L_n}(X_{n+L_n+i})}{h_n}\right). \tag{9}$$

The main question which we want to investigate in this paper is whether there exist situations in which suitably defined estimates based on the complete data (1) achieve simultaneously better rate of convergence results than the estimates (3), (6) and (9).

In the next section we propose a novel method for improving the surrogate models (5) and (8) by using a combination of the real data (4) and the model data (7). Our main result is that the rate of convergence of the corresponding surrogate density estimate is at least as good as the rates of convergence of the density estimates (3), (6) and (9), and is in special situations better than any of the above rates of convergence. The finite sample size behaviour of our estimates is illustrated by using simulated data. The usefulness of our newly proposed estimates for uncertainty quantification is demonstrated by using it to analyze the uncertainty occurring in experiments with a suspension strut.

Throughout this paper we use the following notation: $\mathbb{N}$, $\mathbb{N}_0$ and $\mathbb{R}$ are the sets of positive integers, nonnegative integers and real numbers, respectively. Let $p = k + \beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$, and let $C > 0$. A function $m : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth, if for every $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = k$ the partial derivative $\frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ exists and satisfies

$$\left| \frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(x) - \frac{\partial^k m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^\beta$$

for all $x, z \in \mathbb{R}^d$. If X is a random variable, then $\mathbf{P}_X$ is the corresponding distribution, i.e., the measure associated with the random variable. Let $D \subseteq \mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be a real-valued function defined on $\mathbb{R}^d$. We write $x = \arg\max_{z \in D} f(z)$ if $\max_{z \in \mathcal{D}} f(z)$ exists and if x satisfies

$$x \in D \quad \text{and} \quad f(x) = \max_{z \in \mathcal{D}} f(z).$$

The outline of this paper is as follows: In Section 2 the construction of the improved surrogate model is explained. The main results are presented in Section 3 and proven in Section 5. The finite sample size performance of our estimates is illustrated in Section 4 by applying it to simulated and real data.

2 A new method for improving an imperfect surrogate model by real data

In this section we describe our ideas behind the construction of the improved surrogate model.

In order to construct density estimates on the basis of the data (1), we proceed as follows: We start by defining a surrogate estimate

$$m_{L_n}(\cdot) = m_{L_n}(\cdot, (X_{n+1}, m(X_{n+1})), \ldots, (X_{n+L_n}, m(X_{n+L_n}))) : \mathbb{R}^d \to \mathbb{R} \tag{10}$$

of m. In principle any kind of nonparametric regression estimate can be used at this point. In Section 4 we will use a penalized least squares estimate defined by

$$\tilde{m}_{L_n,(k,\lambda_{L_n})}(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\frac{1}{L_n} \sum_{i=n+1}^{n+L_n} |f(X_i) - m(X_i)|^2 + \lambda_{L_n} \cdot J_k^2(f) \right), \tag{11}$$

where $k \in \mathbb{N}$ with $2k > d$, where

$$J_k^2(f) = \sum_{\alpha_1,\ldots,\alpha_d \in \mathbb{N},\, \alpha_1+\cdots+\alpha_d=k} \frac{k!}{\alpha_1! \cdot \cdots \cdot \alpha_d!} \int_{\mathbb{R}^d} \left| \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) \right|^2 dx$$

is a penalty term penalizing the roughness of the estimate and where $W^k(\mathbb{R}^d)$ denotes the Sobolev space

$$\left\{ f \,:\, \frac{\partial^k f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} \in L_2(\mathbb{R}^d) \text{ for all } \alpha_1, \ldots, \alpha_d \in \mathbb{N} \text{ with } \alpha_1 + \cdots + \alpha_d = k \right\}.$$

The condition $2k > d$ implies that the functions in $W^k(\mathbb{R}^d)$ are continuous and hence the value of a function at a point is well defined. In order to be able to analyze the rate of convergence of this estimate for arbitrary distribution of X and dimension $d > 1$ we will truncate this estimate at some height $\beta > 0$, i.e., we will define

$$m_{L_n}(x) = T_\beta(\tilde{m}_{L_n,(k,\lambda_{L_n})}(x)) \quad (x \in \mathbb{R}^d), \tag{12}$$

where

$$T_\beta z = \begin{cases} \beta, & z > \beta \\ z, & -\beta \le z \le \beta \\ -\beta, & z < -\beta \end{cases}$$

for $z \in \mathbb{R}$.

Next we compute the residuals on the data set (4) of the estimate (10), i.e., we compute

$$\hat{\epsilon}_i = Y_i - m_{L_n}(X_i) \quad (i = 1, \ldots, n). \tag{13}$$

Then we define an estimate

$$\hat{m}_n^{\hat{\epsilon}}(\cdot) : \mathbb{R}^d \to \mathbb{R} \tag{14}$$

which smoothes these residuals (see below) and define our final surrogate model $(X, \hat{m}_n(X))$ (where $\hat{m}_n$ is an estimate of $m^*(\cdot) = \mathbf{E}\{Y|X = \cdot\}$) by setting

$$\hat{m}_n(x) = m_{L_n}(x) + \hat{m}_n^{\hat{\epsilon}}(x) \quad (x \in \mathbb{R}^d). \tag{15}$$

In order to define the estimate (14) we use two kinds of data sets: A first data set corresponding to the residuals of m_{L_n} on $X_1, \ldots, X_n$, i.e., the data set

$$\{(X_1, \hat{\epsilon}_1), \ldots, (X_n, \hat{\epsilon}_n)\} = \{(X_1, Y_1 - m_{L_n}(X_1))), \ldots, (X_n, Y_n - m_{L_n}(X_n))\}. \quad (16)$$

And a second data set corresponding to the residuals of m_{L_n} on the artificial sample with measurement errors

$$\{(X_{n+L_n+1}, m_{L_n}(X_{n+L_n+1})), \ldots, (X_{n+L_n+N_n}, m_{L_n}(X_{n+L_n+N_n}))\} \quad (17)$$

of (X, Y), i.e., the data set

$$\{(X_{n+L_n+1}, 0), \ldots, (X_{n+L_n+N_n}, 0)\}. \quad (18)$$

The second data set will be useful in particular in the case that m_{L_n} is already very close to

$$m^* : \mathbb{R}^d \to \mathbb{R}, m^*(x) = \mathbf{E}\{Y|X = x\},$$

since the sample size n of the data set (16) might be too small in order to detect that $0 \approx m^* - m_{L_n}$ might be the optimal choice for $\hat{m}_n^{\hat{\epsilon}}$.

Since both data sets are not equally trustworthy, we weight them by some weight $w^{(n)} \in [0, 1]$, and set

$$\begin{aligned}\tilde{m}_n^{\hat{\epsilon}}(\cdot) &= \tilde{m}_n^{\hat{\epsilon}}(\cdot, (X_{n+1}, m(X_{n+1})), \ldots, (X_{n+L_n}, m(X_{n+L_n})), X_{n+L_n+1}, \ldots, X_{n+L_n+N_n}) \\ &= \arg\min_{f \in W^k(\mathbb{R}^d)} \left(\frac{w^{(n)}}{n} \sum_{i=1}^{n} (\hat{\epsilon}_i - f(X_i))^2 + \frac{1 - w^{(n)}}{N_n} \sum_{i=1}^{N_n} (0 - f(X_{n+L_n+i}))^2 \right. \\ &\qquad \left. + \lambda_n \cdot J_k^2(f) \right) \end{aligned} \quad (19)$$

and

$$\hat{m}_n^{\hat{\epsilon}}(x) = T_{c_1 \cdot \alpha_n}(\tilde{m}_n^{\hat{\epsilon}}(x)) \quad (x \in \mathbb{R}^d), \quad (20)$$

where $c \geq 1$ and $\alpha_n > 0$. Finally, we use $(X, \hat{m}_n(X))$ as a surrogate model for (X, Y) and estimate the density g of Y by applying a kernel density estimate to a sample of $\hat{m}_n(X)$. To do this, we choose a kernel $K : \mathbb{R}^d \to \mathbb{R}$ and a bandwidth $h_{N_n} > 0$ and define

$$\hat{g}_{N_n}(y) = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K\left(\frac{y - \hat{m}_n(X_{n+L_n+i})}{h_{N_n}} \right). \quad (21)$$

3 Main results

In the next theorem we present bounds on the rate of convergence of our surrogate estimate, which we will use to derive bounds on the rate of convergence of our density estimate. In principle, all of our error bounds are also valid for finite n. In order to simplify the presentation, we consider the case $n \to \infty$ and assume that the distribution of (X, Y) and also the stochastic model $(X, m(X))$ change for increasing n such that $Y - m^*(X)$ and the error $m(X) - m^*(X)$ get smaller for increasing n. In order to simplify the notation we write (X, Y) and m instead of $(X^{(n)}, Y^{(n)})$ and $m^{(n)}$, resp.

Theorem 1 *Let $d, k \in \mathbb{N}$ with $2k > d$. Let (X, Y), (X_1, Y_1), ... be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$–valued random variables such that supp(X) is bounded and $\mathbf{E}\{Y^2\} < \infty$. Assume that Y has a density g with respect to the Lebesgue measure. Let $m^*(x) = \mathbf{E}\{Y|X = x\}$ and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. Let $n \in \mathbb{N}$ with $n \geq 2$ and let $L_n, N_n \in \mathbb{N}$ with $n \leq L_n \leq N_n$ and let $\alpha_n > \alpha_n^* \geq 0$. Assume that*

$$\mathbf{E}\left\{|Y - m^*(X)|^2\right\} \leq (\alpha_n^*)^2 \quad \text{and} \quad \mathbf{E}\left\{|Y - m^*(X)|^3\right\} \leq (\alpha_n^*)^3, \tag{22}$$

that there exists $K, \sigma_0 > 0$ such that

$$K^2 \cdot \left(\mathbf{E}\left\{\exp\left(\frac{(Y - m^*(X))^2}{\alpha_n \cdot K}\right)|X\right\} - 1\right) \leq \sigma_0 \quad a.s., \tag{23}$$

that the regression function $\mathbf{E}\{Y - m^*(X)|X = x\} = (m - m^*)(x)$ satisfies

$$\sup_{x \in \mathbb{R}^d} |m(x) - m^*(x)| \leq \alpha_n \tag{24}$$

and

$$J_k^2(m - m^*) \leq (\alpha_n)^2. \tag{25}$$

Furthermore, assume that

$$\alpha_n \geq \frac{1}{n^l} \quad \text{for some } l \in \mathbb{N} \tag{26}$$

and that for some $c_2 \in \mathbb{R}_+$ and $1 \leq \beta \leq n + L_n$ we have

$$|m(x)| \leq \beta \quad (x \in \mathbb{R}^d) \quad \text{and} \quad J_k^2(m) \leq c_2 < \infty. \tag{27}$$

Define $m_{L_n,(k,\lambda)}$ by (11) and (12), where

$$\lambda_{L_n} = c_3 \cdot \left(\frac{\log L_n}{L_n}\right)^{\frac{2k}{2k+d}}.$$

Define $\hat{m}_n^{\hat{\epsilon}}$ by (19) and (20) for some N_n satisfying $N_n \leq c_4 \cdot n^l$ for some $l \in \mathbb{N}$, choose $\lambda_n > 0$ such that

$$\frac{\log n}{n} \leq \lambda_n \leq \left(\frac{\log n}{n}\right)^{\frac{2k}{d}},$$

let $w^{(n)} \in [0, 1]$ and define $\hat{m}_n$ by (15). Then there exists constants $c_5, \ldots, c_{10} \in \mathbb{R}_+$ such that

$$\begin{aligned}
&\mathbf{E}\left\{|Y - \hat{m}_n(X)|^2\right\} \\
&\leq c_5 \cdot (\alpha_n^*)^2 + c_6 \cdot \alpha_n^2 \cdot \lambda_n + c_7 \cdot w^{(n)} \cdot \alpha_n^2 \cdot \frac{\log n}{n \cdot \lambda_n^{d/k}} + c_8 \cdot \left(\frac{\log L_n}{L_n}\right)^{\frac{2k}{2k+d}} \\
&\quad + c_9 \cdot (1 - w^{(n)}) \cdot \alpha_n^2 \left(1 + \frac{\log N_n}{N_n \cdot \lambda_n^{d/k}}\right) + \frac{c_{10} \cdot \alpha_n^2}{\min\{n, N_n\}} + \frac{c_{10}}{L_n}.
\end{aligned}$$

In particular, in case $w^{(n)} = 1$ *and* $\lambda_n = c_{11} \cdot ((\log n)/n)^{2 \cdot k/(2 \cdot k+d)}$ *we get*

$$\mathbf{E}\left\{|Y - \hat{m}_n(X)|^2\right\} \leq c_{12} \cdot \max\left\{(\alpha_n^*)^2, \alpha_n^2 \cdot \left(\frac{\log n}{n}\right)^{\frac{2k}{2k+d}}, \left(\frac{\log L_n}{L_n}\right)^{\frac{2k}{2k+d}}\right\}.$$

Remark 1. In any application of the estimate in Theorem 1 we have to choose the parameters depending on the data. In Section 3 we will use k–fold cross validation applied to the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ in order to choose $w^{(n)}$ and λ_n, and we choose λ_{L_n} by generalized cross validation applied the data $(X_{n+i}, m(X_{n+i}))$ $(i = 1, \ldots, L_n)$.

