Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neurocomputing 69 (2005) 100–122
NEUROCOMPUTING
www.elsevier.com/locate/neucom

The differogram: Non-parametric noise variance estimation and its use for model selection

Kristiaan Pelckmans[a,*], Jos De Brabanter[a,b], Johan A.K. Suykens[a], Bart De Moor[a]

[a]*K.U. Leuven, ESAT - SCD/SISTA, Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium*
[b]*Hogeschool KaHo Sint-Lieven (Associatie KULeuven), Departement Industrieel Ingenieur, Belgium*

Received 5 April 2004; received in revised form 19 November 2004; accepted 9 February 2005
Available online 29 August 2005

Abstract

Model-free estimates of the noise variance are important in model selection and setting tuning parameters. In this paper a data representation is discussed which leads to such an estimator suitable for multivariate data. Its visual representation—called the *differogram cloud* here—is based on the 2-norm of the differences of input and output data. The crucial concept of locality in this representation is translated as the increasing variance of the difference, which does not rely explicitly on an extra hyper-parameter. Connections with *U*-statistics, Taylor series expansions and other related methods are given. Numerical simulations indicate a convergence of the estimator. This paper extends results towards a time-dependent setting and to the case of non-Gaussian noise models or outliers. As an application, this paper focuses on model selection for Least Squares Support Vector Machines. For this purpose, a variant of the LS-SVM regressor is derived based on Morozov's discrepancy principle relating the regularization constant directly with the (observed) noise level.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Noise level; Complexity criteria; Kernel methods; Least squares support vector machines; Morozov's discrepancy method

*Corresponding author. Tel.: +32 16 328540; fax: +32 16 321970.
E-mail addresses: kristiaan.pelckmans@esat.kuleuven.ac.be (K. Pelckmans), johan.suykens@esat.kuleuven.ac.be (J.A.K. Suykens).
URL: http://www.esat.kuleuven.ac.be/sista/lssvmlab/.

0925-2312/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2005.02.015

1. Introduction

This paper operates in the context of supervised learning (regression) where properties need to be learned from the observed continuous labeled sample. A great deal of effort has gone into developing estimators of the underlying regression model [2,33,36], while the task of estimation of properties of the residuals has often been ignored. However, estimation of the noise variance is also important on its own as it has applications to interval estimation of the function (inference) and determination of the amount of smoothing to be applied. This is readily seen from the important role that it plays in various complexity criteria such as Akaike's information criterion [1] and the C_p-statistic [23]. Another advantage of such estimators is that they give rise to good initial starting values of the tuning parameters in ridge regression, (least squares) support vector machines (SVMs) [6,30,33,36], regularization networks [28] and Nadaraya–Watson kernel estimators; see e.g. [16]. Additional relations between the noise variance and the amount of smoothing required are described in the Bayesian evidence framework and Gaussian Processes [22], statistical learning theory [6,12], splines [38] and regularization theory, e.g. in the form of Morozov's discrepancy principle [25,27].

There are essentially two different approaches to noise variance estimation. The first one is based on an estimated model. This paper focuses on a second, model-free approach, based on differencing the data for removing (local) trends. The proposed method relies on ideas which date back at least from the work of [37]. In the context of geographically distributed measurements [7], a common tool is to visualize second-order properties of the data by inspection of the differences between observations (the so-called variogram-cloud). Furthermore, the variogram presents an interesting relation with the classical auto-covariance function (as frequently used in the context of time-series analysis). This paper extends insights from these fields towards a machine learning and kernel machines setting.

This paper is organized as follows. In Section 2 the literature on noise variance estimators is briefly reviewed. Section 3 deals with the differogram and the derived noise level estimator. Section 4 extends those results to a time-dependent setting and to cases of non-Gaussian noise and outliers. Section 5 gives a direct application useful for the determination of the regularization parameter based on Morozov's discrepancy principle and presents a number of applications in the context of model selection. Section 6 illustrates the methods on a number of examples.

2. Estimating the variance of the noise

2.1. Model-based estimators

Given a random vector (X, Y) where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, let $\{(x_i, y_i)\}_{i=1}^N$ be samples of the random vector satisfying the relation

$$y_i = f(x_i) + e_i, \quad i = 1, \ldots, N. \tag{1}$$

The error terms e_i are assumed to be uncorrelated random variables with zero mean and variance $\sigma^2 < \infty$ (independent and identically distributed, i.i.d.), and $f : \mathbb{R}^d \to \mathbb{R}$ a smooth function. The same setting was adopted e.g. in [9]. An estimate $\hat{f}$ of the underlying function can be used to estimate the noise variance by suitably normalizing the sums of squares of its associated residuals, see e.g. [38]. A broad class of model-based variance estimators can be written as

$$\hat{\sigma}_e^2 = \frac{y^{\mathrm{T}} Q y}{\mathrm{tr}[Q]}$$

with $y = (y_1, \ldots, y_N)^{\mathrm{T}}$ [4]: $\mathrm{tr}(\cdot)$ denotes the trace of the matrix and $Q = (I_N - S)^2$ a symmetric $N \times N$ positive definite matrix. Let $\hat{y}_i = \hat{f}(x_i)$ and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_N)^{\mathrm{T}} \in \mathbb{R}^N$. For most modeling methods, one can determine a smoother matrix $S \in \mathbb{R}^{N \times N}$ with $\hat{y} = Sy$ such as e.g. in the cases of ridge regression, smoothing splines [11] or least squares support vector machines (LS-SVMs) [33].

2.2. Model-free estimators

Model-free variance estimators were proposed in the case of equidistantly ordered data. In the work of [29,13], such estimators of σ^2 have been proposed based on first- and second-order differences of the values of y_i, respectively. For example Rice suggested estimating σ^2 by

$$\hat{\sigma}^2 = \frac{1}{2(N-1)} \sum_{i=1}^{N-1} (y_{i+1} - y_i)^2. \tag{2}$$

Gasser et al. [13] have suggested a similar idea for removing local trend effects by using

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=2}^{N-1} c_i^2 \hat{\varepsilon}_i^2, \tag{3}$$

where $\hat{\varepsilon}_i$ is the difference between y_i and the value at x_i of the line joining (x_{i-1}, y_{i-1}) and (x_{i+1}, y_{i+1}). The values c_i are chosen to ensure that $E[c_i^2 \hat{\varepsilon}_i^2] = \sigma^2$ for all i when the function f in (1) is linear. Note that one assumes that $x_1 < \cdots < x_N$, $x_i \in \mathbb{R}$ in both methods.

In the case of non-equidistant or higher dimensional data an alternative approach is based on a density estimation technique. Consider the regression model as defined in (1). Assume that $e_1, \ldots, e_N$ are i.i.d. with a common probability distribution function F belonging to the family

$$\mathscr{F} = \left\{ F : \int x \, \mathrm{d}F(x) = 0, \; 0 < \int |x|^r \, \mathrm{d}F(x) < \infty \right\},$$
$$r \in \mathbb{N}_0 \text{ and } 1 \leqslant r \leqslant 4. \tag{4}$$

Let $K : \mathbb{R}^d \to \mathbb{R}$ be a function called the kernel function and let $h > 0$ be a bandwidth or smoothing parameter. Then [26] suggested an error variance

estimator given by

$$\hat{\sigma}_e^2 = \frac{1}{N(N-1)h} \sum_{1 \leqslant i < j \leqslant N} \frac{1}{2}(y_i - y_j)^2 \, \frac{1}{2}\left(\frac{1}{\hat{f}_i} + \frac{1}{\hat{f}_j}\right) K\left(\frac{x_i - x_j}{h}\right), \tag{5}$$

where $\hat{f}_i$ is defined as

$$\hat{f}_i = \frac{1}{(N-1)h} \sum_{j \neq i} K\left(\frac{x_i - x_j}{h}\right), \quad i = 1, \ldots, N. \tag{6}$$

The cross-validation principle can be used to select the bandwidth h. This paper is related to (5) and (6) but avoids the need for an extra hyper-parameter such as the bandwidth and is naturally extendible to higher dimensional data.

3. Variogram and differogram

One way to model the correlation structure of a Gaussian process is to estimate the variogram or semi-variogram:

Definition 1 (*Semi-variogram Cressie [7]*). Let $\{Z(x_i), i \in \mathbb{N}\}$ be a stationary Gaussian process with mean $\bar{z}$, $\mathrm{Var}[Z(x_i)] < \infty$ for all $i \in \mathbb{N}$ and a correlation function which only depends on $\Delta_{x,ij}^2 = \|x_i - x_j\|_2^2$ for all $i,j \in \mathbb{N}$. It follows from the stationarity of the process $Z(x_1), \ldots, Z(x_N)$ that

$$\tfrac{1}{2}E[(Z(x_i) - Z(x_j))^2] = \sigma^2 + \tau^2(1 - \rho(\Delta_{x,ij}^2)) = \eta(\Delta_{x,ij}^2), \quad \forall i,j \in \mathbb{N}, \tag{7}$$

where σ^2 is the small-scale variance (the nugget effect), τ^2 is the variance of the serial correlation component and $\rho : \mathbb{R} \to \mathbb{R}$ is the correlation function [10,7]. The function $\eta : \mathbb{R} \to \mathbb{R}^+$ is called the semi-variogram.

