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

Phone: +32-16-32 85 40, Fax: +32-16-32 19 70

E-mail: {kristiaan.pelckmans, johan.suykens}@esat.kuleuven.ac.be

Web: http://www.esat.kuleuven.ac.be/sista/lssvmlab/

[b] *Hogeschool KaHo Sint-Lieven (Associatie KULeuven), Departement Industrieel Ingenieur*

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 here the *differogram cloud* - 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 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.

Key words: Noise Level, Complexity criteria, Kernel Methods, Least Squares Support Vector Machines, Morozov's discrepancy method

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 (Bishop, 1995; Vapnik, 1998; Suykens *et al.*, 2002), 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 (Akaike, 1973) and the C_p-statistic (Mallows, 1973). 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) (Vapnik, 1998; Schölkopf and Smola, 2002; Cherkassky and Ma, 2004; Suykens *et al.*, 2002), regularization networks (Poggio and Girosi, 1990) and Nadaraya-Watson kernel estimators, see e.g. (Hastie *et al.*, 2001). Additional relations between the noise variance and the amount of smoothing required are described in the Bayesian evidence framework and Gaussian Processes (MacKay, 1992), statistical learning theory (Cherkassky and Ma, 2004; Evgeniou *et al.*, 2000), splines (Wahba, 1990) and regularization theory, e.g. in the form of Morozov's discrepancy principle (Morozov, 1984; Neumaier, 1998).

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 (von Neumann

and Schoenberg, 1941). In the context of geographically distributed measurements (Cressie, 1993), 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 (Devroye *et al.*, 2003).

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. (Wahba, 1990). A broad class of model based variance estimators can be written as

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

with $y = (y_1, \ldots, y_N)^T$ (Buckley and Eagleson, 1988), $\mathrm{tr}(\cdot)$ 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 $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_N)^T \in \mathbb{R}^N$. 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 (Eubank, 1999) or Least Squares Support Vector Machines (LS-SVMs) (Suykens *et al.*, 2002).

2.2 Model free estimators

Model-free variance estimators were proposed in the case of equidistantly ordered data. In the work of (Rice, 1984) and (Gasser *et al.*, 1986), 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.* (1986) 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 $E\left[c_i^2 \hat{\varepsilon}_i^2\right] = \sigma^2$ 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

$$\mathcal{F} = \left\{ F : \int x dF(x) = 0,\ 0 < \int |x|^r \, dF(x) < \infty \right\},\ r \in \mathbb{N}_0 \text{ and } 1 \leq r \leq 4. \tag{4}$$

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 (Müller *et al.*, 2003) suggested an error variance estimator given by

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

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

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

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, 1993) *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*

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

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 (Diggle, 1990; Cressie, 1993). 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 (Cressie, 1993). Estimation of the parameters of a variogram model often employs a maximum likelihood criterion (Cressie, 1993) 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 $\Delta^2_{x,ij} = \|x_i - x_j\|_2^2$ for all $i, j = 1, \ldots, N$ be samples of the random variable Δ^2_X and let $\Delta^2_{y,ij} = \|y_i - y_j\|_2^2$ be samples from the random variable Δ^2_Y. The differogram function $\Upsilon : \mathbb{R}^+ \to \mathbb{R}^+$ is defined as*

$$\Upsilon\left(\Delta^2_x\right) = \frac{1}{2}E[\Delta^2_Y \big| \Delta^2_X = \Delta^2_x]. \tag{8}$$

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 (Mood *et al.*, 1963). 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 (Hastie *et al.*, 2001; Devroye *et al.*, 2003).

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 form 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 Figure 1. Figure 2 presents the differogram cloud and the (estimated) differogram function of a non-linear regression, while Section 6 reports on some experiments on higher dimensional data.

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

Lemma 1 *Assume a Lipschitz continuous function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\exists M \in \mathbb{R}^+$ where $\|f(X) - f(X')\|_2^2 \leq M \|X - X'\|_2^2$ with X' 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

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

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 (Mood *et al.*, 1963). 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 (Devroye *et al.*, 2003) where the nearest neighbor paradigm replaces the conditioning on Δ_X^2 and fast rates of convergence were proved.