Theorem 1 implies the following corollary concerning the L_1 error of the density estimate (21):

Corollary 1 *Assume that the density* g *of* Y *is* (r, C)*–smooth for some* $r \in (0, 1]$ *and that its support is compact. Let* $K : \mathbb{R} \to \mathbb{R}$ *be a symmetric and bounded density which decreases monotonically on* $\mathbb{R}_+$ *and define the estimate* $\hat{g}_{N_n}$ *as in Section 2, where* $\hat{m}_n$ *is defined as at the end of Theorem 1. Assume that the assumptions of Theorem 1 are satsified, and that, in addition,*

$$\max\left\{(\alpha_n^*)^2, \left(\frac{\log L_n}{L_n}\right)^{\frac{2k}{2k+d}}\right\} \leq \alpha_n^2 \cdot \left(\frac{\log n}{n}\right)^{\frac{2k}{2k+d}}$$

holds. Set

$$h_{N_n} = c_{13} \cdot \left(\alpha_n \cdot \left(\frac{\log n}{n}\right)^{\frac{k}{2k+d}}\right)^{\frac{1}{r+1}}$$

and assume

$$N_n \geq \frac{1}{\alpha_n^{(2r+1)/(r+1)}} \cdot \left(\frac{n}{\log n}\right)^{\frac{k}{2k+d} \cdot \frac{2r+1}{r}}.$$

Then we have for some $c_{14} \in \mathbb{R}_+$

$$\mathbf{E} \int_{\mathbb{R}} |g_{N_n}(y) - g(y)|\, dy \leq c_{14} \cdot \left(\alpha_n \cdot \left(\frac{\log n}{n}\right)^{\frac{k}{2k+d}}\right)^{\frac{r}{r+1}}$$

Proof. Lemma 1 in Bott, Felber and Kohler (2015) implies that for any $z_1, z_2 \in \mathbb{R}$ we have

$$\int \left| K\left(\frac{y - z_1}{h_n}\right) - K\left(\frac{y - z_2}{h_n}\right)\right| dy \leq 2 \cdot K(0) \cdot |z_1 - z_2|.$$

Consequently,

$$\hat{g}_{Y,N_n}(y) = \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} K\left(\frac{y - Y_{n+L_n+i}}{h_{N_n}}\right)$$

satisfies

$$\int |\hat{g}_{N_n}(y) - \hat{g}_{Y,N_n}(y)|\, dy \leq \frac{1}{N_n \cdot h_{N_n}} \cdot \sum_{i=1}^{N_n} 2 \cdot K(0) \cdot |m_n(X_{n+L_n+i}) - Y_{n+L_n+i}|.$$

From this and standard bounds on the L_1 error of kernel density estimates (cf., e.g., proof of Theorem 1 in Felber, Kohler and Krzyżak (2015a)) we conclude

$$\begin{aligned}
&\mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - g(y)|\, dy \\
&\leq \mathbf{E} \int_{\mathbb{R}} |\hat{g}_{N_n}(y) - \hat{g}_{Y,N_n}(y)|\, dy + \mathbf{E} \int_{\mathbb{R}} |\hat{g}_{Y,N_n}(y) - g(y)|\, dy \\
&\leq \frac{2 \cdot K(0)}{h_{N_n}} \cdot \mathbf{E}\{|m_n(X) - Y|\} + \frac{c_{15}}{\sqrt{N_n \cdot h_{N_n}}} + c_{16} \cdot h_{N_n}^r \\
&\leq \frac{2 \cdot K(0)}{h_{N_n}} \cdot \sqrt{\mathbf{E}\{|m_n(X) - Y|^2\}} + \frac{c_{15}}{\sqrt{N_n \cdot h_{N_n}}} + c_{16} \cdot h_{N_n}^r.
\end{aligned}$$

Application of Theorem 1 yields the assertion. □

Remark 2. It is well-known that the L_1 error of the standard kernel density applied to the data (2) achieves under the assumptions of Corollary 1 the (optimal) rate of convergence

$$n^{-r/(2r+1)}.$$

It follows from the proof of Corollary 1 (together with standard error bounds on the L_2 error of smoothing spline estimates, cf., e.g., Chapter 21 in Györfi et al. (2002)), that the L_1 error of the surrogate density estimate defined in (6) and (9) achieves under the assumptions of Corollary 1 the rates of convergence

$$\left(\alpha_n^* + \left(\frac{\log n}{n} \right)^{\frac{k}{2k+d}} \right)^{\frac{r}{r+1}} \quad \text{and} \quad (\alpha_n)^{\frac{r}{r+1}}, \text{resp.}$$

For α_n suitably small the bound on the rate of convergence in Corollary 1 converges faster to zero than any of the above rates of convergence, which proves, that there exists situations in which our estimate theoretically outperforms the estimates defined in (3), (6) and (9). In the next section we demonstrate with simulated data that this is also the case for finite sample sizes.

4 Application to simulated and real data

In this section we illustrate the finite sample size performance of our estimates by applying them to simulated and real data.

We start with an application to simulated data, where we illustrate how the size of the error of the model influences the performance of our estimate. To do this, we choose X

d-dimensional standard normally distributed and ϵ uniformly distributed on $[0,1]$ such that X and ϵ are independent, set

$$Y = m(X) + \sigma \cdot \epsilon$$

for some $m : \mathbb{R}^d \to \mathbb{R}$ defined below and $\sigma \in \{0.1, 0.5, 1\}$, and let (X_1, Y_1), (X_2, Y_2), ... be independent and identically distributed random variables. Our estimate gets

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

as data from the real technical system,

$$(X_{n+1}, m(X_{n+1})), \ldots, (X_{n+L_n}, m(X_{n+L_n}))$$

as data from the (imperfect) model (where σ controls the maximal error occurring in this model), and the additional X-values

$$X_{n+L_n+1}, \ldots, X_{n+L_n+N_n}.$$

If we compare this setting with Theorem 1 we see that in the notation of Theorem 1 we have $\mathbf{E}\{Y|X = x\} = m(x)$ and consequently $\alpha_n^* = 0$.

In all of our applications we choose $n \in \{10, 20, 40\}$ and $L_n = 500$. As surrogate estimate we use a thin plate spline as implemented in the routine $Tps()$ in the statistics software R, where we use 5–fold cross validation (applied to the data $\mathcal{D}_n$) to choose the degree of freedom *df* of the fitted spline from the set $\{4, 8, 26, \ldots, 256\}$. In the same way we also choose $w^{(n)}$ from the set $\{0, 0.1, \ldots, 1\}$, i.e., we choose simultaneously the degree of freedom *df* and the weight $w^{(n)}$ by 5–fold cross validation.

For our newly proposed density estimate we use a sample of size $N_n = 500,000$ of $\hat{m}_n(X)$ (where $\hat{m}_n$ is the estimate introduced in Section 2) and apply to this sample a kernel density estimate as implemented in the routine *density()* in the statistics package R.

The density of Y is the convolution of the density of $m(X)$ and a uniform density. We do not try to compute its exact form, instead we compute it approximately by applying a kernel density estimate (as implemented in the routine *density*() in R) to a sample of size $1,000,000$ of Y. In order to judge the quality of our density estimates the resulting density is treated in our simulations as if it is the real density.

We compare our estimate (*est. 4*) with three other density estimates. The first one (*est. 1*) is the standard kernel density estimate as implemented in R applied to the sample of size n of Y, cf., (3). The other two estimates are surrogate density estimate, where the kernel density estimate of R is applied to a sample of size $N_n = 500,000$ of the surrogate model. For *est. 2* the surrogate model is chosen by applying a thin plate spline (as implemented in R) to the sample of size n of (X, Y), cf., (6). And for *est. 3* the surrogate model is computed in the same way, but using this time the sample of size $L_n = 500$ of our model $(X, m(X))$, cf., (9).

We consider three different models. In the first model we choose $d = 2$ and

$$m(x_1, x_2) = 2 \cdot x_1 + x_2 + 2.$$

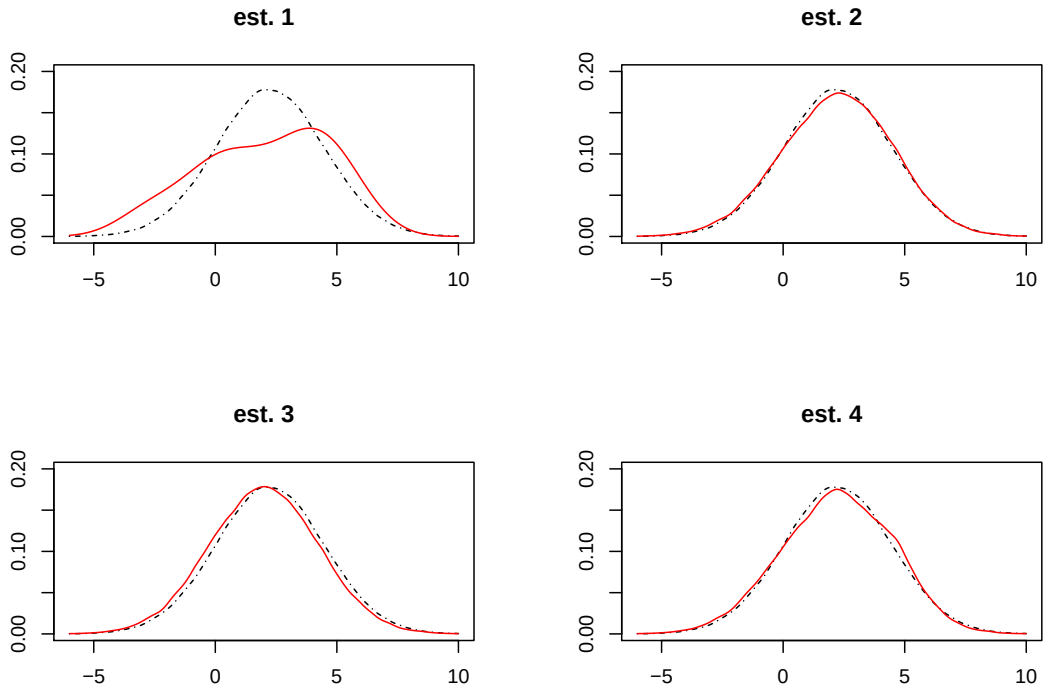

Figure 1: Four different density estimates together with the reference density (dashed-dotted) in simulation from model 1 with parameters $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

Figure 1 shows the plot of four different density estimates together with the reference density for a data set of model 1, where we use $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

In the second model we choose again $d = 2$, but define m this time by

$$m(x_1, x_2) = x_1^2 + x_2^2.$$

Figure 2 shows the plot of four different density estimates together with the reference density for a data set of model 2, where we use $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

In the third model we choose $d = 1$ and define m by

$$m(x) = \exp(x).$$

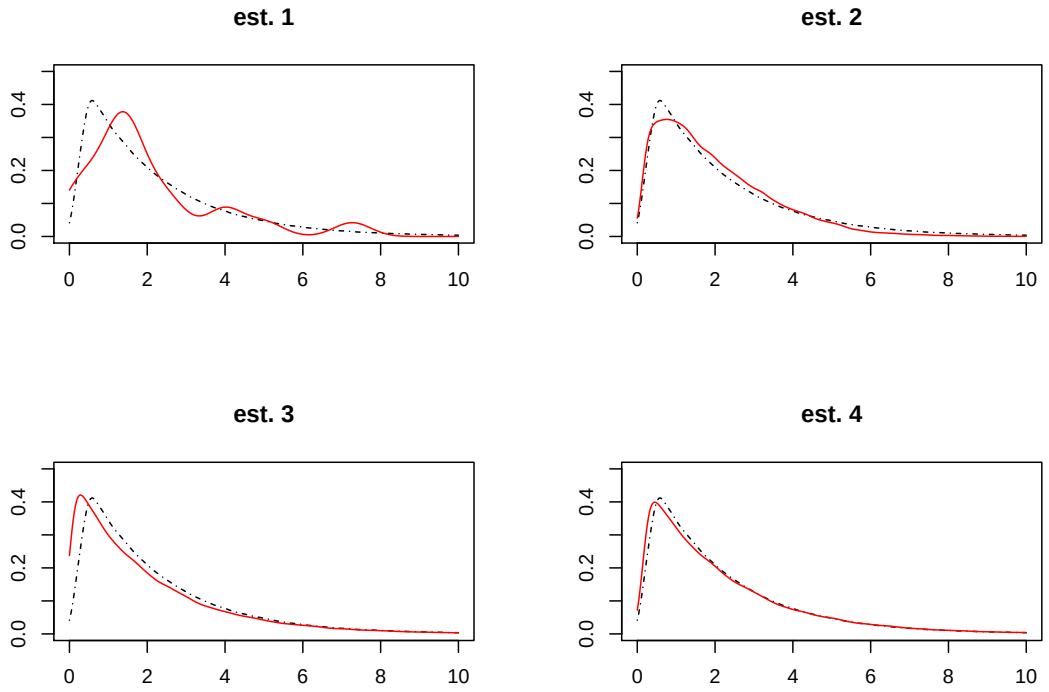

Figure 2: Four different density estimates together with the reference density (dashed-dotted) in simulation from model 1 with parameters $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

Figure 3 shows the plot of four different density estimates together with the reference density for a data set of model 3, where we use $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

We compare the L_1 errors of our four different estimates. To do this, we approximate the integral by a Riemann sum defined on an equidistant partition consisting of 8192 subintervals of the interval $[-6, 10]$ (in model 1) or the interval $[0, 10]$ (in models 2 and 3). Since this L_1 error is random, we repeat each simulation 100 times and report in Table 1 the median (and in brackets the interquartile range) of the 100 L_1 errors for each of our four estimates.