The prefix *semi*-refers to the constant $\frac{1}{2}$ in the definition. A scatter-plot of the differences is referred to as the variogram cloud. A number of parametric models were proposed to model η [7]. Estimation of the parameters of a variogram model often employs a maximum likelihood criterion [7] leading (in most cases) to non-convex optimization problems. The variogram can be considered as being complementary to the auto-covariance function of a Gaussian process as $E(Z(x_i) - Z(x_j))^2 = 2E(Z(x_i))^2 - 2E(Z(x_i)Z(x_j))$. The auto-covariance function is often employed in an equidistantly sampled setting in time-series analysis and stochastic system identification, while the variogram allows to handle non-equidistantly sampled data.

Instead of working with a Gaussian process Z, machine learning is concerned (amongst others) with learning an unknown smooth regression function $f : \mathbb{R}^d \to \mathbb{R}$ from observations $\{(x_i, y_i)\}_{i=1}^N$ sampled from the random vector (X, Y). We now define the differogram similar to the semi-variogram as follows:

Definition 2 (*Differogram*). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous function such that $y_i = f(x_i) + e_i$. Let $\Delta_{x,ij}^2 = \|x_i - x_j\|_2^2$ for all $i,j = 1, \ldots, N$ be samples of the

random variable Δ_X^2 and let $\Delta_{y,ij}^2 = \|y_i - y_j\|_2^2$ be samples from the random variable Δ_Y^2. The differogram function Υ: $\mathbb{R}^+ \to \mathbb{R}^+$ is defined as

$$\Upsilon(\Delta_x^2) = \tfrac{1}{2} E[\Delta_Y^2 \mid \Delta_X^2 = \Delta_x^2]. \tag{8}$$

This function is well-defined as the expectation operator results in a unique value for each different conditioning $\Delta_X^2 = \Delta_x^2$ by definition [24]. A main difference with the semi-variogram is that the differogram does not assume an isotropic structure of the regression function f. A motivation for this choice is that the differogram will be of main interest in the direct region of $\Delta_X^2 = 0$ where the isotropic structure emerges because of the Lipschitz assumption. A similar reasoning lies at the basis of the use of RBF-kernels and nearest-neighbor methods [16,9].

Although the definition is applicable to the multivariate case, some intuition is given by considering the case of one-dimensional inputs. Let $\Delta_{e_{ij}}^2 = (e_i - e_j)^2$ be samples from the random variable Δ_e^2. For one-dimensional linear models $y_i = wx_i + b + e_i$ where $w, b \in \mathbb{R}$ and $\{e_i\}_{i=1}^N$ is an i.i.d. sequence where the inputs are standardized (zero mean and unit variance), the differogram equals $\Upsilon_w(\Delta_x^2) = \frac{1}{2} w \Delta_X^2 + \frac{1}{2} E[\Delta_e^2]$, as illustrated in Fig. 1. Fig. 2 presents the differogram

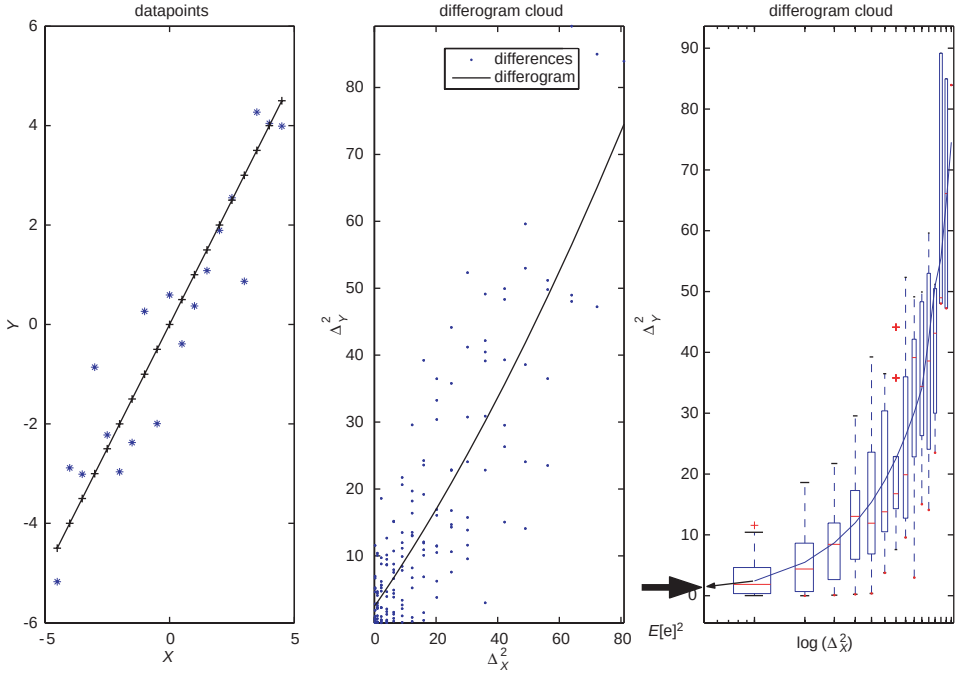


Fig. 1. Differogram of a linear function: (a) data are generated from $y_i = x_i + e_i$ with $e_i \sim \mathcal{N}(0,1)$, i.i.d. and $i = 1, \ldots, N = 25$; (b) all differences $\Delta_{x,ij}^2 = \|x_i - x_j\|_2^2$ and $\Delta_{y,ij}^2 = \|y_i - y_j\|_2^2$ for $i<j = 1, \ldots, N$. The solid line represents the estimated differogram model and (c) all differences boxed using a log scale for $\Delta_{x,ij}^2$. The intercept of the curve crossing the Y-axis corresponds to twice the estimated noise variance $2\hat{\sigma}_e^2$.

cloud and the (estimated) differogram function of a non-linear regression, while Section 6 reports on some experiments on higher dimensional data.

Equivalent to the nugget effect in the variogram, one can prove the following lemma relating the differogram function to the noise variance.

Lemma 1. *Assume a Lipschitz continuous function* $f : \mathbb{R}^d \to \mathbb{R}$ *such that* $\exists M \in \mathbb{R}^+$ *where* $\|f(X) - f(X')\|_2^2 \leqslant M\|X - X'\|_2^2$ *with* X' *being a copy of the random variable* X. *Let* $\{(x_i, y_i)\}_{i=1}^N$ *be sampled from the random vector* (X, Y) *and* e *obeying the relation* $Y = f(X) + e$. *Assume that the random variable* e *has bounded moments and is independent of* $f(X)$. *Under these assumptions, the limit* $\lim_{\Delta_x^2 \to 0} \Upsilon(\Delta_x^2)$ *exists and equals* σ_e^2.

Proof. Let $\Delta_{e,ij}^2 = (e_i - e_j)^2$ be samples of the random variable $\Delta_e^2 = (e - e')^2$ where e' is a copy of the random variable e. As the residuals are not correlated, it follows that $E[\Delta_e^2] = E[e^2] + 2E[ee'] + E[e'^2] = 2\sigma_e^2$. Substitution of the definition of the Lipschitz continuity into the definition of the differogram gives

$$\begin{aligned} 2\Upsilon(\Delta_x^2) &= E[\Delta_Y^2 \mid \Delta_X^2 = \Delta_x^2] \\ &= E\left[(f(X) + e - f(X') - e')^2 \mid \|X - X'\|_2^2 = \Delta_x^2\right] \\ &= E\left[(e - e')^2 + (f(X) - f(X'))^2 \mid \|X - X'\|_2^2 = \Delta_x^2\right] \\ &\leqslant E[\Delta_e^2 + M\|X - X'\|_2^2 \mid \|X - X'\|_2^2 = \Delta_x^2] \\ &= 2\sigma_e^2 + E[M\|X - X'\|_2^2 \mid \|X - X'\|_2^2 = \Delta_x^2] \\ &= 2\sigma_e^2 + M\Delta_x^2, \end{aligned} \tag{9}$$

where the independence between the residuals and the function f (and hence between Δ_e^2 and $(f(X) - f(X'))^2$), and the linearity of the expectation operator E is used [24]. From this result, it follows that $\lim_{\Delta_x^2 \to 0} \Upsilon(\Delta_x^2) \to \sigma_e^2$. □

The differogram function will only be of interest near the limit $\Delta_x^2 \to 0$ in the sequel. A similar approach was presented in [9] where the nearest-neighbor paradigm replaces the conditioning on Δ_X^2 and fast rates of convergence were proved.