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 $l \geq 2$. One may motivate the use of an r-th order Taylor series approximation of the differogram function with center $m = 0$ as a suitable model because one is only interested in the case $\Delta_x^2 \to 0$:

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

where the parameter vector $a = (a_0, a_1, \ldots, a_r)^T \in \mathbb{R}^{+,r+1}$ 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

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

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 upperbound 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

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

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. (Sen and Srivastava, 1997)). As the parameter vector a is positive, the weighting function is monotonically decreasing and as such represents always 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, 1948) *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*

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

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

It is shown (Lee, 1990) 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 $l = 2$

$$\hat{\sigma}_e^2 = U(g; e_1, \ldots, e_N) = \frac{2}{N(N-1)} \sum_{1 \leq i \leq j \leq N} g_1(e_i, e_j)$$
$$\text{and} \quad g_1(e_i, e_j) = \frac{1}{2}(e_i - e_j)^2 = \frac{1}{2}\Delta_{e,ij}^2. \qquad (15)$$

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 (Vapnik, 1998) or the global correlation structure (Cressie, 1993). 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

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

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: $2c(\sum_{i \le j}^N 1/\bar{\vartheta}(\Delta_{x,ij}^2)) = N(N-1)$.

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

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

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

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

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. (Hettmansperger and McKean, 1994). As previously, one can write

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

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

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

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

$$2\hat{\sigma}_e^2 = \arg\min_{a_0 \geq 0} \sum_{k=1}^{N(N-1)/2} \frac{c}{\bar{\vartheta}(\Delta_{x,k}^2)} \left(\Delta_{y,k}^2 - \mathcal{A}(\Delta_{x,k}^2) - a_0\right)^2$$
$$= \arg\min_{a_0 \geq 0} \sum_{i<j} \frac{c}{\bar{\vartheta}(\Delta_{x,ij}^2)} \left(\Delta_{y,ij}^2 - \mathcal{A}(\Delta_{x,ij}^2) - a_0\right)^2. \quad (22)$$

In both cases, the function $\mathcal{A} : \mathbb{R}^+ \to \mathbb{R}^+$ 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. (Cressie, 1993). Both are considered as random variables X and T. The observations are generated as

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

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 is assumed to be independent (Hastie *et al.*, 2001). 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

$$\Upsilon^{st}\left(\Delta_x^2,\Delta_t\right)=\frac{1}{2}E[\Delta_Y^2\ \Big|\ \Delta_X^2=\Delta_x^2,\ \ \Delta_T=\Delta_t]. \tag{24}$$

As in the previous Section, we consider the differogram models

$$\begin{aligned}\Upsilon^{st}_{\mathcal{A},\mathcal{B}}\left(\Delta_x^2,\Delta_t\right)=a_0+\mathcal{A}(\Delta_x^2)+\mathcal{B}(\Delta_t),\ \ \mathcal{A}(\Delta_x^2)=\sum_{i=1}^{r_1}a_i(\Delta_x^2)^i,\\ \mathcal{B}(\Delta_t)=\sum_{j=1}^{r_2}b_j(\Delta_t)^j\quad\text{where}\quad a_0,\ldots,a_{r_1},b_1,\ldots,b_{r_2}\in\mathbb{R}^+,\end{aligned} \tag{25}$$

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

$$\begin{aligned}(a^*,b^*)&=\arg\min_{0\leq a\in\mathbb{R}^{r_1},0\leq b\in\mathbb{R}^{r_2}}\mathcal{J}^{st}(a,b)\\&=\sum_{i\leq j}\frac{c}{\bar{\vartheta}^2_{\mathrm{st}}(\Delta^2_{x,ij},\Delta_{t,ij})}\left(\Delta^2_{y,ij}-\Upsilon^{st}_{\mathcal{A},\mathcal{B}}(\Delta^2_{x,ij},\Delta_{t,ij};a,b)\right)^2,\end{aligned} \tag{26}$$