From Table 1 we see that our estimate outperforms all other estimates in 20 out of 27 settings, and in these cases often its error is by a factor 2 till 3 smaller than the errors of all other estimates. And in the seven cases where it does not achieves the smallest error, its error is approximately at the same size as the smallest error (and at most 20 percent larger). These larger error occur only in model 1, where the function m is a linear function which can be easily estimated even from a small sample of observation,

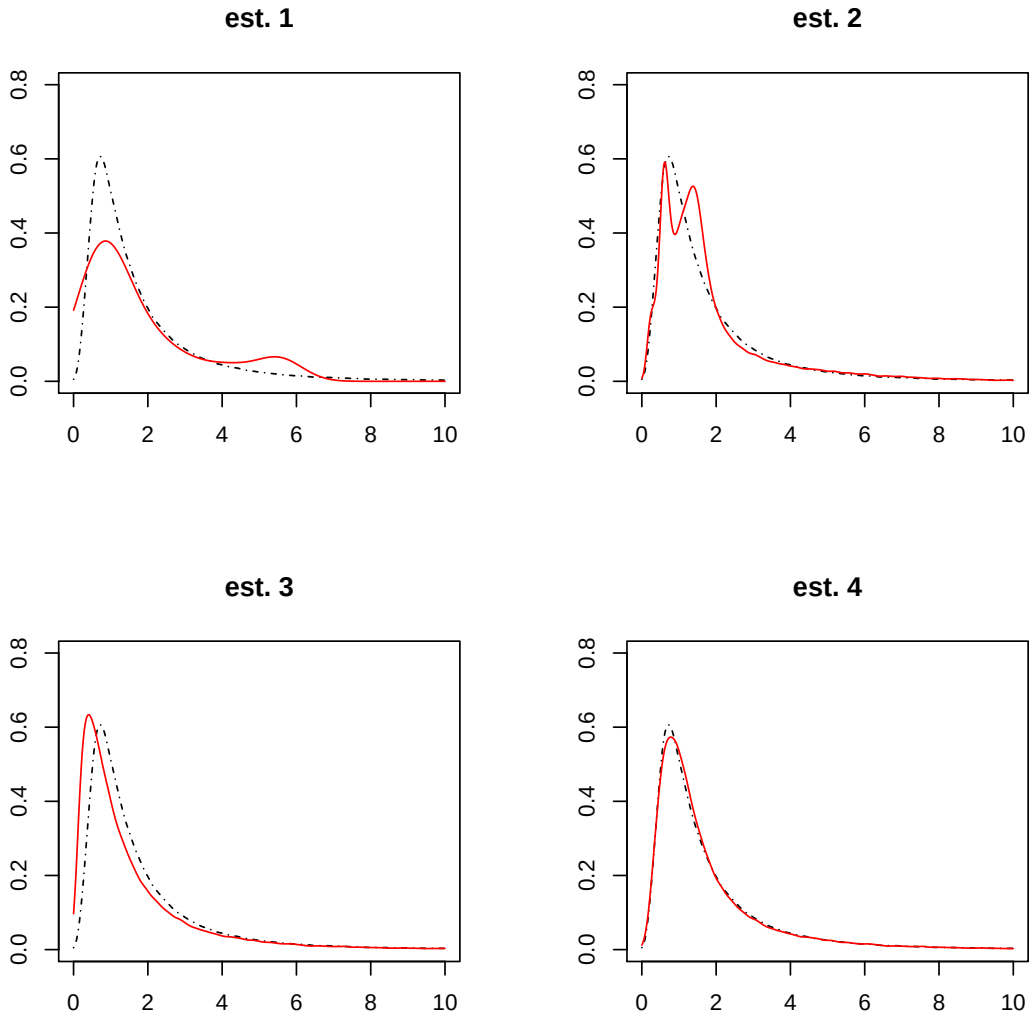

Figure 3: Four different density estimates together with the reference density (dashed-dotted) in simulation from model 3 with parameters $n = 20$, $\sigma = 0.5$, $L_n = 500$ and $N_n = 500,000$.

and where therefore the surrogate density estimate based on an estimated surrogate is rather good.

Next we illustrate the fact that the use of two samples with a data-dependent selected weight $w^{(n)}$ improves our estimate. To do this, we do additional simulations for $n = 10$ and $\sigma = 0.1$ in each of the three models, and compute also an estimate which uses a fixed weight $w^{(n)} = 1$ but is otherwise defined as our newly introduced est. 4. In Table

model	d	σ	n	est. 1	est. 2	est. 3	est. 4
1	2	0.1	10	0.324 (0.179)	0.010 (0.005)	0.019 (0.003)	**0.009** (0.004)
1	2	0.1	20	0.260 (0.151)	**0.008** (0.003)	0.019 (0.003)	0.009 (0.003)
1	2	0.1	40	0.223 (0.114)	**0.008** (0.003)	0.019 (0.003)	**0.008** (0.003)
1	2	0.5	10	0.389 (0.199)	**0.034** (0.028)	0.090 (0.003)	0.037 (0.029)
1	2	0.5	20	0.248 (0.139)	**0.022** (0.020)	0.090 (0.003)	0.025 (0.018)
1	2	0.5	40	0.233 (0.128)	**0.015** (0.010)	0.090 (0.003)	0.019 (0.012)
1	2	1	10	0.349 (0.221)	**0.066** (0.070)	0.178 (0.003)	**0.066** (0.048)
1	2	1	20	0.269 (0.132)	0.053 (0.050)	0.178 (0.003)	**0.051** (0.039)
1	2	1	40	0.213 (0.128)	**0.030** (0.026)	0.178 (0.003)	0.034 (0.020)
2	2	0.1	10	0.422 (0.189)	0.296 (0.150)	0.034 (0.002)	**0.015** (0.010)
2	2	0.1	20	0.344 (0.147)	0.135 (0.072)	0.034 (0.003)	**0.012** (0.008)
2	2	0.1	40	0.261 (0.102)	0.065 (0.027)	0.034 (0.002)	**0.011** (0.005)
2	2	0.5	10	0.444 (0.178)	0.317 (0.167)	0.188 (0.003)	**0.067** (0.046)
2	2	0.5	20	0.300 (0.115)	0.181 (0.078)	0.188 (0.003)	**0.053** (0.033)
2	2	0.5	40	0.246 (0.086)	0.111 (0.050)	0.188 (0.003)	**0.042** (0.018)
2	2	1	10	0.415 (0.201)	0.343 (0.164)	0.351 (0.003)	**0.145** (0.086)
2	2	1	20	0.304 (0.134)	0.220 (0.083)	0.351 (0.003)	**0.119** (0.049)
2	2	1	40	0.231 (0.072)	0.165 (0.054)	0.352 (0.003)	**0.108** (0.024)
3	1	0.1	10	0.483 (0.192)	0.141 (0.073)	0.064 (0.003)	**0.047** (0.050)
3	1	0.1	20	0.404 (0.140)	0.109 (0.078)	0.064 (0.003)	**0.030** (0.034)
3	1	0.1	40	0.293 (0.096)	0.070 (0.029)	0.064 (0.003)	**0.022** (0.021)
3	1	0.5	10	0.459 (0.170)	0.316 (0.182)	0.304 (0.003)	**0.168** (0.118)
3	1	0.5	20	0.389 (0.131)	0.256 (0.171)	0.304 (0.003)	**0.127** (0.132)
3	1	0.5	40	0.304 (0.086)	0.196 (0.090)	0.304 (0.003)	**0.105** (0.084)
3	1	1	10	0.430 (0.149)	0.401 (0.211)	0.528 (0.003)	**0.335** (0.213)
3	1	1	20	0.333 (0.119)	0.340 (0.178)	0.529 (0.003)	**0.255** (0.192)
3	1	1	40	0.278 (0.114)	0.316 (0.134)	0.528 (0.003)	**0.245** (0.159)

Table 1: Simulation results in the three different models.

2 we present the medians (and in brackets the interquartile ranges) of the error of this modified est. 4 together with the errors of the original est. 4 in 100 simulations for each of the above three settings.

From Table 2 we see that the adaptive choice of the weight slightly improves the error of our newly proposed estimate in two out of three cases, and that our newly proposed estimate achieves the same error as the estimate with the fixed weight in the remaining case.

Finally we illustrate the usefulness of our newly proposed method for uncertainty quantification by using it to analyze the uncertainty occurring in experiments with a suspension strut (cf., Figure 4), which serves as an academic demonstrator to study uncertainty in load distributions and the ability to control vibrations, stability and load

	model 1, $n = 10$, $\sigma = 0.1$	model 2, $n = 10$, $n = 10$, $\sigma = 0.1$	model 3, $n = 10$, $\sigma = 0.1$
est. 4 with $w^n = 1$	0.010 (0.006)	0.015 (0.011)	0.054 (0.078)
original est. 4	0.009 (0.004)	0.015 (0.010)	0.047 (0.050)

Table 2: Comparison of est. 4 with $w^{(n)} = 1$ fixed and the original est. 4 with data-dependent $w^{(n)}$ chosen from the set $\{0, 0.1, \ldots, 1\}$.

paths in suspension struts such as aircraft landing gears. A CAD illustration of this

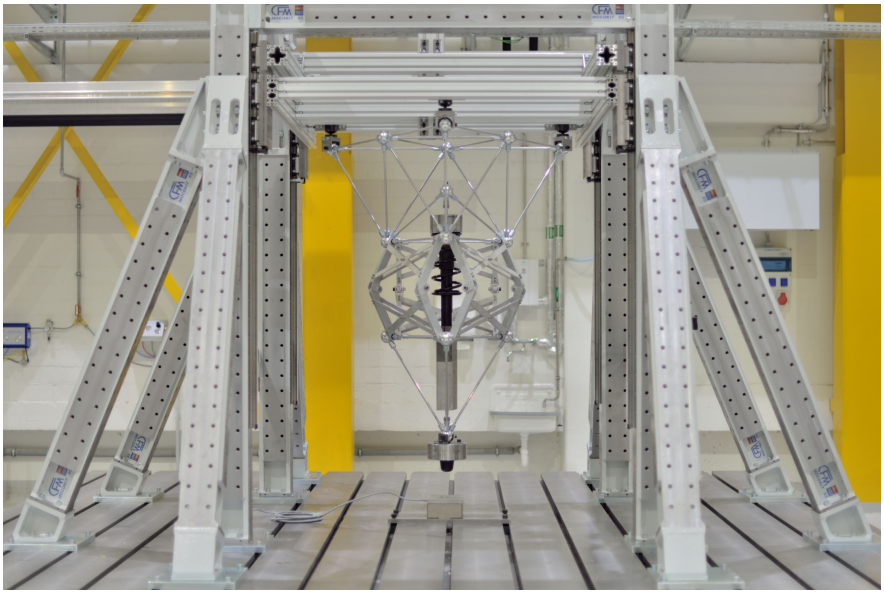

Figure 4: A photo of the demonstrator of a suspension strut and its experimental test setup.

suspension strut can be found in Figure 5 (left). This suspension strut consists of an upper and lower structure, where the lower structure contains a spring–damper component. The spring–damper component transmits the axial forces between the upper and lower structures of the suspension strut. The aim of our analysis is the analysis of the behaviour of the maximum relative compression of the spring damper component in case that the free fall height is chosen randomly. Here we assume that the free fall heights are independent normally distributed with mean 0.05 meter and standard deviation 0.0057 meter.