3.1. Differogram models based on Taylor series expansions

Consider the Taylor series expansion of order r centered at $m \in \mathbb{R}$ for local approximation in $x_i \in \mathbb{R}$ for all $i = 1, \ldots, N$

$$T_r[f(x_i)](m) = f(m) + \sum_{l=1}^{r} \frac{1}{l!} \nabla^{(l)} f(m)(x_i - m)^l + \mathcal{O}(x_i - m)^{r+1}, \tag{10}$$

where $\nabla f(x) = \frac{\partial f}{\partial x}$, $\nabla^2 f(x) = \frac{\partial^2 f}{\partial x^2}$, etc. for $l \geqslant 2$. One may motivate the use of an rth order Taylor series approximation of the differogram function with the center $m = 0$ as a suitable model because one is only interested in the case $\Delta_x^2 \to 0$:

$$\Upsilon_{\mathscr{A}}(\Delta_x^2) = a_0 + \mathscr{A}(\Delta_x^2), \quad \text{where } \mathscr{A}(\Delta_x^2) = \sum_{l=1}^{r} a_l (\Delta_x^2)^l, \ a_0, \ldots, a_r \in \mathbb{R}^+, \tag{11}$$

where the parameter vector $a = (a_0, a_1, \ldots, a_r)^{\mathrm{T}} \in \mathbb{R}^{+,r+1}$ is assumed to exist uniquely. The elements of the parameter vector a are enforced to be positive as the (expected) differences should always be strictly positive. The function ϑ of the mean absolute deviation of the estimate can be bounded as follows:

$$\begin{aligned}\vartheta(\Delta_x^2; a) &= E[|\Delta_Y^2 - \Upsilon_{\mathscr{A}}(\Delta_X^2; a)| \,|\, \Delta_X^2 = \Delta_x^2] \\ &= E\left[|\Delta_Y^2 - a_0 - \sum_{l=1}^{r} a_l (\Delta_X^2)^l| \,|\, \Delta_X^2 = \Delta_x^2\right] \\ &\leqslant E\left[\left|a_0 + \sum_{l=1}^{r} a_l (\Delta_x^2)^l\right|\right] + E[|\Delta_Y^2| \,|\, \Delta_X^2 = \Delta_x^2] \\ &= 3\left(a_0 + \sum_{l=1}^{r} a_l (\Delta_x^2)^l\right) \triangleq \bar{\vartheta}(\Delta_x^2; a), \end{aligned} \tag{12}$$

where, respectively, the triangle inequality, the property $|\Delta_Y^2| = \Delta_Y^2$ and Definition 2 are used. The function $\bar{\vartheta} : \mathbb{R}^+ \to \mathbb{R}^+$ is defined as an upper bound to the spread of the samples Δ_Y^2 from the function $\Upsilon(\Delta_x^2)$. Instead of deriving the parameter vector a from the (estimated) underlying function f, it is estimated immediately based on the observed differences $\Delta_{x,ij}^2$ and $\Delta_{y,ij}^2$ for $i<j = 1, \ldots, N$. The following weighted least squares method can be used:

$$a^* = \arg\min_{a \in \mathbb{R}_+^{r+1}} \mathscr{J}(a) = \sum_{i \leqslant j}^{N} \frac{c}{\bar{\vartheta}(\Delta_{x,ij}^2; a)} (\Delta_{y,ij}^2 - \Upsilon_{\mathscr{A}}(\Delta_{x,ij}^2; a))^2, \tag{13}$$

where the constant $c \in \mathbb{R}_0^+$ normalizes the weighting function such that $1 = \sum_{i<j} c/\bar{\vartheta}(\Delta_{x,ij}^2; a)$. The function $\bar{\vartheta}$ corrects for the heteroscedastic variance structure inherent to the differences (see e.g. [32]). As the parameter vector a is positive, the weighting function is monotonically decreasing and as such always represents a local weighting function.

3.2. The differogram for noise variance estimation

A U-statistic is proposed to estimate the variance of the noise from observations.

Definition 3 (*U-statistic Hoeffding [18]*). Let $g : \mathbb{R}^l \to \mathbb{R}$ be a measurable and symmetric function and let $\{u_i\}_{i=1}^N$ be i.i.d. samples drawn from a fixed but unknown distribution. The function

$$U_N = U(g; u_1, \ldots, u_N) = \frac{1}{\binom{N}{l}} \sum_{1 \leqslant i_1 \leqslant \cdots \leqslant i_l \leqslant N} g(u_{i_1}, \ldots, u_{i_l}), \tag{14}$$

for $l<N$, is called a U-statistic of degree l with kernel g.

It is shown [21] that for every unbiased estimator based on the same observations, a U-statistic exists with a smaller variance of the corresponding estimator. If the regression function was known, the errors e_i for all $i = 1, \ldots, N$ were observable and

the sample variance can be written as a U-statistic of order $l = 2$

$$\hat{\sigma}_e^2 = U(g; e_1, \ldots, e_N) = \frac{2}{N(N-1)} \sum_{1 \leqslant i \leqslant j \leqslant N} g_1(e_i, e_j) \quad \text{and}$$

$$g_1(e_i, e_j) = \tfrac{1}{2}(e_i - e_j)^2 = \tfrac{1}{2}\Delta_{e,ij}^2. \tag{15}$$

However, the true function f is not known in practice. A key step deviating from classical practice is to abandon trying to estimate the global function [36] or the global correlation structure [7]. Instead, knowledge of the average local behavior is sufficient for making a distinction between smoothness in the data and unpredictable noise. As an example, consider $r = 0$, the 0th-order Taylor polynomial of f centered at x_i evaluated at x_j for all $i, j = 1, \ldots, N$. This approximation scheme is denoted as $T_0[f(x_j)](x_i) = f(x_i)$ such that (15) becomes

$$\begin{aligned}
\hat{\sigma}_e^2 &= \frac{2}{N(N-1)} \sum_{1 \leqslant i \leqslant j \leqslant N} \frac{1}{2}(y_i - y_j)^2 \\
&\approx \frac{2}{N(N-1)} \sum_{1 \leqslant i \leqslant j \leqslant N} \frac{1}{2}(e_i + f(x_i) - e_j - T_0[f(x_j)](x_i))^2 \\
&= \frac{2}{N(N-1)} \sum_{1 \leqslant i \leqslant j \leqslant N} \frac{1}{2}\Delta_{e,ij}^2,
\end{aligned} \tag{16}$$

where the approximation improves as $x_i \to x_j$. To correct for this, a localized second-order isotropic kernel $g_2 : \mathbb{R}^2 \to \mathbb{R}$ can be used

$$g_2(y_i, y_j) = \frac{c}{2\bar{\vartheta}(\Delta_{x,ij}^2)} \Delta_{y,ij}^2, \tag{17}$$

where the decreasing weighting function $1/\bar{\vartheta}(\Delta_x^2)$ is taken from (12) in order to favor good (local) estimates. The constant $c \in \mathbb{R}_0^+$ is chosen such that the sum of the weighting terms are constant: $2c(\sum_{i \leqslant j}^N 1/\bar{\vartheta}(\Delta_{x,ij}^2)) = N(N-1)$.

From these derivations one may motivate the following kernel for a U-statistic based on the differogram model (11) and weighting function as derived in (12):

$$g_3(y_i, y_j) = \frac{c}{2\bar{\vartheta}(\Delta_{x,ij}^2)} (\Delta_{y,ij}^2 - \mathscr{A}(\Delta_{x,ij}^2)) \quad \text{with } \bar{\vartheta}(\Delta_{x,ij}^2) = 3(a_0 + \mathscr{A}(\Delta_{x,ij}^2)), \tag{18}$$

where $c \in \mathbb{R}_0^+$ is a normalization constant. The resulting U-statistic becomes

$$\hat{\sigma}_e^2 = \frac{2}{N(N-1)} \sum_{1 \leqslant i \leqslant j \leqslant N} g_3(y_i, y_j). \tag{19}$$

One can show that this U-estimator equals the estimated intercept of the differogram model (11):

Lemma 2. *Let $x_1, \ldots, x_N \in \mathbb{R}^d$ and $y_1, \ldots, y_N \in \mathbb{R}$ be samples drawn according to the distribution of the random vector (X, Y) with joint distribution F. Consider a U-statistic as in Definition 3 with kernel g such that $g : \mathbb{R}^l \to \mathbb{R}$ is a measurable and symmetric function. Consider the differogram according to Definition 2 and the differogram model*

(11). *The estimator of the weighted U-statistic* (18) *of the noise variance estimator* (19) *equals the intercept* a_0 *of the estimated differogram model using the weighted least squares estimate* (13).