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

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

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

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

A common view on robustness is to provide alternatives to least squares methods when deviating from the assumptions of Gaussian distributions as in Subsection 3.1. This part considers the extension of the differogram towards contaminated noise models (Huber, 1964), 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 $\mathcal{H} : \mathbb{R} \rightarrow \mathbb{R}^+$ given as

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

with $\beta \in \mathbb{R}_0^+$. Note that the weighting function need not to 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)

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

where c is a normalizing constant for the weighting function. If $\bar{\vartheta}(\Delta^2_{x,ij}; a)$ were known, the corresponding convex optimization problem can be solved efficiently as noted e.g. in (Vapnik, 1998). In the case of large scale datasets, 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 least squares support vector machine (LS-SVM) regressor as described in (Suykens *et al.*, 2002). A closely related derivation can be made based on Morozov's discrepancy principle (Morozov, 1984; Neumaier, 1998). Let again $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ be samples dawn 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 $f(x) = w^T \varphi(x) + b$ where $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ denotes the potentially infinite ($n_h = \infty$) dimensional feature map.

Morozov's discrepancy principle (Morozov, 1984) 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 (Suykens *et al.*, 2002), it is given by

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

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

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

The conditions for optimality are

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

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

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

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. (Chapelle *et al.*, 2002; Schölkopf and Smola, 2002). Typical examples are the use of a linear kernel $K(x_i, x_j) = x_i^T x_j$ 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) (Golub and Loan, 1989) of Ω is given as

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

where $U \in \mathbb{R}^{N \times N}$ is an orthonormal matrix $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_N)$ and $\gamma_i \geq 0$ 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, the conditions (34)

can then be rewritten as

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

where $p = U^T y \in \mathbb{R}^N$. 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 with 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);

(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, this suggests to use the obtained regularization constant as a good starting value for local search of a more powerful

data-driven method (Chapelle *et al.*, 2002; Cherkassky and Ma, 2004; Suykens *et al.*, 2002).

5.2 Other applications

A model-free estimate of the noise variance plays an important role for doing 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 (Akaike, 1973), the Bayesian Information Criterion (Schwartz, 1979) and C_p statistic (Mallows, 1973) 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:

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

where S denotes the smoother matrix, see (De Brabanter *et al.*, 2002). 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 (Wahba, 1990; Suykens *et al.*, 2002)) of the model $\hat{f}$ grows.

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

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

Using the Bayesian interpretation (MacKay, 1992; Van Gestel *et al.*, 2001) of ridge regression and under i.i.d. Gaussian assumptions, the posterior can be written as $p(w,b \mid x_i,y_i,\mu,\zeta) \propto \exp(-\zeta(wx_i+b-y_i)^2)\exp(-\mu(w^Tw))$, 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 (Hoerl *et al.*, 1975): $\hat{\gamma} = \hat{w}_{LS}^T \hat{w}_{LS}/(\hat{\sigma}_e d)$ 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 (Suykens *et al.*, 2002) 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 $h_{opt} = C\hat{\sigma}^2 N^{-\frac{1}{5}}$ where $C \approx 6\sqrt{\pi}/25$, see e.g. (Härdle, 1989).

(4) We note that $\hat{\sigma}_e^2$ also plays an important role in setting the tuning parameters of SVMs, see e.g. (Vapnik, 1998; Cherkassky and Ma, 2004).

6 Experiments

Figure 1 and 2 show artificially generated data and the differogram cloud of a linear and nonlinear toy dataset respectively. The latter example was taken from (Wahba, 1990) 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 (2.b) 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 datasets of N ranging from 10 to 1000 are shown in Figure 3.a. Figure 3.b 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 dataset 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 $\mathcal{N}(0, 1)$.