We analyze the uncertainty in the maximum relative compression in our suspension strut using a simplified mathematical model of the suspension strut (cf., Figure 5 (right)),

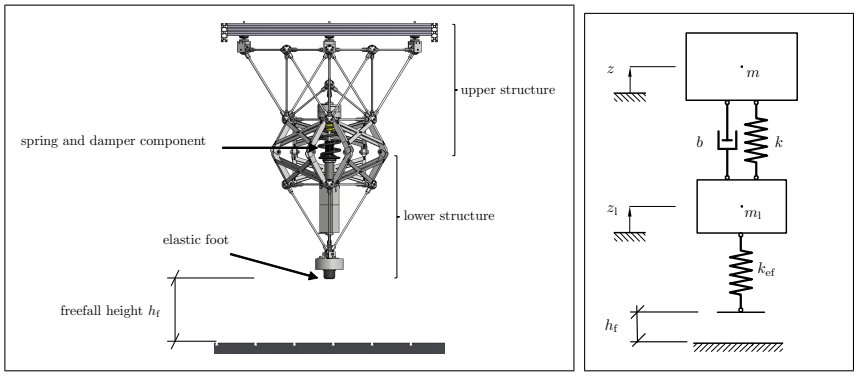

Figure 5: A CAD illustration of the suspension strut (left) and illustration of a simplified model of the suspension strut (right).

where the upper and the lower structures of the suspension strut are two lump masses m and m_1, the spring damper component is represented by a stiffness parameter k and a suitable damping coefficient b, and the elastic foot is represented by another stiffness parameter k_{ef}. Using a linear stiffness and an axiomatic damping it is possible to compute the maximum relative compression by solving a differential equation using Runge-Kutta algorithm (cf., model a) in Mallapur and Platz (2017)). We use the results of $L_n = 500$ corresponding computer experiments to construct a surrogate estimate m_{L_n} as described above.

In Figure 6 we see in the upper left panel data from $L_n = 500$ computer experiments together with a corresponding surrogate model (solid line), and in the upper right panel the corresponding surrogate density estimate. In the lower left panel we see again surrogate (dashed-dotted) based on the data from the computer experiments together with $n = 10$ real data points from the experiment. Clearly, our (dashed-dotted) surrogate model based only on the computer experiment is imperfect since it does not really fit the real data. By the methodology introduced in this paper we can improve this imperfect surrogate model, which yields the solid line in the lower left panel of Figure 6. The corresponding surrogate density estimate is shown in the lower right panel of Figure 6, and we see that the use of $n = 10$ additional data points leads in this example clearly to a different density estimate than the estimate based only on the model data in the upper right panel of Figure 6.

5 Proofs

5.1 Auxiliary results

In this subsection we present various auxiliary results on smoothing spline estimates, which we use in the next subsection in order to derive a new error bound on smoothing splines applied to weighted data with additional measurement errors in the dependent variable, cf., Theorem 2 below. This result will be used to proof Theorem 1.

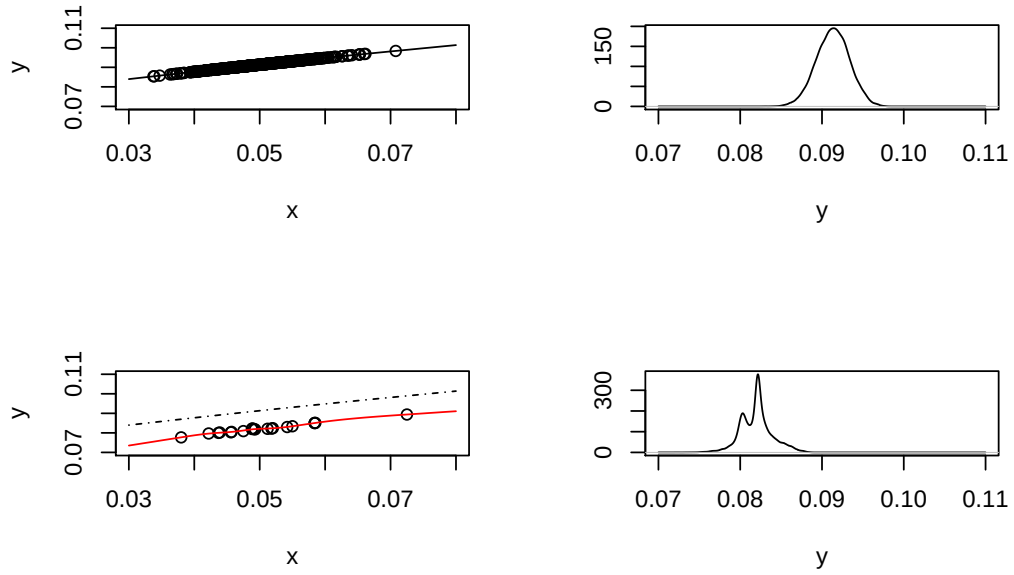

Figure 6: Data from $L_n = 500$ computer experiments together with a corresponding surrogate model (upper left panel), the corresponding density estimate (upper right model), the surrogate model based only on the data from the computer experiments (dashed-dotted) together with $n = 10$ experimental data points and the corresponding surrogate model proposed in this paper (solid line) (lower left panel) and the corresponding density estimate proposed in this paper (lower right panel).

5.1.1 A deterministic lemma

Lemma 1 *Let $d, N \in \mathbb{N}$, $t > 0$, $w_1, \ldots, w_N \in \mathbb{R}_+$, $x_1, \ldots, x_N \in \mathbb{R}^d$, $\beta_N \geq L > 0$, $z_1, \ldots, z_N \in \mathbb{R}$ and $\bar{z}_1, \ldots, \bar{z}_N \in [-L, L]$. Let $m : \mathbb{R}^d \to \mathbb{R}$ be a function. Let $\mathcal{F}_N$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$ and for $f \in \mathcal{F}_N$ let*

$$pen^2(f) \geq 0$$

be a penalty term. Define

$$\tilde{m}_N = \arg\min_{f \in \mathcal{F}_N} \left(\sum_{i=1}^{N} w_i \cdot |f(x_i) - \bar{z}_i|^2 + pen^2(f) \right)$$

(where we tacitly assume that the above minimum exists), and

$$\hat{m}_N(x) = T_{\beta_N}(\tilde{m}_N(x)) \quad (x \in \mathbb{R}^d)$$

and let $m_N^* \in \mathcal{F}_N$ *be arbitrary. Then*

$$\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2\left(\tilde{m}_N\right) \tag{28}$$
$$\geq 3\left(\sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + pen^2\left(m_N^*\right)\right) + 128\sum_{i=1}^{N} w_i \cdot |z_i - \bar{z}_i|^2 + t$$

implies

$$\sum_{i=1}^{N} w_i \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) \cdot (z_i - m(x_i))$$
$$\geq \frac{1}{24}\left(\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m_N^*(x_i)|^2 + pen^2\left(\tilde{m}_N\right)\right) + \frac{t}{6}. \tag{29}$$

Proof. Using

$$\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - \bar{z}_i|^2 \leq \sum_{i=1}^{N} w_i \cdot |\tilde{m}_N(x_i) - \bar{z}_i|^2$$

the assertion follows as in the proof of Lemma 5 in Furer and Kohler (2015). A complete proof is available from the authors upon request. □

5.1.2 A bound on a covering number

Definition 1 *Let* $l \in \mathbb{N}$ *and let* $\mathcal{F}$ *be a class of functions* $f : \mathbb{R}^l \to \mathbb{R}$. *The covering number* $\mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n)$ *is defined for any* $\epsilon > 0$ *and* $x_1^n = (x_1, ..., x_n) \in (\mathbb{R}^l)^n$ *as the smallest integer* k *such that there exist functions* $g_1, ..., g_k : \mathbb{R}^l \to \mathbb{R}$ *with*

$$\min_{1 \leq i \leq k} \left(\frac{1}{n}\sum_{j=1}^{n} |f(x_j) - g_i(x_j)|^2\right)^{1/2} \leq \epsilon$$

for each $f \in \mathcal{F}$.

Lemma 2 *Let* $L, A, c > 0$ *and set*

$$\mathcal{F} = \left\{T_L f \ : \ f \in W^k(\mathbb{R}^d) \text{ and } J_k^2(f) \leq c\right\}.$$

Then there exists constants $c_{17}, c_{18}, c_{19} \in \mathbb{R}_+$ *depending only on* A, k *and* d *such that for any* $\epsilon > 0$ *and all* $x_1, \ldots, x_n \in [-A, A]^d$

$$\log \mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n) \leq \left(c_{17} \cdot \left(\frac{\sqrt{c}}{\epsilon}\right)^{d/k} + c_{18}\right) \cdot \log\left(c_{19} \cdot \frac{L^2 n}{\epsilon^2}\right). \tag{30}$$

Proof. See Lemma 20.6 and Problem 20.9 in Györfi et al. (2002), or Lemma 3 in Kohler, Krzyżak and Schäfer (2002). □

5.1.3 A bound on the error for smoothing spline estimates for fixed design regression

Let $L \geq 0$ and

$$Y_i = m(x_i) + W_i \quad (i = 1, \ldots, n)$$

for some $x_1, \ldots, x_n \in \mathbb{R}^d$, $m : \mathbb{R}^d \to \mathbb{R}$ and some random variables $W_1, \ldots, W_n$ which are independent and have expectation zero. We assume that the W_i's are sub-Gaussian in the sense that

$$\max_{i=1,\ldots,n} K^2 \mathbf{E}\{e^{W_i^2/K^2} - 1\} \leq \sigma_0^2 \tag{31}$$

for some $K, \sigma_0 > 0$. Our goal is to estimate m from $(x_1, \bar{Y}_{1,n}), \ldots, (x_n, \bar{Y}_{n,n})$, where $\bar{Y}_{1,n}, \ldots, \bar{Y}_{n,n} \in [-L, L]$ are arbitrary (bounded) random variables with the property that the average squared measurement error

$$\frac{1}{n} \sum_{i=1}^{n} |Y_i - \bar{Y}_{i,n}|^2$$

is "small". Let $\mathcal{F}_n$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$ and consider the least squares estimate with complexity penalty

$$\tilde{m}_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - \bar{Y}_{i,n}|^2 + pen_n^2(f) \right) \quad \text{and} \quad m_n = T_{\beta_n} \tilde{m}_n, \tag{32}$$

where for $f \in \mathcal{F}_n$

$$pen_n^2(f) \geq 0$$

is a penalty term penalizing the complexity of f and where $\beta_n \geq L$. Set

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} |f(x_i)|^2.$$

Lemma 3 *Assume that the sub-Gaussian condition (31) holds and let the estimate be defined by (32). Then there exist constants $c_{20}, c_{21}, c_{22} > 0$ which depend only on σ_0 and K such that for any $\delta_n > c_{20}/n$ with*

$$\sqrt{n} \cdot \delta \geq c_{21} \int_{\delta/(12\sigma_0)}^{\sqrt{48\delta}} \left(\log \mathcal{N}_2 \left(u, \{T_{\beta_n} f - g : f \in \mathcal{F}_n, \right. \right. \tag{33}$$

$$\left. \left. \frac{1}{n} \sum_{i=1}^{n} |T_{\beta_n} f(x_i) - g(x_i)|^2 + pen_n^2(f) \leq 48 \cdot \delta \}, x_1^n \right) \right)^{1/2} du$$

for all $\delta \geq \delta_n/6$ and all $g \in \mathcal{F}_n$ we have for any $m_n^* \in \mathcal{F}_n$

$$\mathbf{P}\left\{ \|m_n - m\|_n^2 + pen_n^2(\tilde{m}_n) + 4 \cdot \delta_n \leq \frac{24}{n} \cdot \sum_{i=1}^{n} (m_n(x_i) - m_n^*(x_i)) \cdot W_i \right\}$$

$$\leq c_{22} \cdot \exp\left(-\frac{n \cdot \min\{\delta_n, \sigma_0^2\}}{c_{22}} \right).$$

Proof. The result follows from the proof of Lemma 2 in Kohler and Krzyżak (2012). A detailed proof is available from the authors upon request. □

5.1.4 A bound on the deviation between the L_2 error and the empirical L_2 error for smoothing splines

Let $(X, Y), (X_1, Y_1), \ldots$ be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$ valued random variables with $\mathbf{E}Y^2 < \infty$. Let $m(x) = \mathbf{E}\{Y|X = x\}$ be the corresponding regression function. Let $\bar{Y}_{1,n}, \ldots, \bar{Y}_{n,n}$ be $\mathbb{R}$–valued random variables and define the estimate m_n by

$$\tilde{m}_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - \bar{Y}_{i,n}|^2 + pen_n^2(f) \right),$$

where $\mathcal{F}_n$ is a set of functions $f : \mathbb{R}^d \to \mathbb{R}$ and for $f \in \mathcal{F}_n$

$$pen_n^2(f) \geq 0$$

is a penalty term penalizing the complexity of f. Set

$$m_n = T_{\beta_n} \tilde{m}_n$$

for some $\beta_n > 0$. Then the following result holds.