Proof. This can be readily seen as the expectation can be estimated empirically in two equivalent ways. Consider for example the mean μ of the error terms $e_1,\ldots,e_N$ which can be estimated as $\hat{\mu}=\arg\min_\mu \sum_{i=1}^N (e_i-\mu)^2$ and as $\hat{\mu}=\frac{1}{N}\sum_{i=1}^N e_i$, see e.g. [17]. As previously, one can write

$$\begin{aligned} 2\hat{\sigma}_e^2 &= \lim_{\Delta_x^2\to 0} E[\Delta_Y^2 \mid \Delta_X^2=\Delta_x^2] \\ &= \lim_{\Delta_x^2\to 0} E\left[\frac{c}{\bar{\vartheta}(\Delta_X^2)}(\Delta_Y^2-\mathscr{A}(\Delta_X^2))\,\middle|\,\Delta_X^2=\Delta_x^2\right], \end{aligned} \tag{20}$$

if $\lim_{\Delta_x^2\to 0}\mathscr{A}(\Delta_x^2)=0$. The sample mean estimator becomes

$$\begin{aligned} 2\hat{\sigma}_e^2 &= \frac{2}{N(N-1)}\sum_{k=1}^{N(N-1)/2}\frac{c}{2\bar{\vartheta}(\Delta_{x,k}^2)}(\Delta_{y,k}^2-\mathscr{A}(\Delta_{x,k}^2)) \\ &= \frac{2}{N(N-1)}\sum_{i<j}^{N}\frac{c}{2\bar{\vartheta}(\Delta_{x,ij}^2)}(\Delta_{y,ij}^2-\mathscr{A}(\Delta_{x,ij}^2)) \\ &= U(g_3;u_1,\ldots,u_N), \end{aligned} \tag{21}$$

where a unique index $k=1,\ldots,N(N-1)/2$ corresponds with every distinct pair $1\leqslant i<j\leqslant N$. Alternatively, using the least squares estimate

$$\begin{aligned} 2\hat{\sigma}_e^2 &= \arg\min_{a_0\geqslant 0}\sum_{k=1}^{N(N-1)/2}\frac{c}{\bar{\vartheta}(\Delta_{x,k}^2)}(\Delta_{y,k}^2-\mathscr{A}(\Delta_{x,k}^2)-a_0)^2 \\ &= \arg\min_{a_0\geqslant 0}\sum_{i<j}\frac{c}{\bar{\vartheta}(\Delta_{x,ij}^2)}(\Delta_{y,ij}^2-\mathscr{A}(\Delta_{x,ij}^2)-a_0)^2. \end{aligned} \tag{22}$$

In both cases, the function $\mathscr{A}:\mathbb{R}^+\to\mathbb{R}^+$ of the differogram model and the weighting function $\bar{\vartheta}:\mathbb{R}^+\to\mathbb{R}^+$ are assumed to be known from (13). □

4. Extensions

4.1. Spatio-temporal data

This subsection extends the results towards a setting of mixed inputs. Let the input variables consist of a temporal part and a spatial part, e.g. [7]. Both are considered as random variables X and T. The observations are generated as

$$y_{t,i}=f(x_i)+g(e_{t-1})+e_{t,i}, \tag{23}$$

where $y_{t,i}$ is sampled from random variable Y and $\{e_i\}_{i,t}$ contains an i.i.d. sample from e. More specifically, one works under the assumption that the spatial local

correlation induced by the underlying smooth function $f : \mathbb{R}^d \to \mathbb{R}$ and the temporal correlation structure induced by $g : \mathbb{R} \to \mathbb{R}$ of the data are uncorrelated. The additive structure can be motivated when time and space are assumed to be independent [16]. The graphical representation of the differogram cloud is extended in a third direction, the (absolute value of) the difference in time denoted as $\Delta_{t,ij} = |t_i - t_j|$. Let $\Delta_{t,ij}$ be samples of the random variable Δ_T. The definition of the differogram is extended similarly

$$\Upsilon^{\mathrm{st}}(\Delta_x^2, \Delta_t) = \tfrac{1}{2} E[\Delta_Y^2 \mid \Delta_X^2 = \Delta_x^2,\ \Delta_T = \Delta_t]. \tag{24}$$

As in the previous section, we consider the differogram models

$$\Upsilon^{\mathrm{st}}_{\mathscr{A},\mathscr{B}}(\Delta_x^2, \Delta_t) = a_0 + \mathscr{A}(\Delta_x^2) + \mathscr{B}(\Delta_t), \quad \mathscr{A}(\Delta_x^2) = \sum_{i=1}^{r_1} a_i (\Delta_x^2)^i,$$

$$\mathscr{B}(\Delta_t) = \sum_{j=1}^{r_2} b_j (\Delta_t)^j \quad \text{where } a_0, \ldots, a_{r_1}, b_1, \ldots, b_{r_2} \in \mathbb{R}^+, \tag{25}$$

where we denote $a = (a_0, a_1, \ldots, a_{r_1})^{\mathrm{T}}$ and $b = (b_1, \ldots, b_{r_2})^{\mathrm{T}}$. Similar to the derivation of Eq. (13), one can motivate the use of the following cost-function:

$$\begin{aligned}(a^*, b^*) &= \arg \min_{0 \leqslant a \in \mathbb{R}^{r1}, 0 \leqslant b \in \mathbb{R}^{r_2}} \mathscr{J}^{\mathrm{st}}(a, b) \\ &= \sum_{i \leqslant j} \frac{c}{\bar{\vartheta}^2_{\mathrm{st}}(\Delta^2_{x,ij}, \Delta_{t,ij})} (\Delta^2_{y,ij} - \Upsilon^{\mathrm{st}}_{\mathscr{A},\mathscr{B}}(\Delta^2_{x,ij}, \Delta_{t,ij}; a, b))^2,\end{aligned} \tag{26}$$

where $\bar{\vartheta}_{\mathrm{st}}(\Delta^2_{x,ij}, \Delta_{t,ij})$ is an upper bound on the variance of the model:

$$\begin{aligned}\vartheta_{\mathrm{st}}(\Delta_x^2, \Delta_t; a, b) &= E[|\Delta_Y^2 - \Upsilon^{\mathrm{st}}_{\mathscr{A},\mathscr{B}}(\Delta_X^2, \Delta_T; a, b)| \mid \Delta_X^2 = \Delta_x^2, \Delta_T = \Delta_t] \\ &\leqslant \left| a_0 + \sum_{i=1}^{r_1} a_i (\Delta_x^2)^i + \sum_{j=1}^{r_2} b_j (\Delta_t)^j \right| + E[|\Delta_Y^2| \mid \Delta_X^2 = \Delta_x^2, \Delta_T = \Delta_t] \\ &= 3 \left(a_0 + \sum_{i=1}^{r_1} a_i (\Delta_x^2)^i + \sum_{j=1}^{r_2} b_j (\Delta_t)^j \right) \triangleq \bar{\vartheta}_{\mathrm{st}}(\Delta_x^2, \Delta_t; a, b),\end{aligned} \tag{27}$$

and c normalizes the weighting terms such that the sum equals $0.5N(N-1)$. A U-statistic of order 2 is used for estimating the noise level using the following kernel in (19) which is localized in space as well as in time:

$$g_4(y_i, y_j) = \frac{c}{\bar{\vartheta}_{\mathrm{st}}(\Delta^2_{x,ij}, \Delta_{t,ij}; a, b)} (\Delta^2_{y,ij} - \mathscr{A}(\Delta^2_{x,ij}) - \mathscr{B}(\Delta_{t,ij})). \tag{28}$$

4.2. Robust estimation

A common view on robustness is to provide alternatives to least squares methods when deviating from the assumptions of Gaussian distributions as in Section 3.1.

This part considers the extension of the differogram towards contaminated noise models [20], where a typical observations (outliers) occur in addition to the normal noise model. A robust cost-function was designed based on a Huber loss function $\mathscr{H} : \mathbb{R} \to \mathbb{R}^+$ given as

$$\mathscr{H}(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leqslant \beta, \\ \beta|x| - \frac{1}{2}\beta^2 & |x| > \beta, \end{cases} \tag{29}$$

with $\beta \in \mathbb{R}_0^+$. Note that the weighting function need not be robustified as it only depends on the (already robustified) differogram model using (29). This results in the following modification of the weighted cost-function (13):

$$a^* = \arg \min_{0 \leqslant a \in \mathbb{R}^{r+1}} \mathscr{J}^{\mathscr{H}}(a) = \sum_{i \leqslant j} \frac{c}{\bar{\vartheta}(\Delta_{x,ij}^2; a)} \mathscr{H}(\Delta_{y,ij}^2 - \Upsilon_{\mathscr{A}}(\Delta_{x,ij}^2; a)), \tag{30}$$

where c is a normalizing constant for the weighting function. If $\bar{\vartheta}(\Delta_{x,ij}^2; a)$ were known, the corresponding convex optimization problem can be solved efficiently as noted e.g. in [36]. In the case of large-scale data sets, one can employ an iteratively re-weighted least squares method instead. The estimated value for a_0 is then a robust estimate for the noise variance.