Lemma 4 *Let $\beta_n \geq L \geq 1$ and assume that the m is bounded in absolute value by L. Let $\mathcal{F}_n$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$ and define the estimate m_n as above. Then there exist constants $c_{23}, c_{24}, c_{25}, c_{26} > 0$ such that for any $\delta_n > 0$ which satisfies*

$$\delta_n > c_{23} \cdot \frac{\beta_n^2}{n}$$

and

$$c_{24} \frac{\sqrt{n}\delta}{\beta_n^2} \geq \int_{c_{25}\delta/\beta_n^2}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(u, \{(T_{\beta_n} f - m)^2 \, : \, f \in \mathcal{F}_n, \right. \right.$$
$$\left. \left. \frac{1}{n} \sum_{i=1}^{n} |T_{\beta_n} f(x_i) - m(x_i)|^2 \leq \frac{\delta}{\beta_n^2}, pen_n^2(f) \leq \delta \}, x_1^n \right) \right)^{1/2} du$$

for all $\delta \geq \delta_n$ and all $x_1, \ldots, x_n \in \mathbb{R}^d$, we have for $n \in \mathbb{N}$

$$\mathbf{P} \left\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) > \delta_n + 3 \cdot pen_n^2(\tilde{m}_n) + 3 \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - m(X_i)|^2 \right\}$$
$$\leq c_{26} \cdot \exp\left(-\frac{n \cdot \delta_n}{c_{26} \beta_n^2} \right).$$

Proof. The result follows from the bound on $P_{1,n}$ presented in the proof of Lemma 3 in Kohler and Krzyżak (2012). A detailed proof is available from the authors on request. □

5.2 A general result on penalized least squares estimates

Theorem 2 *Let $d, k, n, L_n \in \mathbb{N}$, $w^{(n)} \in [0,1]$ with $n \geq 2$ and $1 \leq \beta \leq n + L_n$. Let (X, Y), (X_1, Y_1), ... be independent and identically distributed $\mathbb{R}^d \times \mathbb{R}$–valued random variables with $\mathbf{E}\{Y^2\} < \infty$ and with $supp(X)$ bounded. Set $m(x) = \mathbf{E}\{Y|X = x\}$. Let $\bar{Y}_{1,n}, \ldots, \bar{Y}_{n+L_n,n}$ be arbitrary $\mathbb{R}$–valued random variables satisfying*

$$\max_{i=1,\ldots,n+L_n} \mathbf{E}\left\{|\bar{Y}_{i,n}|^3\right\} \leq c_{27} < \infty. \tag{34}$$

Set

$$w_i = \frac{w^{(n)}}{n} \quad \text{for } i = 1, \ldots, n$$

and

$$w_i = \frac{1 - w^{(n)}}{L_n} \quad \text{for } i = n+1, \ldots, n+L_n.$$

Assume $2 \cdot k > d$ and define the estimate m_n by

$$\tilde{m}_n(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - \bar{Y}_{i,n}|^2 + \lambda_n \cdot J_k^2(f) \right)$$

and

$$m_n(x) = T_\beta \tilde{m}_n(x) \quad (x \in \mathbb{R}^d).$$

Assume

$$K^2 \cdot \left(\mathbf{E}\left\{ \exp\left(\frac{(Y - m(X))^2}{K^2} \right) \Big| X \right\} - 1 \right) \leq \sigma_0^2 \quad a.s. \tag{35}$$

for some $K, \sigma_0 > 0$,

$$|m(x)| \leq \beta \quad (x \in \mathbb{R}^d) \tag{36}$$

and

$$J_k^2(m) < \infty. \tag{37}$$

Choose $\lambda_n \in \mathbb{R}_+$ such that

$$\frac{\log n}{n} \leq \lambda_n \leq \left(\frac{1}{\log L_n} \right)^{\frac{2k}{d}}. \tag{38}$$

Assume furthermore

$$n \leq L_n \leq n^l \tag{39}$$

for some $l \in \mathbb{N}$. Then there exists constants $c_{28}, c_{29}, c_{30}, c_{31} \in \mathbb{R}_+$ such that

$$\begin{aligned}
&\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\leq c_{28} \cdot \lambda_n \cdot J_k^2(m) + c_{29} \cdot w^{(n)} \cdot \left(\frac{\log n}{n \cdot \lambda_n^{d/2k}} + \mathbf{E}\left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} |\bar{Y}_{i,n} - Y_i|^2 \right\} \right) \\
&\quad + c_{30} \cdot (1 - w^{(n)}) \cdot \left(\frac{\log L_n}{L_n \cdot \lambda_n^{d/2k}} + \mathbf{E}\left\{ \frac{1}{L_n} \cdot \sum_{i=n+1}^{n+L_n} |\bar{Y}_{i,n} - Y_i|^2 \right\} \right) + \frac{c_{31}}{n}.
\end{aligned}$$

Proof. Set $\beta_n = n + L_n$.

In the first step of the proof we show that we can assume w.l.o.g.

$$\bar{Y}_{i,n} \in [-\beta_n, \beta_n] \quad \text{for all} \quad i = 1, \ldots, n + L_n. \tag{40}$$

To do this, we let

$$A_n = \left\{ |\bar{Y}_{i,n}| \leq \beta_n \quad \text{for all } i = 1, \ldots, n + L_n \right\}$$

be the event that all $\bar{Y}_{i,n}$ be bounded in absolutely value by β_n. The union bound together with Markov inequality implies

$$\begin{aligned} \mathbf{P}(A_n^c) &\leq (n + L_n) \cdot \max_{i=1,\ldots,n+L_n} \mathbf{P}\{|\bar{Y}_{i,n}| > \beta_n\} \leq (n + L_n) \cdot \frac{\max_{i=1,\ldots,n+L_n} \mathbf{E}\left\{|\bar{Y}_{i,n}|^3\right\}}{\beta_n^3} \\ &\leq \frac{c_{27}}{n}. \end{aligned}$$

On the event A_n the estimate m_n coincides with the estimate $m_n^{(trunc)}$ defined by

$$\tilde{m}_n^{(trunc)}(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\sum_{i=1}^{n+L_n} w_i \cdot |f(X_i) - T_{\beta_n} \bar{Y}_{i,n}|^2 + \lambda_n \cdot J_k^2(f) \right)$$

and

$$m_n^{(trunc)}(x) = T_\beta \tilde{m}_n^{(trunc)}(x) \quad (x \in \mathbb{R}^d).$$

From this we can conclude that

$$\begin{aligned} &\mathbf{E} \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\ &\leq \mathbf{E} \left\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot I_{A_n} \right\} + 4 \cdot \beta^2 \cdot \mathbf{P}(A_n^c) \\ &= \mathbf{E} \left\{ \int |m_n^{(trunc)}(x) - m(x)|^2 \mathbf{P}_X(dx) \cdot I_{A_n} \right\} + 4 \cdot \beta^2 \cdot \mathbf{P}(A_n^c) \\ &\leq \mathbf{E} \int |m_n^{(trunc)}(x) - m(x)|^2 \mathbf{P}_X(dx) + 4 \cdot \beta^2 \cdot \frac{c_{27}}{n}, \end{aligned}$$

which completes the first step of the proof.

So from now on we assume that (40) holds. Set

$$\delta_n = c_{32} \cdot \frac{\log n}{n \cdot \lambda_n^{d/(2k)}}, \quad \delta_{L_n} = c_{32} \cdot \frac{\log L_n}{L_n \cdot \lambda_n^{d/(2k)}}, \quad \gamma_n = w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}$$

and

$$T_n = \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) - \left(9 \cdot \lambda_n \cdot J_k^2(m) + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right).$$

In the second step of the proof we show that the assertion follows from

$$\int_{36\cdot\gamma_n}^{\infty} \mathbf{P}\{T_n > t\}\, dt \le \frac{c_{33}}{n}.$$

To do this, we observe

$$\begin{aligned}
&\mathbf{E}\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \\
&\le \mathbf{E}\left\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) - \left(9\cdot\lambda_n \cdot J_k^2(m) + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right) \right\} \\
&\qquad +9\cdot\lambda_n \cdot J_k^2(m) + 384\cdot \mathbf{E}\left\{ \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_{i,n}|^2 \right\} \\
&\le 36\cdot\gamma_n + \int_{36\cdot\gamma_n}^{\infty} \mathbf{P}\{T_n > t\}\, dt + 9\cdot\lambda_n\cdot J_k^2(m) + 384\cdot\mathbf{E}\left\{ \sum_{i=1}^{n+L_n} w_i\cdot|Y_i - \bar{Y}_{i,n}|^2 \right) \right\}.
\end{aligned}$$

The definition of γ_n and of the weights implies the assertion of step 2.

In the third step of the proof we show that we have for $t > 0$

$$\mathbf{P}\{T_n > t\} \le P_{1,n}(t) + P_{2,n}(t),$$

where

$$\begin{aligned}
P_{1,n}(t) &= \mathbf{P}\left\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \right. \\
&\qquad \left. > \frac{t}{2} + 3\cdot\lambda_n\cdot J_k^2(\tilde{m}_n) + 3\cdot \sum_{i=1}^{n+L_n} w_i \cdot |m_n(X_i) - m(X_i)|^2 \right\}
\end{aligned}$$

and

$$\begin{aligned}
P_{2,n}(t) &= \mathbf{P}\left\{ 3\cdot \sum_{i=1}^{n+L_n} w_i\cdot|m_n(X_i) - m(X_i)|^2 + 3\cdot\lambda_n\cdot J_k^2(\tilde{m}_n) \right. \\
&\qquad > \frac{t}{2} + 9\cdot\left(\sum_{i=1}^{n+L_n} w_i\cdot|m(X_i) - m(X_i)|^2 + \lambda_n J_k^2(m) \right) \\
&\qquad\qquad \left. +384\cdot\sum_{i=1}^{n+L_n} w_i\cdot|Y_i - \bar{Y}_{i,n}|^2 \right\}.
\end{aligned}$$

Using

$$T_n = \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) - 3\cdot\lambda_n\cdot J_k^2(\tilde{m}_n) - 3\cdot\sum_{i=1}^{n+L_n} w_i\cdot|m_n(X_i) - m(X_i)|^2$$

$$
\begin{aligned}
&+3\cdot\sum_{i=1}^{n+L_n} w_i\cdot|m_n(X_i)-m(X_i)|^2+3\cdot\lambda_n\cdot J_k^2(\tilde m_n)\\
&-\left(9\cdot\left(\sum_{i=1}^{n+L_n} w_i\cdot|m(X_i)-m(X_i)|^2+\lambda_n J_k^2(m)\right)+384\cdot\sum_{i=1}^{n+L_n} w_i\cdot|Y_i-\bar Y_{i,n}|^2\right)\\
=:\;& T_{1,n}+T_{2,n}
\end{aligned}
$$

this immediately follows from

$$\mathbf{P}\{T_n>t\}=\mathbf{P}\{T_{1,n}+T_{2,n}>t\}\le\mathbf{P}\{T_{1,n}>t/2\}+\mathbf{P}\{T_{2,n}>t/2\}.$$

In the fourth step of the proof we derive a upper bound on

$$\int_{36\cdot\gamma_n}^{\infty} P_{1,n}(t)\,dt.$$

Let $t\ge 36\cdot\gamma_n$. The definition of the weights together with

$$a+b>c+d\quad\Rightarrow\quad(a>c\text{ or }b>d)$$

implies that we have

$$
\begin{aligned}
P_{1,n}(t)\;\le\;&\mathbf{P}\Bigg\{w^{(n)}\cdot\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)>\frac{w^{(n)}\cdot\delta_n}{w^{(n)}\cdot\delta_n+(1-w^{(n)})\cdot\delta_{L_n}}\cdot\frac{t}{2}\\
&\qquad+w^{(n)}\cdot3\cdot\lambda_n\cdot J_k^2(\tilde m_n)+w^{(n)}\cdot3\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}|m_n(X_i)-m(X_i)|^2\Bigg\}\\
&+\mathbf{P}\Bigg\{(1-w^{(n)})\cdot\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)\\
&\qquad>\frac{(1-w^{(n)})\cdot\delta_{L_n}}{w^{(n)}\cdot\delta_n+(1-w^{(n)})\cdot\delta_{L_n}}\cdot\frac{t}{2}+(1-w^{(n)})\cdot3\cdot\lambda_n\cdot J_k^2(\tilde m_n)\\
&\qquad+(1-w^{(n)})\cdot3\cdot\frac{1}{L_n}\cdot\sum_{i=n+1}^{n+L_n}|m_n(X_i)-m(X_i)|^2\Bigg\}\\
\le\;&\mathbf{P}\Bigg\{\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)>\frac{\delta_n}{w^{(n)}\cdot\delta_n+(1-w^{(n)})\cdot\delta_{L_n}}\cdot\frac{t}{2}\\
&\qquad+3\cdot\lambda_n\cdot J_k^2(\tilde m_n)+3\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}|m_n(X_i)-m(X_i)|^2\Bigg\}\\
&+\mathbf{P}\Bigg\{\int|m_n(x)-m(x)|^2\mathbf{P}_X(dx)>\frac{\delta_{L_n}}{w^{(n)}\cdot\delta_n+(1-w^{(n)})\cdot\delta_{L_n}}\cdot\frac{t}{2}\\
&\qquad+3\cdot\lambda_n\cdot J_k^2(\tilde m_n)+3\cdot\frac{1}{L_n}\cdot\sum_{i=n+1}^{n+L_n}|m_n(X_i)-m(X_i)|^2\Bigg\}.
\end{aligned}
$$

We show next that Lemma 4 is applicable to the two different probabilities, where both times β_n is replaced by β and where we use sample sizes n and L_n, resp. Since $t \geq 36 \cdot \gamma_n \geq 2 \cdot \gamma_n$ implies

$$\frac{\delta_n}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2} \geq \delta_n$$

and

$$\frac{\delta_{L_n}}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2} \geq \delta_{L_n},$$

in order to show that Lemma 4 is applicable to the first probability, it suffices to show

$$\delta_n > c_{34} \cdot \frac{\beta^2}{n}$$

and

$$\begin{aligned} & c_{35} \frac{\sqrt{n}\delta}{\beta^2} \\ & \geq \int_{c_{36}\delta/\beta^2}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(u, \{(T_\beta f - m)^2 \,:\, f \in W_k(\mathbb{R}^d), J_k^2(f) \leq \frac{\delta}{\lambda_n}\}, x_1^n \right) \right)^{1/2} du \end{aligned}$$

for all $\delta \geq \delta_n$ and all $x_1, \ldots, x_n \in \mathbb{R}^d$. Using $|a^2 - b^2|^2 \leq (|a| + |b|)^2 \cdot |a - b|^2$ $(a, b \in \mathbb{R})$ (which we apply with $a = T_\beta f(x_i) - m(x_i)$ and $b = g(x_i)$, where g is approximating $T_\beta f - m$), we see that we have

$$\begin{aligned} & \mathcal{N}_2 \left(u, \left\{ (T_\beta f - m)^2 \,:\, f \in W_k(\mathbb{R}^d), J_k^2(f) \leq \frac{\delta}{\lambda_n} \right\}, x_1^n \right) \\ & \leq \mathcal{N}_2 \left(\frac{u}{16\beta^2}, \left\{ T_{\beta_n} f - m \,:\, f \in W_k(\mathbb{R}^d), J_k^2(f) \leq \frac{\delta}{\lambda_n} \right\}, x_1^n \right). \end{aligned}$$

Using this together with Lemma 2 we see that Lemma 4 is applicable to the first probability, if $\delta_n > c_{34} \cdot \frac{\beta^2}{n}$ and the following inequality hold:

$$\frac{\sqrt{n} \cdot \delta}{\beta^2} \geq c_{37} \cdot \int_0^{\sqrt{\delta}} \left(\left(\left(\frac{\sqrt{\delta/\lambda_n}}{\frac{u}{16\beta^2}} \right)^{d/k} + 1 \right) \cdot \log(c_{38} \cdot \beta^2 \cdot n^3) \right)^{1/2} du.$$

The last condition is implied by

$$\sqrt{n} \cdot \delta \geq c_{39} \cdot \sqrt{\log(c_{38} \cdot \beta^2 \cdot n^3)} \cdot \left(\left(\frac{\delta}{\lambda_n} \right)^{d/(4k)} \cdot \delta^{\frac{1}{2} - \frac{d}{4k}} + \sqrt{\delta} \right),$$

which in turn follows from

$$\delta \geq c_{40} \cdot \frac{\log n}{n \cdot \lambda_n^{d/(2k)}} \quad \text{and} \quad \delta \geq c_{40} \cdot \frac{\log n}{n}.$$

In case that

$$\lambda_n \leq 1$$

the last two conditions hold for all $\delta \geq \delta_n$, provided c_{32} is chosen large enough.

In the same way one can show that Lemma 4 is also applicable to the second probability above.

By applying Lemma 4 to the two different probabilities we get

$$\begin{aligned} P_{1,n}(t) \quad &\leq \quad c_{41} \cdot \exp\left(-c_{42} \cdot n \cdot \frac{\delta_n}{w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2}\right) \\ &\quad + c_{41} \cdot \exp\left(-c_{42} \cdot L_n \cdot \frac{\delta_{L_n}}{w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{2}\right), \end{aligned}$$

which implies

$$\begin{aligned} &\int_{36 \cdot \gamma_n}^{\infty} P_{1,n}(t)\, dt \\ &\leq \frac{c_{43}}{n} \cdot \frac{w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}}{\delta_n} \cdot \exp\left(-c_{44} \cdot n \cdot \delta_n\right) \\ &\quad + \frac{c_{43}}{L_n} \cdot \frac{w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}}{\delta_{L_n}} \cdot \exp\left(-c_{44} \cdot L_n \cdot \delta_{L_n}\right) \leq \frac{c_{45}}{n}. \end{aligned}$$

Here the last inequality follows from the assumptions (38) and (39), from which we can conclude

$$n \cdot \delta_n \geq c_{32} \cdot \log^2(n), \quad L_n \cdot \delta_{L_n} \geq c_{32} \cdot \log^2(n),$$

$$\frac{w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}}{n \cdot \delta_n} \leq n^s \quad \text{and} \quad \frac{w^{(n)} \cdot \delta_n + (1-w^{(n)}) \cdot \delta_{L_n}}{L_n \cdot \delta_{L_n}} \leq n^s$$

for some $s > 0$.

In the fifth step of the proof we derive a upper bound on

$$\int_{36 \cdot \gamma_n}^{\infty} P_{2,n}(t)\, dt.$$

Since $|m(x)| \leq \beta \leq \beta_n$ $(x \in \mathbb{R}^d)$ and $w_i \geq 0$ $(i \in \{1, \ldots, n\})$ we have

$$\sum_{i=1}^{n+L_n} w_i \cdot |m_n(X_i) - m(X_i)|^2 \leq \sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n} \tilde{m}_n(X_i) - m(X_i)|^2.$$

This together with (40) and Lemma 1 (applied with $m_n^* = m$) implies

$$\begin{aligned} &P_{2,n}(t) \\ &\leq \mathbf{P}\left\{3 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)|^2 + 3 \cdot \lambda_n \cdot J_k^2(\tilde{m}_n)\right. \end{aligned}$$

$$> \frac{t}{2} + 9 \cdot \left(\sum_{i=1}^{n+L_n} w_i \cdot |m(X_i) - m(X_i)|^2 + \lambda_n J_k^2(m) \right) + 384 \cdot \sum_{i=1}^{n+L_n} w_i \cdot |Y_i - \bar{Y}_i|^2 \Bigg\}$$

$$\leq \mathbf{P}\Big\{ \sum_{i=1}^{n+L_n} w_i \cdot (T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)) \cdot (Y_i - m(X_i)) \geq$$

$$\frac{1}{24} \left(\sum_{i=1}^{n+L_n} w_i \cdot |T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2\left(\bar{m}_n\right) \right) + \frac{t}{36} \Big\}.$$

Proceeding as in the proof of step 4 we can conclude from the definition of the weights that the last probability is bounded by

$$\mathbf{P}\Big\{ \frac{1}{n} \cdot \sum_{i=1}^{n} (T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)) \cdot (Y_i - m(X_i)) \geq$$

$$\frac{1}{n} \cdot \frac{1}{24} \left(\sum_{i=1}^{n} |T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2\left(\bar{m}_n\right) \right)$$

$$+ \frac{\delta_n}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{36} \Big\}$$

$$+\mathbf{P}\Big\{ \frac{1}{L_n} \cdot \sum_{i=n+1}^{n+L_n} (T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)) \cdot (Y_i - m(X_i)) \geq$$

$$\frac{1}{L_n} \cdot \frac{1}{24} \left(\sum_{i=n+1}^{n+L_n} |T_{\beta_n}(\tilde{m}_n)(X_i) - m(X_i)|^2 + \lambda_n \cdot J_k^2\left(\bar{m}_n\right) \right)$$

$$+ \frac{\delta_{L_n}}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{36} \Big\},$$

and that Lemma 3 can be applied to both probabilities. From this we can conclude that the above probabilities are bounded by

$$c_{46} \cdot \exp\left(-c_{47} \cdot n \cdot \frac{\delta_n}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{36} \right)$$

$$+ c_{46} \cdot \exp\left(-c_{47} \cdot L_n \cdot \frac{\delta_{L_n}}{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}} \cdot \frac{t}{36} \right),$$

which implies as above

$$\int_{36 \cdot \gamma_n}^{\infty} P_{2,n}(t)\, dt$$

$$\leq \frac{c_{48}}{n} \cdot \frac{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}}{\delta_n} \cdot \exp\left(-c_{49} \cdot n \cdot \delta_n \right)$$

$$+ \frac{c_{48}}{L_n} \cdot \frac{w^{(n)} \cdot \delta_n + (1 - w^{(n)}) \cdot \delta_{L_n}}{\delta_{L_n}} \cdot \exp\left(-c_{49} \cdot L_n \cdot \delta_{L_n} \right) \leq \frac{c_{50}}{n}.$$

Summarizing the above results we get the assertion. □

5.3 Proof of Theorem 1

Using the definition of $\hat{m}_n$, $(a+b+c)^2 \le 3a^2+3b^2+3c^2$ $(a,b,c \in \mathbb{R})$, (22), the definition of m_n and (27) we get

$$\begin{aligned}
&\mathbf{E}\left\{|Y-\hat{m}_n(X)|^2\right\} \\
&= \mathbf{E}\left\{\left|(Y-m^*(X))+(m^*(X)-m(X)-\hat{m}_n^{\hat{\epsilon}}(X))+(m(X)-m_{L_n}(X))\right|^2\right\} \\
&\le 3\cdot\mathbf{E}\left\{|Y-m^*(X)|^2\right\}+3\cdot\mathbf{E}\left\{\left|m^*(X)-m(X)-\hat{m}_n^{\hat{\epsilon}}(X)\right|^2\right\} \\
&\quad +3\cdot\mathbf{E}\left\{|m(X)-m_{L_n}(X)|^2\right\} \\
&\le 3(\alpha_n^*)^2+3\cdot\mathbf{E}\int\left|\hat{m}_n^{\hat{\epsilon}}(x)-(m^*-m)(x)\right|^2\mathbf{P}_X(dx)+3\cdot\mathbf{E}\int|m_{L_n}(x)-m(x)|^2\,\mathbf{P}_X(dx).
\end{aligned}$$

Hence in order to prove the assertion it suffices to show

$$\mathbf{E}\int|m_{L_n}(x)-m(x)|^2\,\mathbf{P}_X(dx) \le c_{51}\cdot\lambda_{L_n}\cdot J_k^2(m)+c_{52}\cdot\frac{\log L_n}{L_n\cdot\lambda_{L_n}^{d/(2k)}}+\frac{c_{53}}{L_n} \tag{41}$$

and

$$\begin{aligned}
&\mathbf{E}\int\left|\hat{m}_n^{\hat{\epsilon}}(x)-(m^*-m)(x)\right|^2\mathbf{P}_X(dx) \\
&\le c_{54}\cdot\alpha_n^2\cdot\lambda_n+c_{55}\cdot w^{(n)}\cdot\alpha_n^2\cdot\frac{\log n}{n\cdot\lambda_n^{d/(2k)}}+c_{56}\cdot\left(\frac{\log L_n}{L_n}\right)^{\frac{2k}{2k+d}} \\
&\quad +c_{57}\cdot(1-w^{(n)})\cdot\alpha_n^2\cdot\left(1+\frac{\log N_n}{N_n\cdot\lambda_n^{d/k}}\right)+\frac{c_{58}\cdot\alpha_n^2}{\min\{n,N_n\}}.
\end{aligned} \tag{42}$$

Inequality (41) follows from Theorem 2 applied with $(X,Y)=(X,m(X))$, $n=L_n$, $w^{(n)}=1$ and $\bar{Y}_{i,L_n+\bar{L}_n}=Y_i=m(X_{n+i})$ $(i=1,\dots,L_n)$ and suitably chosen $\bar{Y}_{L_n+1,L_n+\bar{L}_n}$, $\dots,\bar{Y}_{L_n+\bar{L}_n,L_n+\bar{L}_n}$.