5. Applications

5.1. Morozov's discrepancy principle

As a possible application of the noise variance estimator, consider the derivation of the LS-SVM regressor as described in [33]. A closely related derivation can be made based on Morozov's discrepancy principle [25,27]. Let again $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ be samples drawn from the distribution of the random vector (X, Y), denoted as the training data with inputs x_i and outputs y_i. Consider the regression model $y_i = f(x_i) + e_i$ where $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown real-valued smooth function and $e_1, \ldots, e_N$ satisfy $E[e_i] = 0$, $E[e_i^2] = \sigma_e^2 < \infty$ and $E[e_i e_j] = \delta_{ij}$ where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The primal LS-SVM model for regression is given as $f(x) = w^{\mathrm{T}}\varphi(x) + b$ where $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ denotes the potentially infinite ($n_h = \infty$)-dimensional feature map (Table 1).

Morozov's discrepancy principle [25] is explained now in a kernel context. It considers minimizing $\|w\|_2$ given a fixed noise level σ^2. Formulated in the primal-dual LS-SVM context [33], it is given by

$$\min_{w,b,e_i} \mathscr{J}_M(w) = \frac{1}{2} w^{\mathrm{T}} w \text{ s.t.} \begin{cases} w^{\mathrm{T}}\varphi(x_i) + b + e_i = y_i, & \forall i = 1, \ldots, N, \\ N\sigma^2 = \sum_{i=1}^N e_i^2. \end{cases} \tag{31}$$

Table 1
Numerical results from the experiment comparing different model selection strategies for regularization parameter tuning in the context of Morozov-based LS-SVM regressors

	Morozov	Differogram	10-fold CV	Leaveoneout	Bayesian	C_p	"true"
Toy example: 25 datapoints							
Mean (MSE)	0.4238	0.4385	0.3111	0.3173	0.3404	1.0072	0.2468
std (MSE)	1.4217	1.9234	0.3646	1.5926	0.3614	1.0727	0.1413
Toy example: 200 datapoints							
Mean (MSE)	0.1602	0.2600	0.0789	0.0785	0.0817	0.0827	0.0759
std (MSE)	0.0942	0.5240	0.0355	0.0431	0.0289	0.0369	0.0289
Boston housing data set							
Mean (MSE)	–	0.1503	0.1538	0.1518	0.1522	0.3563	0.1491
std (MSE)	–	0.0199	0.0166	0.0217	0.0152	0.1848	0.0184

The different methods are based, respectively, on exact prior knowledge of the noise level, a model-free estimate using the differogram and some standard data-driven methods such as L-fold cross-validation, leave-one-out, Mallows C_p statistic, and Bayesian inference. Results from experiments on regularization constant tuning in the context of LS-SVMs. The experiment compares results when using Morozov's discrepancy principle in combination with exact prior knowledge of the noise level, using the estimate of the differogram. For comparison, results are included from classical data-driven techniques such as 10-fold cross-validation, leave-one-out, Mallows C_p statistic and Bayesian inference of the hyper-parameters.

The Lagrangian can be written using Lagrange multipliers $\alpha_1, \ldots, \alpha_N, \xi \in \mathbb{R}$

$$\mathscr{L}_M(w, b, e_i; \alpha_i, \xi) = \frac{1}{2} w^{\mathrm{T}} w - \xi \left(\sum_{i=1}^{N} e_i^2 - N\sigma^2 \right) - \sum_{i=1}^{N} \alpha_i (w^{\mathrm{T}} x_i + b + e_i - y_i). \tag{32}$$

The conditions for optimality are

$$\begin{aligned}
&\frac{\partial \mathscr{L}_M}{\partial w} = 0 \to w = \sum_{i=1}^{N} \alpha_i \varphi(x_i), \quad i = 1, \ldots, N, \\
&\frac{\partial \mathscr{L}_M}{\partial b} = 0 \to \sum_{i=1}^{N} \alpha_i = 0, \\
&\frac{\partial \mathscr{L}_M}{\partial e_i} = 0 \to 2\xi e_i = \alpha_i, \quad i = 1, \ldots, N, \\
&\frac{\partial \mathscr{L}_M}{\partial \alpha_i} = 0 \to w^{\mathrm{T}} \varphi(x_i) + b + e_i = y_i, \quad i = 1, \ldots, N, \\
&\frac{\partial \mathscr{L}_M}{\partial \xi} = 0 \to \sum_{i=1}^{N} e_i^2 = N\sigma^2.
\end{aligned} \tag{33}$$

This set of equations can be rewritten in matrix notation as the dual problem

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + \frac{1}{2\xi} I_N \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \text{s.t.} \quad N\sigma^2 = \frac{\alpha^{\mathrm{T}}\alpha}{4\xi^2}, \tag{34}$$

where $\Omega \in \mathbb{R}^{N\times N}$ and $\Omega_{ij} = K(x_i, x_j)$. For the choice of the kernel $K(\cdot,\cdot)$, see e.g. [5,30]. Typical examples are the use of a linear kernel $K(x_i, x_j) = x_i^{\mathrm{T}} x_j$ or the RBF kernel $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2/h^2)$ where h denotes the bandwidth of the kernel. The final regressor $\hat{f}$ can be evaluated at a new data-point x_* using the dual expression as

$$f(x_*) = \sum_{i=1}^{N} \alpha_i K(x_i, x_*) + b, \tag{35}$$

where α_i and b are the solutions to (34). The singular value decomposition (SVD) [14] of Ω is given as

$$\Omega = U\Gamma U^{\mathrm{T}} \quad \text{s.t.} \quad U^{\mathrm{T}}U = I_N, \tag{36}$$

where $U \in \mathbb{R}^{N\times N}$ is an orthonormal matrix $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_N)$ and $\gamma_i \geqslant 0$ for all $i = 1, \ldots, N$. In the following, the intercept term b is omitted from the derivations $b = 0$ for reasons of clarity. Using orthonormality, conditions (34) can then be rewritten as

$$\alpha = U\left(\Gamma + \frac{1}{2\xi} I_N\right)^{-1} p,$$

$$N\sigma^2 = \frac{\alpha^{\mathrm{T}}\alpha}{4\xi^2} = \frac{1}{4\xi^2} p^{\mathrm{T}} \left(\Gamma + \frac{1}{2\xi} I_N\right)^{-2} p = \sum_{i=1}^{N} \left(\frac{p_i}{2\xi\gamma_i + 1}\right)^2, \tag{37}$$

where $p = U^{\mathrm{T}} y \in \mathbb{R}^N$. The nonlinear constraint is monotonically decreasing in ξ, and as a result there is a one-to-one correspondence between the specified noise level σ_e^2 (in the appropriate range) and the Lagrange multiplier ξ. A zero finding algorithm (e.g. a bisection algorithm) can for example be used for finding the Lagrange multiplier corresponding to the given noise variance level.

Combining the described noise variance estimator with the LS-SVM regression based on Morozov's discrepancy principle leads to the following procedure for selecting an appropriate regularization trade-off.

Algorithm 1 (*Morozov's model selection*). For regularization parameter tuning:

(1) Estimate the noise level $\hat{\sigma}_e^2$ using the method of the differogram via (18), (19) or the robust version (30);
(2) For an initial choice of ξ, determine the corresponding noise level denoted as σ_ξ^2 by training the model via (34); and

(3) Use a zero-finding algorithm to find the unique value ξ^* by taking σ_ξ^2 equal to $\hat{\sigma}_e^2$ according to (34).

The advantage of this model selection criterion over classical data-driven methods as cross-validation is that the algorithm converges globally by solving a few linear systems (usually in 5 to 10 steps). However, a drawback for this simplicity is often a decrease in generalization performance. Hence, it suggests the use of the obtained regularization constant as a good starting value for local search of a more powerful data-driven method [5,6,33].

5.2. Other applications

A model-free estimate of the noise variance plays an important role for making model selection and setting tuning parameters. Examples of such applications are given:

(1) Well-known complexity criteria (or model selection criteria) such as the Akaike Information Criterion [1], the Bayesian Information Criterion [31] and C_p statistic [23] take the form of a prediction error criterion which consists of the sum of a training set error (e.g. the residual sum of squares) and a complexity term. In general:

$$J(S) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i;S))^2 + \lambda(Q_N(\hat{f}))\hat{\sigma}_e^2, \tag{38}$$

where S denotes the smoother matrix; see [8]. The complexity term $Q_N(\hat{f})$ represents a penalty term which grows proportionally with the number of free parameters (in the linear case) or the effective number of parameters (in the nonlinear case [38,33]) of the model $\hat{f}$ grows.

(2) Consider the linear ridge regression model $y = w^{\mathrm{T}}x + b$ with w and b optimized w.r.t.

$$J_{RR,\gamma}(w,b) = \frac{1}{2}w^{\mathrm{T}}w + \frac{\gamma}{2}\sum_{i=1}^{N}(y_i - w^{\mathrm{T}}x_i - b)^2. \tag{39}$$

Using the Bayesian interpretation [22,35] of ridge regression and under i.i.d. Gaussian assumptions, the posterior can be written as $p(w,b\,|\,x_i,y_i,\mu,\zeta) \propto \exp(-\zeta(wx_i+b-y_i)^2)\exp(-\mu(w^{\mathrm{T}}w))$, the estimate of the noise variance $\zeta = 1/\hat{\sigma}_e^2$ and the expected variance of the first derivative $\mu = 1/\sigma_w^2$ can be used to set, respectively, the expected variance of the likelihood $p(y_i\,|\,x_i,w,b)$ and on the prior $p(w,b)$. As such, a good guess for the regularization constant when the input variables are independent becomes $\hat{\gamma} = \hat{a}_1^2/\hat{\sigma}_e^2$.
Another proposed guess for the regularization constant $\hat{\gamma}$ in ridge regression (39) can be derived as in [19]: $\hat{\gamma} = \hat{w}_{LS}^{\mathrm{T}}\hat{w}_{LS}/(\hat{\sigma}_e d)$ where $\hat{\sigma}_e$ is the estimated variance of the noise, d is the number of free parameters and $\hat{w}_{LS}$ are the estimated parameters of the ordinary least squares problem. These guesses can also be used

to set the regularization constant in the parametric step in fixed-size LS-SVMs [33] where the estimation is done in the primal space instead of the dual via a Nÿstrom approximation of the feature map.

(3) Given the non-parametric Nadaraya–Watson estimator $\hat{f}(x) = [\sum_{i=1}^{N} (K((x - x_i)/h)y_i)]/[\sum_{i=1}^{N} K((x - x_i)/h)]$, the plugin estimator for the bandwidth h is calculated under the assumption that a Gaussian kernel is to be used and the noise is Gaussian. The derived plugin estimator becomes $h_{\text{opt}} = C\hat{\sigma}^2 N^{-\frac{1}{5}}$ where $C \approx 6\sqrt{\pi}/25$; see e.g. [15].

(4) We note that $\hat{\sigma}_e^2$ also plays an important role in setting the tuning parameters of SVMs; see e.g. [36,6].

6. Experiments

Figs. 1 and 2 show artificially generated data and the differogram cloud of a linear and nonlinear toy data set, respectively. The latter example was taken from [38] using the function $f(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x})$ such that $y_i = f(x_i) + e_i$ for all $i = 1, \ldots, N$. The random white noise terms are distributed as $\mathcal{N}(0, 0.1)$. The differogram cloud in (1) of the one-dimensional linear case has a reasonably good fit as explained in Section 3, while the differogram cloud of the nonlinear function (2b) is reasonably good (solid line) close to $\Delta_x^2 = 0$ which is quantified by the decreasing weighting function (dashed line).

6.1. Noise variance estimation and model selection

To randomize the experiments, the following class of underlying functions is considered:

$$f(\cdot) = \sum_{i=1}^{N} \bar{\alpha}_i K(x_i, \cdot), \tag{40}$$

where $\bar{\alpha}_i$ is an i.i.d. sequence of uniformly distributed terms. The kernel is fixed as $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2)$ for any $i,j = 1, \ldots, N$. Data-points were generated as $y_i = f(x_i) + e_i$ for $i = 1, \ldots, N$ where e_i are N i.i.d. samples.

The first example illustrates the behavior of the estimator for an increasing number of given data-points. At first, a Gaussian noise distribution F of unit variance is generated. Results of the proposed model-free noise estimator on a Monte–Carlo simulation study of 500 iterations for data sets of N ranging from 10 to 1000 are shown in Fig. 3a. Fig. 3b reports numerical results when working in the case of contaminated data ($F = 0.95\mathcal{N}(0, 1) + 0.05\mathcal{N}(0, 10)$) instead. While the non-robust method tends to consider the complete data set as noise (the reported noise level is close to the variance of the original signal), the robustified method converges faster to the true noise level of the nominal noise model $\mathcal{N}(0, 1)$.

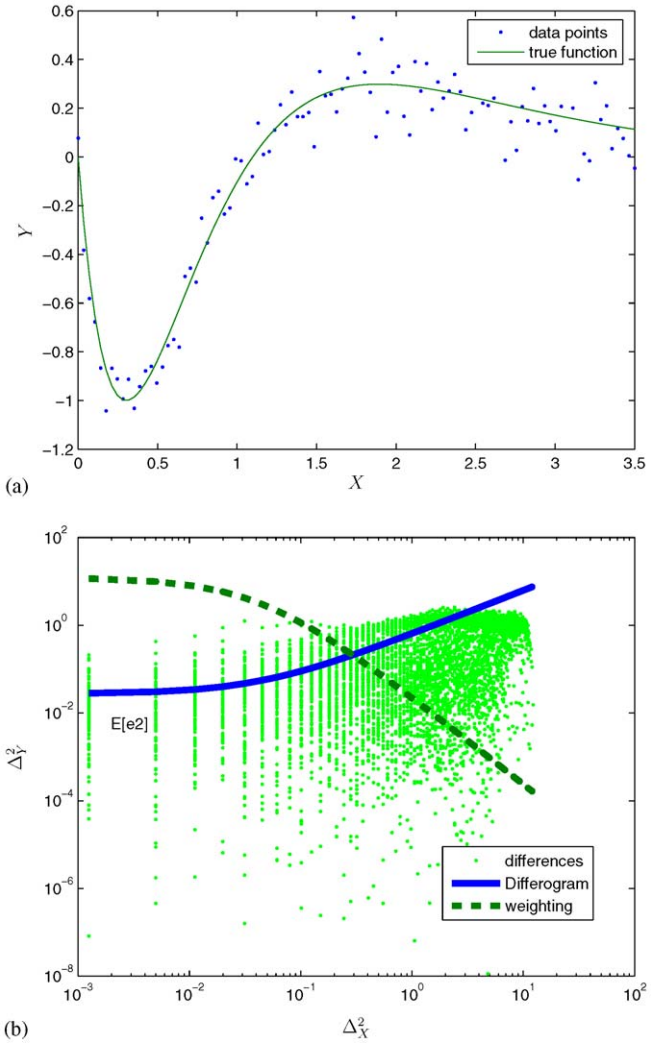


Fig. 2. Differogram of a nonlinear function: (a) data are generated according to the nonlinear data set described in [38] with a noise standard deviation of 0.1 and $N = 100$ and (b) differogram cloud of all differences $\Delta_{x,ij}^2 = \|x_i - x_j\|_2^2$ and $\Delta_{y,ij}^2 = \|y_i - y_j\|_2^2$ for all $i,j = 1,\ldots,N$. The solid line represents the estimated differogram $\hat{\Upsilon}(\Delta_x^2)$ and the dashed line denotes the corresponding weighting function $1/\vartheta(\Delta_x^2)$. The estimate of the noise variance is $\hat{\sigma}_e^2 = \hat{\Upsilon}(0) = 0.1086$.

The second experiment investigates the accuracy of regularization constant tuning using Morozov's discrepancy principle. Again functions were generated by (40). Note that now the fixed kernel makes it possible to draw conclusions based on the experiments only regarding the regularization parameter as kernel-parameter tuning is no longer involved. The experiment compares different model

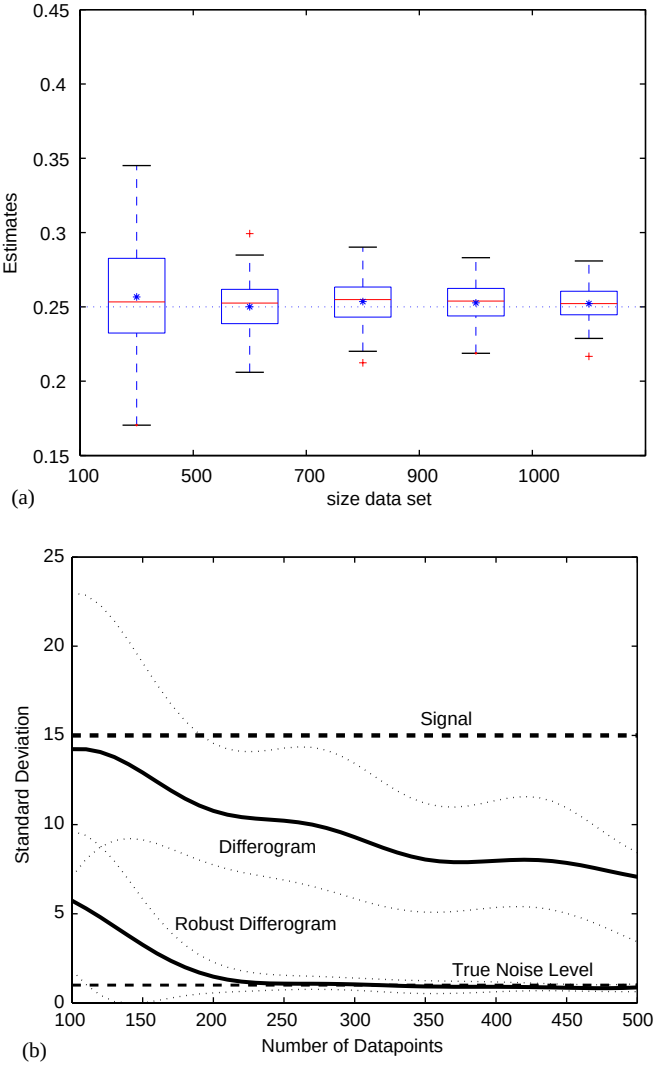


Fig. 3. Accuracy of the differogram estimator when increasing the number of data-points (a) when data were generated from (40) using a Gaussian noise model and (b) using a contaminated noise model, using, respectively, the standard and the robustified estimator.