In order to prove (42) we first observe that

$$\mathbf{E}\{Y-m(X)|X=x\}=m^*(x)-m(x),$$

hence m^*-m is the regression function to $(X,Y-m(X))$, and $(m^*-m)/\alpha_n$ is the regression function to $(X,(Y-m(X))/\alpha_n)$. Clearly,

$$\int\left|\hat{m}_n^{\hat{\epsilon}}(x)-(m^*-m)(x)\right|^2\mathbf{P}_X(dx)=\alpha_n^2\cdot\int\left|\frac{1}{\alpha_n}\cdot\hat{m}_n^{\hat{\epsilon}}(x)-\frac{1}{\alpha_n}\cdot(m^*-m)(x)\right|^2\mathbf{P}_X(dx).$$

By definition of $\hat{m}_n^{\hat{\epsilon}}$ we have

$$\frac{1}{\alpha_n}\cdot\hat{m}_n^{\hat{\epsilon}}(x)=\frac{1}{\alpha_n}\cdot T_{c_1\cdot\alpha_n}(\tilde{m}_n^{\hat{\epsilon}}(x))=T_{c_1}\left(\frac{1}{\alpha_n}\cdot\tilde{m}_n^{\hat{\epsilon}}(x)\right)\quad(x\in\mathbb{R}^d),$$

where

$$\frac{1}{\alpha_n} \cdot \tilde{m}_n^{\hat{\epsilon}}(\cdot) = \arg \min_{f \in W^k(\mathbb{R}^d)} \left(\frac{w^{(n)}}{n} \sum_{i=1}^{n} \left(\frac{1}{\alpha_n} \cdot \hat{\epsilon}_i - f(X_i) \right)^2 + \frac{1 - w^{(n)}}{N_n} \sum_{i=1}^{N_n} (0 - f(X_{n+L_n+i}))^2 + \lambda_n \cdot J_k^2(f) \right).$$

The assumptions in Theorem 1 together with (41) imply that we have

$$\sup_{x \in \mathbb{R}^d} \left| \frac{1}{\alpha_n} \cdot (m^* - m)(x) \right| \leq 1 \leq c_1$$

and

$$\begin{aligned}
&\max_{i=1,\ldots,n} \mathbf{E} \left\{ \left| \frac{Y_i - m_{L_n}(X_i)}{\alpha_n} \right|^3 \right\} \\
&\leq \frac{27}{\alpha_n^3} \cdot \left(\mathbf{E} \left\{ |Y - m^*(X)|^3 \right\} + \mathbf{E} \left\{ |m^*(X) - m(X)|^3 \right\} + \mathbf{E} \left\{ |m(X) - m_{L_n}(X)|^3 \right\} \right) \\
&\leq \frac{(\alpha_n^*)^3}{\alpha_n^3} + 1 + \frac{c_{59} \cdot \left(\frac{\log L_n}{L_n} \right)^{\frac{2k}{2k+d}}}{\alpha_n^3} \leq 2 + c_{59}
\end{aligned}$$

We consider

$$\frac{1}{\alpha_n} \cdot \hat{\epsilon}_i = \frac{1}{\alpha_n} \cdot (Y_i - m_{L_n}(X_i)) = \frac{1}{\alpha_n} \cdot (Y_i - m(X_i)) + \frac{1}{\alpha_n} \cdot (m(X_i) - m_{L_n}(X_i))$$

as an observation of $(Y_i - m(X_i))/\alpha_n$ with an additional measurement error

$$\frac{1}{\alpha_n} \cdot (m(X_i) - m_{L_n}(X_i))$$

$(i = 1, \ldots, n)$. And we consider

$$0 = \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i})) - \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i}))$$

as an observation of $\frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i}))$ with an additional measurement error

$$(-1) \cdot \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i}))$$

$(i = 1, \ldots, N_n)$.

From inequality (41) we can conclude

$$\mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \frac{1}{\alpha_n} \cdot (m(X_i) - m_{L_n}(X_i)) \right|^2 \right\}$$

$$\leq \frac{1}{\alpha_n^2} \cdot \left(c_{51} \cdot \lambda_{L_n} \cdot J_k^2(m) + c_{52} \cdot \frac{\log L_n}{L_n \cdot \lambda_{L_n}^{d/(2k)}} + \frac{c_{53}}{L_n} \right),$$

and the assumptions in Theorem 1 imply

$$\begin{aligned}
&\mathbf{E}\left\{ \frac{1}{N_n} \sum_{i=1}^{N_n} \left| \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m(X_{n+L_n+i})) \right|^2 \right\} \\
&\leq 2 \cdot \mathbf{E}\left\{ \frac{1}{N_n} \sum_{i=1}^{N_n} \left| \frac{1}{\alpha_n} \cdot (Y_{n+L_n+i} - m^*(X_{n+L_n+i})) \right|^2 \right\} \\
&\quad +2 \cdot \mathbf{E}\left\{ \frac{1}{N_n} \sum_{i=1}^{N_n} \left| \frac{1}{\alpha_n} \cdot (m^*(X_{n+L_n+i}) - m(X_{n+L_n+i})) \right|^2 \right\} \\
&\leq 2 \cdot \frac{(\alpha_n^*)^2}{\alpha_n^2} + 2 \leq 4.
\end{aligned}$$

Application of Theorem 2 yields

$$\begin{aligned}
&\mathbf{E} \int \left| \frac{1}{\alpha_n} \cdot \hat{m}_n^{\hat{\epsilon}}(x) - \frac{1}{\alpha_n} \cdot (m^* - m)(x) \right|^2 \mathbf{P}_X(dx) \\
&\leq c_{28} \cdot \lambda_n \cdot J_k^2 \left(\frac{1}{\alpha_n} \cdot (m^* - m) \right) \\
&\quad +c_{29} \cdot w^{(n)} \cdot \left(\frac{\log n}{n \cdot \lambda_n^{d/(2k)}} + \frac{1}{\alpha_n^2} \cdot \left(c_{51} \cdot \lambda_{L_n} \cdot J_k^2(m) + c_{52} \cdot \frac{\log L_n}{L_n \cdot \lambda_{L_n}^{d/(2k)}} + \frac{c_{53}}{L_n} \right) \right) \\
&\quad +c_{30} \cdot (1 - w^{(n)}) \cdot \left(\frac{\log N_n}{N_n \cdot \lambda_n^{d/(2k)}} + 4 \right) + \frac{c_{31}}{\min\{n, N_n\}},
\end{aligned}$$

which implies (42). □

References

[1] Bayarri, M. J., Berger, J. O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R. J., Paulo, R., Sacks, J., and Walsh, D. (2007). Computer model validation with functional output. *The Annals of Statistics*, **35**, pp. 1874–1906.

[2] Bichon, B., Eldred, M., Swiler, M., Mahadevan, S. and McFarland, J. (2008). Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA Journal*, **46**, pp. 2459–2468.

[3] Bott, A. K., Felber, T., and Kohler, M. (2015). Estimation of a density in a simulation model. *Journal of Nonparametric Statistics*, **27**, pp. 271-285.

[4] Bourinet, J.-M., Deheeger, F. and Lemaire, M. (2011). Assessing small failure probabilities by combined subset simulation and support vector machines. *Structural Safety*, **33**, pp. 343–353.

[5] Bucher, C. and Bourgund, U. (1990). A fast and efficient response surface approach for structural reliability problems. *Structural Safety*, **7**, pp. 57-66.

[6] Das, P.-K. and Zheng, Y. (2000). Cumulative formation of response surface and its use in reliability analysis. *Probabilistic Engineering Mechanics*, **15**, pp. 309-315.

[7] Deheeger, F. and Lemaire, M. (2010). Support vector machines for efficient subset simulations: ^{2}SMART method. In: *Proceedings of the 10th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP10)*, Tokyo, Japan.

[8] Devroye, L., Felber, T., and Kohler, M. (2013). Estimation of a density using real and artificial data. *IEEE Transactions on Information Theory*, **59**, No. 3, pp. 1917-1928.

[9] Felber, T., Kohler, M., and Krzyżak, A. (2015a). Adaptive density estimation based on real and artificial data. *Journal of Nonparametric Statistics*, **27**, pp. 1-18.

[10] Felber, T., Kohler, M., and Krzyżak, A. (2015b). Density estimation with small measurement errors. *IEEE Transactions on Information Theory*, **61**, pp. 3446-3456.

[11] Furer, D. and Kohler, M. (2015). Smoothing spline regression estimation based on real and artificial data. *Metrika*, **78**, pp. 711-746.

[12] Goh, J., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C., and Rutter, E. (2013). Prediction and computer model calibration using outputs from mulitfidelity simulators. *Technometrics*, **55**, pp. 501-512.

[13] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). A Distribution-Free Theory of Nonparametric Regression. *Springer-Verlag*, New York.

[14] Han, G., Santner, T. J., Rawlinson, J. J. (2009). Simultaneous determination of tuning and calibration parameters for computer experiments. *Technometrics*, **51**, pp. 464-474.

[15] Higdon, D., Gattiker, J., Lawrence, E., Jackson, C., Tobis, M., Pratola, M., Habib, S., Heitmann, K., and Price, S. (2013). Computer model calibration using the ensemble kalman filter. *Technometrics*, **55**, pp. 488–500.

[16] Hurtado, J. (2004). *Structural Reliability – Statistical Learning Perspectives.* Vol. 17 of lecture notes in applied and computational mechanics. Springer.

[17] Kaymaz, I. (2005). Application of Kriging method to structural reliability problems. *Strutural Safety*, **27**, pp. 133–151.

[18] Kennedy, M. C., and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society: Series B*, **63**, pp. 425-464.

[19] Kim, S.-H. and Na, S.-W. (1997). Response surface method using vector projected sampling points. *Structural Safety*, **19**, pp. 3–19.

[20] Kohler, M., Krzyżak, A., and Schäfer, D. (2002). Application of structural risk minimization to multivariate smoothing spline regression estimates. *Bernoulli* **8**, pp. 475-489.

[21] Kohler, M., and Krzyżak, A. (2012). Pricing of American options in discrete time using least squares estimates with complexity penalties. *Journal of Statistical Planning and Inference* **142**, pp. 2289-2307.

[22] Mallapur, S., and Platz, R. (2017). Quantification and Evaluation of Uncertainty in the Mathematical Modelling of a Suspension Strut using Bayesian Model Validation Approach. Proceedings of the International Modal Analysis Conference IMAC-XXXV, Garden Grove, California, USA, Paper 117, 30. Jan - 2. Feb., 2017.

[23] Papadrakakis, M. and Lagaros, N. (2002). Reliability–based structural optimization using neural networks and Monte Carlo simulation. *Computer Methods in Applied Mechanics and Engineering*, **191**, pp. 3491–3507.

[24] Santner, T. J., Williams, B. J. and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York.

[25] Tuo, R., and Wu, C. F. J. (2015). Efficient calibration for imperfect computer models. *Annals of Statistics* **43**, pp. 2331-2352.

[26] Wang, S., Chen, W., and Tsui, K. L. (2009). Bayesian validation of computer models. *Technometrics*, **51**, pp. 439-451.

Supplementary material for the referees

Proof of Lemma 1. Since $|\bar{z}_i| \leq L \leq \beta_n$ $(i = 1, \ldots, N)$ and $w_i \geq 0$ $(i = 1, \ldots, N)$ we have

$$\begin{aligned}\sum_{i=1}^{N} w_i \cdot |\bar{z}_i - \hat{m}_N(x_i)|^2 &= \sum_{i=1}^{N} w_i \cdot |\bar{z}_i - T_{\beta_N}(\tilde{m}_N(x_i))|^2 \\ &\leq \sum_{i=1}^{N} w_i \cdot |\bar{z}_i - \tilde{m}_N(x_i)|^2.\end{aligned}$$

This together with the definition of the estimate implies

$$\sum_{i=1}^{N} w_i \cdot |\bar{z}_i - \hat{m}_N(x_i)|^2 + pen^2\left(\tilde{m}_N\right) \leq \sum_{i=1}^{N} w_i \cdot |\bar{z}_i - m_N^*(x_i)|^2 + pen^2\left(m_N^*\right),$$

hence

$$\begin{aligned}&\sum_{i=1}^{N} w_i \cdot |\bar{z}_i - m(x_i)|^2 + 2\sum_{i=1}^{N} w_i \cdot (m(x_i) - \hat{m}_N(x_i)) \cdot (\bar{z}_i - m(x_i)) \\ &\quad + \sum_{i=1}^{N} w_i \cdot |m(x_i) - \hat{m}_N(x_i)|^2 + pen^2\left(\tilde{m}_N\right) \\ &\leq \sum_{i=1}^{N} w_i \cdot |\bar{z}_i - m(x_i)|^2 + 2\sum_{i=1}^{N} w_i \cdot (m(x_i) - m_N^*(x_i)) \cdot (\bar{z}_i - m(x_i)) \\ &\quad + \sum_{i=1}^{N} w_i \cdot |m(x_i) - m_N^*(x_i)|^2 + pen^2\left(m_N^*\right),\end{aligned}$$

which implies

$$\begin{aligned}&\sum_{i=1}^{N} w_i \cdot |m(x_i) - \hat{m}_N(x_i)|^2 + pen^2\left(\tilde{m}_N\right) - \sum_{i=1}^{N} w_i \cdot |m(x_i) - m_N^*(x_i)|^2 - pen^2(m_N^*) \\ &\leq 2\sum_{i=1}^{N} w_i \cdot (\bar{z}_i - m(x_i)) \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) \\ &= 2\sum_{i=1}^{N} w_i \cdot (\bar{z}_i - z_i) \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) + 2\sum_{i=1}^{N} w_i \cdot (z_i - m(x_i)) \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) \\ &=: T_1 + T_2.\end{aligned}$$