selection strategies based on respectively, exact prior knowledge of the noise level, a model-free estimate using the differogram and some standard data-driven methods such as L-fold cross-validation, leave-one-out, Mallows C_p statistic [23,8] and Bayesian inference [33]. An important remark is that the method based on the differogram is orders of magnitudes faster than any data-driven method and can therefore be used for picking good starting-values for a local search based on

a more powerful and computationally intensive way to achieve a good generalization performance. The Boston Housing data set [3] concerning the housing values in suburbs of Boston was used to benchmark the proposed method on a real-world data set. This set contains 506 instances of 12 continuous and 1 binary input variables and one dependent variable. A Monte–Carlo experiment of 500 runs (with standardized inputs and outputs) suggests that the proposed measure can be sufficiently good as a model selection criterion on its own. For this experiment, one-third of the data was reserved for test purposes, while the remaining data were used for the training and selection of the regularization parameter.

6.2. The Lynx time-series

This example illustrates how to use the spatio-temporal extension of the differogram in the context of noisy time-series. This classical data set was described and benchmarked e.g. in [34]. In order to motivate the use of the spatio-temporal differogram in this case, all 114 samples were used for the purpose of recovering the amount of noise in the data and related to the level reported in [34]. At first, the observed data were pre-processed using a $\log_{10}$ transformation as classical, resulting in the given data $\{y_t\}_{t=1}^{114}$ (see Fig. 4a). The Lynx data set can be ordered in time (temporal) as well as based on its first lagged variable (spatial) which is often used as a threshold variable in a Threshold Auto-Regressive (TAR) model [34] defined as

$$\hat{y}_t = \begin{cases} b_1 + \sum_{l=1}^{q} w_{1,l} y_{t-l} & \text{if } y_{t-1} \leqslant t_1, \\ b_2 + \sum_{l=1}^{q} w_{2,l} y_{t-l} & \text{if } t_1 < y_{t-1} \leqslant t_1, \\ \dots & \\ b_r + \sum_{l=1}^{q} w_{r,l} y_{t-l} & \text{if } t_{r-1} < y_{t-1}, \end{cases} \tag{41}$$

where t_i for $i = 1, \dots, r-1$ are the discrete thresholds for the first lagged variable and $q \in \mathbb{N}_+$ is the order of the linear models. These nonlinear models are characterized by different (linear) operating regimes where the threshold variable makes the distinction between regions. The differogram then becomes:

$$\Upsilon^{\mathrm{st}}(\Delta^2_{x,ij}, \Delta_{t,ij}) = \tfrac{1}{2} E[\Delta^2_Y \mid \Delta^2_X = \|y_{i-1} - y_{j-1}\|_2^2,\ \Delta_{t,ij} = |t_i - t_j|]. \tag{42}$$

As motivated by earlier applications [34], the lag of this model was fixed to the value one. Using only the temporal order to build a differogram gives an estimate of the noise with standard deviation 0.1900, while using the spatial ordering alone results in a noise level equal to 0.3940. The spatial and the temporal differogram are visualized, respectively, in Fig. 4a and b. A combination of both orderings can be done by using the spatio-temporal differogram as described in Section 4.1. The spatio-temporal model indicates a noise standard deviation equal to 0.12 which is closer to the reported level [34].

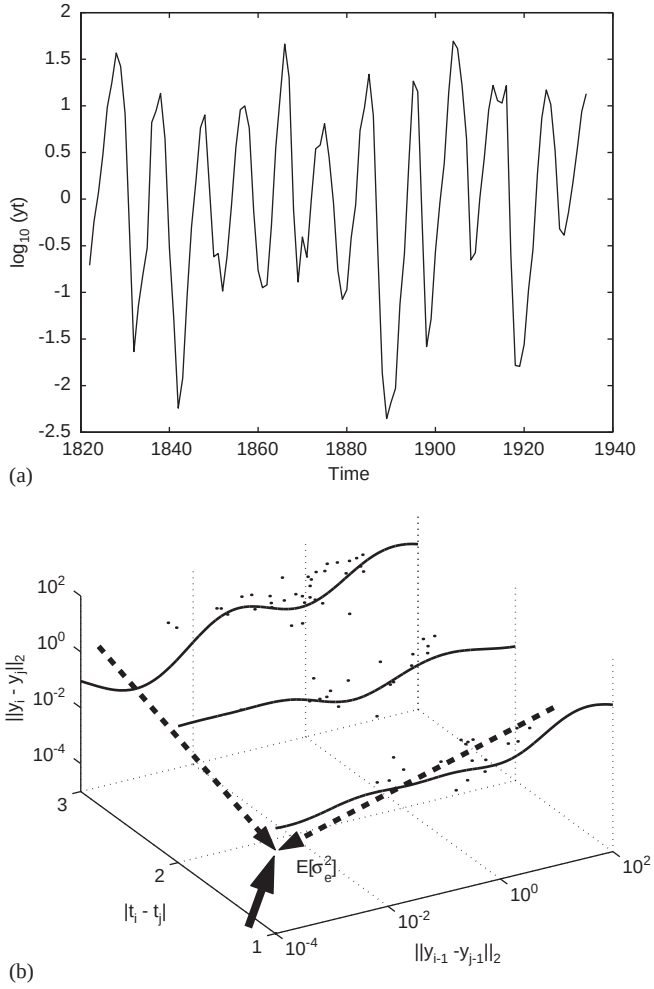


Fig. 4. (a) Log transformation of the Lynx data and (b) the spatio-temporal differogram cloud displays the data as a function of their distance in time and space. In this autonomous case, a good candidate for the spatial indicator is the first lagged observation as suggested by the use of TAR models which have regimes conditionally defined on their first lagged variable.

7. Conclusions

This paper proposed the use of a non-parametric data analysis tool for noise variance estimation in the context of machine learning. By modeling the variation in the data for observations that are located close to each other, properties of the data can be extracted without relying explicitly on an estimated model of the data. These ideas are translated by considering the differences of the data instead of the data itself in the so-called differogram cloud. As only the behavior of the differogram

function towards the origin is of concern here, a model for the differogram function can be motivated without the need for extra hyper-parameters such as the bandwidth. Extensions towards a spatio-temporal setting and an approach for contaminated data have been given. Furthermore, an application is added where the estimated noise level results in a fast determination of the regularization trade-off for LS-SVMs based on Morozov's discrepancy principle. Finally, a number of applications of model-free noise variance estimators for model selection and hyper-parameter tuning have been discussed.

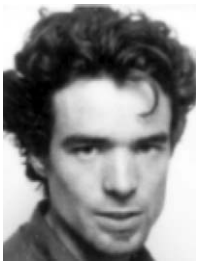
Acknowledgements

This research work was carried out at the ESAT laboratory of the KUL. Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666, GOA-Ambiorics, IDO, several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, FWO projects G.0407.02, G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0499.04, G.0211.05, G.0080.01, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s, STWW-Genprom, GBOU-McKnow, Eureka-Impact, Eureka-FLiTE, several PhD grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996–2001) and IUAP V-10-29 (2002–2006)), Program Sustainable Development PODO-II (CP/40); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS is an associate professor and BDM is a full professor at K.U.Leuven Belgium, respectively.

References

[1] H. Akaike, Statistical predictor identification, Ann. Inst. Statist. Math. 22 (1973) 203–217.
[2] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.
[3] C.L. Blake, C.J. Merz, UCI repository of machine learning databases, 1998.
[4] M.J. Buckley, G.K. Eagleson, The estimation of residual variance in nonparametric regression, Biometrika 75 (2) (1988) 189–199.
[5] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Machine Learning 46 (1–3) (2002) 131–159.
[6] V. Cherkassky, Ma. Yunqian, Practical selection of SVM parameters and noise estimation for svm regression, Neural Networks 17 (2004) 113–126.
[7] N.A.C. Cressie, Statistics for Spatial Data, Wiley, New York, 1993.
[8] J. De Brabanter, K. Pelckmans, J.A.K. Suykens, J. Vandewalle, B. De Moor, Robust cross-validation score function for non-linear function estimation, in: International Conference on Artificial Neural Networks, Madrid, Spain, 2002, pp. 713–719.
[9] L. Devroye, L. Györfi, D. Schäfer, H. Walk, The estimation problem of minimum mean squared error, Statistics and Decisions 21 (2003) 15–28.
[10] P.J. Diggle, Time-series: A Biostatistical Introduction, Oxford University Press, Oxford, 1990.
[11] R.L. Eubank, Nonparametric Regression and Spline Smoothing, Marcel Dekker, New York, 1999.
[12] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (1) (2000) 1–50.
[13] T. Gasser, L. Sroka, C. Jennen-Steinmetz, Residual variance and residual pattern in nonlinear regression, Biometrika 73 (1986) 625–633.

[14] G.H. Golub, C.F. Van Loan, Matrix Computations, The John Hopkins University Press, Baltimore, MD, 1989.
[15] W. Härdle, Applied Nonparametric Regression, Econometric Society Monographs, Cambridge University Press, Cambridge, 1989.
[16] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, Heidelberg, 2001.
[17] T.P. Hettmansperger, J.W. McKean, Robust Nonparametric Statistical Methods, Kendall's Library of Statistics, vol. 5, Arnold, 1994.
[18] W. Hoeffding, A class of statistics with asymptotically normal distribution, Ann. Math. Stat. 19 (1948) 293–325.
[19] A.E. Hoerl, R.W. Kennard, K.F. Baldwin, Ridge regression: some simulations, Commun. Stat. Part A - Theory and Methods 4 (1975) 105–123.
[20] P.J. Huber, Robust estimation of a location parameter, Ann. Math. Statist. 35 (1964) 73–101.
[21] A.J. Lee, U-Statistics, Theory and Practice, Marcel Dekker, New York, 1990.
[22] D.J.C. MacKay, The evidence framework applied to classification networks, Neural Comput. 4 (1992) 698–714.
[23] C.L. Mallows, Some comments on Cp, Technometrics 15 (1973) 661–675.
[24] A.M. Mood, F.A. Graybill, D.C. Boes, Introduction to the theory of statistics, Series in Probability and Statistics, McGraw-Hill, 1963.
[25] V.A. Morozov, Methods for Solving Incorrectly Posed Problems, Springer, Heidelberg, 1984.
[26] U.U. Müller, A. Schick, W. Wefelmeyer, Estimating the error variance in nonparametric regression by a covariate-matched U-statistic, Statistics 37 (3) (2003) 179–188.
[27] A. Neumaier, Solving ill-conditioned and singular linear systems: a tutorial on regularization, SIAM Review 40 (3) (1998) 636–666.
[28] T. Poggio, F. Girosi, Networks for approximation and learning, Proc. IEEE 78 (9) (1990) 1481–1497.
[29] J. Rice, Bandwidth choice for nonparametric regression, Ann. Stat. (12) (1984) 1215–1230.
[30] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.
[31] G. Schwartz, Estimating the dimension of a model, Ann. Stat. 6 (1979) 461–464.
[32] A. Sen, M. Srivastava, Regression Analysis, Theory, Methods and Applications, Springer, Heidelberg, 1997.
[33] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, 2002.
[34] H. Tong, Nonlinear Time Series Analysis: A Dynamic Approach, Oxford University Press, Oxford, 1990.
[35] T. Van Gestel, J.A.K. Suykens, D. Baestaens, A. Lambrechts, G. Lanckriet, B. Vandaele, B. De Moor, J. Vandewalle, Financial time series prediction using least squares support vector machines within the evidence framework, IEEE Trans. Neural Networks (special issue on Neural Networks in Financial Engineering) 12 (4) (2001) 809–821.
[36] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[37] J. von Neumann, I.N. Schoenberg, Fourier integrals and metric geometry, Trans. Am. Math. Soc. (50) (1941) 226–251.
[38] G. Wahba, Spline Models for Observational Data, SIAM, 1990.



Kristiaan Pelckmans was born at 3 November 1978 in Merksplas, Belgium. He received a M.Sc. degree in Computer Science in 2000 from the Katholieke Universiteit Leuven. After a projectwork for an implementation of kernel machines and LS-SVMs (LS-SVMlab), he currently pursues a PhD at the KULeuven in the faculty of Applied Sciences, department of Electrical Engineering in the SCD/SISTA laboratory. His research mainly focuses on machine learning and statistical inference using primal-dual kernel machines.



Jos De Brabanter was born in Ninove Belgium, January 11 1957. He received the degree in Electronical Engineering (Industriële hogeschool van het Rijk Brabant) in 1990, Safety Engineer (Vrije Universiteit Brussel) in 1992, Master of Environment, Human Ecology (Vrije Universiteit Brussel) in 1993, Master in Artificial Intelligence (Katholieke Universiteit Leuven) in 1996, Master of Statistics (Katholieke Universiteit Leuven) in 1997 and the Ph.D. degree in Applied Sciences (Katholieke Universiteit Leuven) in 2004. His research mainly focuses on Statistics and nonlinear systems.



Johan A.K. Suykens was born in Willebroek Belgium, May 18 1966. He received the degree in Electro-Mechanical Engineering and the Ph.D. degree in Applied Sciences from the Katholieke Universiteit Leuven, in 1989 and 1995, respectively. In 1996 he has been a Visiting Postdoctoral Researcher at the University of California, Berkeley. He has been a Postdoctoral Researcher with the Fund for Scientific Research FWO Flanders and is currently an Associate Professor with K.U.Leuven. His research interests are mainly in the areas of the theory and application of neural networks and nonlinear systems. He is author of the books "Artificial Neural Networks for Modelling and Control of Non-linear Systems" (Kluwer Academic Publishers) and "Least Squares Support Vector Machines" (World Scientific) and editor of the books "Nonlinear Modeling: Advanced Black-Box Techniques" (Kluwer Academic Publishers) and "Advances in Learning Theory: Methods, Models and Applications" (IOS Press). In 1998 he organized an International Workshop on Nonlinear Modelling with Time-series Prediction Competition. He is a Senior IEEE member and has served as associate editor for the IEEE Transactions on Circuits and Systems- Part I (1997–1999) and Part II (since 2004) and since 1998 he is serving as associate editor for the IEEE Transactions on Neural Networks. He received an IEEE Signal Processing Society 1999 Best Paper (Senior) Award and several Best Paper Awards at International Conferences. He is a recipient of the International Neural Networks Society INNS 2000 Young Investigator Award for significant contributions in the field of neural networks. He has served as Director and Organizer of a NATO Advanced Study Institute on Learning Theory and Practice (Leuven 2002) and as a program co-chair for the International Joint Conference on Neural Networks IJCNN 2004.



Bart De Moor was born Tuesday July 12, 1960 in Halle, Belgium. He is married and has three children. In 1983, he obtained his Master (Engineering) Degree in Electrical Engineering at the Katholieke Universiteit Leuven, Belgium, and a PhD in Engineering at the same university in 1988. He spent 2 years as a Visiting Research Associate at Stanford University (1988–1990) at the departments of EE (ISL, Prof. Kailath) and CS (Prof.Golub). Currently, he is a full professor at the Department of Electrical Engineering (http://www.esat.kuleuven.ac.be) of the K.U.Leuven. His research interests are in numerical linear algebra and optimization, system theory and identification, quantum information theory, control theory, data-mining, information retrieval and bio-informatics, areas in which he has (co)authored several books and hundreds of research papers (consult the publication search engine at http://www.esat.kuleuven.ac.be/sista-cosic-docarch/template.php). Currently, he is leading a research group of 39 PhD students and 8 postdocs and in the recent past, 16 PhDs were obtained under his guidance. He has been teaching at and been a member of PhD jury's in several universities in Europe and the US. He is also a member of several scientific and professional organizations. His work has won him several scientific awards (Leybold-Heraeus Prize (1986), Leslie Fox Prize (1989), Guillemin-Cauer best paper Award of the IEEE Transaction on Circuits and Systems (1990),

Laureate of the Belgian Royal Academy of Sciences (1992), bi-annual Siemens Award (1994), best paper award of Automatica (IFAC, 1996), IEEE Signal Processing Society Best Paper Award (1999). Since 2004 he is a fellow of the IEEE (www.ieee.org). He is an associate editor of several scientific journals. From 1991–1999 he was the chief advisor on Science and Technology of several ministers of the Belgian Federal Government (Demeester, Martens) and the Flanders Regional Governments (Demeester, Van den Brande). He was and/or is in the board of 3 spin-off companies (www.ipcos.be, www.data4s.com, www.tml.be), of the Flemish Interuniversity Institute for Biotechnology (www.vib.be), the Study Center for Nuclear Energy (www.sck.be) and several other scientific and cultural organizations. He was a member of the Academic Council of the Katholieke Universiteit Leuven, and still is a member of its Research Policy Council.