We show next that $T_1 \leq T_2$. Assume to the contrary that this is not true. Then

$$\sum_{i=1}^{N} w_i \cdot |m(x_i) - \hat{m}_N(x_i)|^2 + pen^2\left(\tilde{m}_N\right) - \sum_{i=1}^{N} w_i \cdot |m(x_i) - m_N^*(x_i)|^2 - pen^2(m_N^*)$$

$$
\begin{aligned}
&< 4 \sum_{i=1}^{N} w_i \cdot (\bar{z}_i - z_i) \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) \\
&\leq 4 \cdot \sqrt{\sum_{i=1}^{N} w_i \cdot (\bar{z}_i - z_i)^2} \cdot \sqrt{\sum_{i=1}^{N} w_i \cdot (\hat{m}_N(x_i) - m_N^*(x_i))^2} \\
&\leq 4 \cdot \sqrt{\sum_{i=1}^{N} w_i \cdot (\bar{z}_i - z_i)^2} \\
&\cdot \sqrt{2 \sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + 2pen^2\left(\tilde{m}_N\right) + 2 \sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + 2pen^2(m_N^*)}.
\end{aligned}
$$

Using (28) we see that

$$
\begin{aligned}
&\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2\left(\tilde{m}_N\right) - \sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 - pen^2(m_N^*) \\
&\geq \frac{1}{2} \cdot \left(\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2\left(\tilde{m}_N\right) \right) \\
&\quad + \frac{1}{2} \cdot \left(3 \left(\sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + pen^2(m_N^*) \right) + 128 \cdot \sum_{i=1}^{N} w_i \cdot |z_i - \bar{z}_i|^2 + t \right) \\
&\quad - \sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 - pen^2(m_N^*) \\
&\geq \frac{1}{2} \cdot \left(\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2\left(\tilde{m}_N\right) + \sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + pen^2(m_N^*) \right),
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\frac{1}{2} \cdot \sqrt{\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2\left(\tilde{m}_N\right) + \sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + pen^2(m_N^*)} \\
&< 4 \cdot \sqrt{2} \cdot \sqrt{\sum_{i=1}^{N} w_i \cdot |z_i - \bar{z}_i|^2}
\end{aligned}
$$

i.e.,

$$
\begin{aligned}
&\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2\left(\tilde{m}_N\right) + \sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + pen^2(m_N^*) \\
&< 128 \cdot \sum_{i=1}^{N} w_i \cdot |z_i - \bar{z}_i|^2
\end{aligned}
$$

But this is a contradiction to (28), so we have indeed proved $T_1 \leq T_2$. As a consequence we can conclude from (28)

$$
\begin{aligned}
&4\sum_{i=1}^{N} w_i \cdot (\hat{m}_N(x_i) - m_N^*(x_i)) \cdot (z_i - m(x_i)) \\
&\geq \sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2\left(\tilde{m}_N\right) - \sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 - pen^2(m_N^*) \\
&\geq \frac{1}{3}\left(\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + pen^2\left(\tilde{m}_N\right)\right) \\
&\quad + \frac{2}{3}\left(2\sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + 2pen^2(m_N^*) + t\right) \\
&\quad - \sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 - pen^2(m_N^*) \\
&= \frac{1}{3}\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m(x_i)|^2 + \frac{1}{3}pen^2\left(\tilde{m}_N\right) \\
&\quad + \frac{1}{3}\sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + \frac{1}{3}pen^2(m_N^*) + \frac{2}{3}t \\
&= \frac{1}{3}\sum_{i=1}^{N} w_i \cdot |\left(\hat{m}_N(x_i) - m_N^*(x_i)\right) - \left(m(x_i) - m_N^*(x_i)\right)|^2 \\
&\quad + \frac{1}{3}pen^2\left(\tilde{m}_N\right) + \frac{1}{3}\sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + \frac{1}{3}pen^2(m_N^*) + \frac{2}{3}t \\
&\geq \frac{1}{6}\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m_N^*(x_i)|^2 - \frac{1}{3}\sum_{i=1}^{N} w_i \cdot |m(x_i) - m_N^*(x_i)|^2 \\
&\quad + \frac{1}{3}pen^2\left(\tilde{m}_N\right) + \frac{1}{3}\sum_{i=1}^{N} w_i \cdot |m_N^*(x_i) - m(x_i)|^2 + \frac{1}{3}pen^2(m_N^*) + \frac{2}{3}t \\
&\geq \frac{1}{6}\left(\sum_{i=1}^{N} w_i \cdot |\hat{m}_N(x_i) - m_N^*(x_i)|^2 + pen^2\left(\tilde{m}_N\right)\right) + \frac{2}{3}t.
\end{aligned}
$$

In the next to last inequality we have used, that $a^2/2 - b^2 \leq (a-b)^2$ $(a, b \in \mathbb{R})$ with $a = \hat{m}_N(x_i) - m_N^*(x_i)$ and $b = m(x_i) - m_N^*(x_i)$. □

Proof of Lemma 3. We have

$$
\begin{aligned}
&\mathbf{P}\left\{\|m_n - m_n^*\|_n^2 + pen_n^2(\tilde{m}_n) + 4\delta_n \leq \frac{24}{n}\sum_{i=1}^{n}(m_n(x_i) - m_n^*(x_i)) \cdot W_i\right\} \\
&\leq P_1 + P_2
\end{aligned}
$$

where

$$P_1 = \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n} W_i^2 > 2\sigma_0^2\right\}$$

and

$$P_2 = \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n} W_i^2 \le 2\sigma_0^2, \|m_n - m_n^*\|_n^2 + pen_n^2(\tilde{m}_n) + 4\delta_n \le \frac{24}{n}\sum_{i=1}^{n}(m_n(x_i) - m_n^*(x_i)) \cdot W_i\right\}.$$

Application of Chernoff's exponential bounding method (cf. Chernoff (1952)) together with (31) yields

$$\begin{aligned}
P_1 &= \mathbf{P}\left\{\sum_{i=1}^{n} W_i^2/K^2 > 2n\sigma_0^2/K^2\right\} \\
&\le \mathbf{P}\left\{\exp\left(\sum_{i=1}^{n} W_i^2/K^2\right) > \exp\left(2n\sigma_0^2/K^2\right)\right\} \\
&\le \exp\left(-2n\sigma_0^2/K^2\right) \cdot \mathbf{E}\left\{\exp(\sum_{i=1}^{n} W_i^2/K^2)\right\} \\
&\le \exp\left(-2n\sigma_0^2/K^2\right) \cdot \left(1 + \sigma_0^2/K^2\right)^n \\
&\le \exp\left(-2n\sigma_0^2/K^2\right) \cdot \exp\left(n \cdot \sigma_0^2/K^2\right) = \exp\left(-n\sigma_0^2/K^2\right).
\end{aligned}$$

To bound P_2, we observe first that $1/n\sum_{i=1}^{n} W_i^2 \le 2\sigma_0^2$ together with the Cauchy-Schwarz inequality implies

$$\begin{aligned}
\frac{24}{n}\sum_{i=1}^{n}(m_n(x_i) - m_n^*(x_i)) \cdot W_i &\le 24 \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_n(x_i) - m_n^*(x_i))^2} \cdot \sqrt{2\sigma_0^2} \\
&\le 24 \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_n(x_i) - m_n^*(x_i))^2 + pen_n^2(\tilde{m}_n)} \cdot \sqrt{2\sigma_0^2}
\end{aligned}$$

hence inside of P_2 we have

$$\frac{1}{n}\sum_{i=1}^{n}(m_n(x_i) - m_n^*(x_i))^2 + pen_n^2(\tilde{m}_n) \le 1152\sigma_0^2.$$

Set

$$S = \min\{s \in \mathbb{N}_0 \,:\, 4 \cdot 2^s\delta_n > 1152\sigma_0^2\}.$$

Application of the peeling device (cf. Section 5.3 in van de Geer (2000)) yields

$$P_2 = \sum_{s=1}^{S} \mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n} W_i^2 \le 2\sigma_0^2, 4 \cdot 2^{s-1}\delta_n \cdot I_{\{s\neq 1\}} \le \|m_n - m_n^*\|_n^2 + pen_n^2(\tilde{m}_n) < 4 \cdot 2^s\delta_n,\right.$$

$$\|m_n - m_n^*\|_n^2 + pen_n^2(\tilde{m}_n) + 4\delta_n \le \frac{24}{n}\sum_{i=1}^n (m_n(x_i) - m_n^*(x_i)) \cdot W_i \Bigg\}$$

$$\le \sum_{s=1}^S \mathbf{P}\Bigg\{ \frac{1}{n}\sum_{i=1}^n W_i^2 \le 2\sigma_0^2, \|m_n - m_n^*\|_n^2 + pen_n^2(\tilde{m}_n) < 4 \cdot 2^s \delta_n,$$

$$\frac{1}{12} \cdot 2^s \delta_n \le \frac{1}{n}\sum_{i=1}^n (m_n(x_i) - m_n^*(x_i)) \cdot W_i \Bigg\}$$

The probabilities in the above sum can be bounded by Corollary 8.3 in van de Geer (2000) (use there $R = \sqrt{4 \cdot 2^s \delta_n}$, $\delta = \frac{1}{12} \cdot 2^s \delta_n$ and $\sigma = \sqrt{2}\sigma_0$). This yields

$$P_2 \le \sum_{s=1}^\infty c_{60} \exp\left(-\frac{n \cdot (\frac{1}{12} \cdot 2^s \delta_n)^2}{4c_{60} \cdot 4 \cdot 2^s \delta_n} \right) = \sum_{s=1}^\infty c_{60} \exp\left(-\frac{n \cdot 2^s \cdot \delta_n}{c_{61}} \right)$$

$$\le \sum_{s=1}^\infty c_{60} \exp\left(-\frac{n \cdot (s+1) \cdot \delta_n}{c_{60}} \right) \le c_{62} \exp\left(-\frac{n\delta_n}{c_{62}} \right).$$

□

Proof of Lemma 4. For $f : \mathbb{R}^d \to \mathbb{R}$ set

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n |f(X_i)|^2.$$

We have

$$\mathbf{P}\left\{ \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) > \delta_n + 3 \cdot pen_n^2(\tilde{m}_n) + 3\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - m(X_i)|^2 \right\}$$

$$= \mathbf{P}\Bigg\{ 2\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) - 2\|m_n - m\|_n^2$$

$$> \delta_n + 3 \cdot pen_n^2(\tilde{m}_n) + \int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) + \|m_n - m\|_n^2 \Bigg\}$$

$$\le \mathbf{P}\left\{ \exists f \in \mathcal{F}_n : \frac{\left| \int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) - \|T_{\beta_n} f - m\|_n^2 \right|}{\delta_n + 3 \cdot pen_n^2(f) + \int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) + \|T_{\beta_n} f - m\|_n^2} > \frac{1}{2} \right\}$$

$$\le \sum_{s=1}^\infty \mathbf{P}\Bigg\{ \exists f \in \mathcal{F}_n : I_{\{s \ne 0\}} \cdot 2^{s-1} \cdot \delta_n \le pen_n^2(f) \le 2^s \delta_n,$$

$$\frac{\left| \int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) - \|T_{\beta_n} f - m\|_n^2 \right|}{\delta_n + 3 \cdot pen_n^2(f) + \int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) + \|T_{\beta_n} f - m\|_n^2} > \frac{1}{2} \Bigg\}$$

$$\le \sum_{s=1}^\infty \mathbf{P}\Bigg\{ \exists f \in \mathcal{F}_n : pen_n^2(f) \le 2^s \delta_n,$$

$$\frac{\left|\int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) - \|T_{\beta_n} f - m\|_n^2\right|}{2^{s-1}\delta_n + \int |T_{\beta_n} f(x) - m(x)|^2 \mathbf{P}_X(dx) + \|T_{\beta_n} f - m\|_n^2} > \frac{1}{2}\Bigg\}.$$

The probabilities in the above sum can be bounded by Theorem 19.2 in Györfi et al. (2002) (which we apply with

$$\mathcal{F} = \left\{(T_{\beta_n} f - m)^2 \,:\, f \in \mathcal{F}_n,\, pen_n^2(f) \le 2^s \delta_n\right\},$$

$K = 4\beta_n^2$, $\epsilon = 1/2$, and $\alpha = 2^{s-1}\delta_n$. Here in the integral of the covering number we use the fact that for $\delta \ge \alpha \cdot K/2 \ge 2 \cdot \alpha = 2^s \cdot \delta_n$ the condition $pen_n^2(f) \le 2^s \delta_n$ inside $\mathcal{F}$ implies $pen_n^2(f) \le \delta$.) This yields

$$P_{1,n} \le \sum_{s=1}^{\infty} 15 \cdot \exp\left(-\frac{n \cdot 2^s \cdot \delta_n}{c_{63} \cdot \beta_n^2}\right) \le c_{64} \cdot \exp\left(-\frac{n \cdot \delta_n}{c_{64} \cdot \beta_n^2}\right).$$

□