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 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 (Mallows, 1973; De Brabanter *et al.*, 2002) and Bayesian inference (Suykens *et al.*, 2002). An important remark is that the method based on the differogram is orders of magnitudes faster than any data-driven method and can therefor 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 dataset (Blake and Merz, 1998) concerning the housing values in suburbs of Boston was used to benchmark the proposed method on a real world dataset. 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 a context of noisy time-series. This classical dataset was described and benchmarked e.g. in (Tong, 1990). 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 (Tong, 1990). 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 Figure 4.a). The Lynx dataset 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 (Tong, 1990), defined as

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

where t_i for $i = 1, \ldots, r-1$ 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 becomes then:

$$\Upsilon^{\text{st}}(\Delta^2_{x,ij}, \Delta_{t,ij}) = \frac{1}{2} E\left[\Delta^2_Y \;\middle|\; \Delta^2_X = \|y_{i-1} - y_{j-1}\|_2^2,\ \ \Delta_{t,ij} = |t_i - t_j|\right]. \tag{42}$$

As motivated by earlier applications (Tong, 1990), 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 Figure 4.a and Figure 4.b. Combination of both orderings can be done by using the spatio-temporal differogram as described in Subsection 4.1. The spatio-temporal model indicates a noise standard deviation equal to 0.12 which is closer to the reported level (Tong, 1990).

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.

Acknowledgments. 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), (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

Akaike, H. (1973). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–217.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Blake, C.L. and C.J. Merz (1998). UCI repository of machine learning databases.

Buckley, M.J. and G.K. Eagleson (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75**(2), 189–199.

Chapelle, O., V. Vapnik, O. Bousquet and S. Mukherjee (2002). Choosing multiple parameters for support vector machines. *Machine Learning* **46**(1-3), 131–159.

Cherkassky, V. and Yunqian Ma (2004). Practical selection of SVM parameters and noise estimation for svm regression. *Neural Networks* **17**, 113–126.

Cressie, N.A.C. (1993). *Statistics for spatial data*. Wiley.

De Brabanter, J., K. Pelckmans, J.A.K. Suykens, J. Vandewalle and B. De Moor (2002). Robust cross-validation score function for non-linear function estimation. In: *International Conference on Artificial Neural Networks*. Madrid, Spain. pp. 713–719.

Devroye, L., L. Györfi, D. Schäfer and H. Walk (2003). The estimation problem of minimum mean squared error. *Statistics and Decisions* **21**, 15–28.

Diggle, P.J. (1990). *Time-series: A biostatistical introduction*. Oxford University Press, Oxford.

Eubank, R.L. (1999). *Nonparameteric Regression and Spline Smoothing*. Marcel Dekker, Inc. New York.

Evgeniou, T., M. Pontil and T. Poggio (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics* **13**(1), 1–50.

Gasser, T., L. Sroka and C. Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

Golub, G.H. and C.F. Van Loan (1989). *Matrix Computations*. The John Hopkins University Press.

Härdle, W. (1989). *Applied Nonparametric Regression*. Econometric Society Monographs. Cambridge University Press.

Hastie, T., R. Tibshirani and J. Friedman (2001). *The Elements of Statistical Learning*. Springer-Verlag. Heidelberg.

Hettmansperger, T.P. and J.W. McKean (1994). *Robust Nonparametric Statistical Methods*. Vol. 5 of *Kendall's Library of Statistics*. Arnold.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**, 293–325.

Hoerl, A.E., R.W. Kennard and K.F. Baldwin (1975). Ridge regression: Some simulations. *Communications in Statistics, Part A - Theory and Methods* **4**, 105– 123.

Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Lee, A.J. (1990). *U-Statistics, Theory and Practice*. Marcel Dekker, New York.

MacKay, D.J.C. (1992). The evidence framework applied to classification networks. *Neural Computation* **4**, 698–714.

Mallows, C.L. (1973). Some comments on Cp. *Technometrics* **15**, 661–675.

Mood, A.M., F.A. Graybill and D.C. Boes (1963). *Introduction to the Theory of Statistics*. Series in Probability and Statistics. McGraw-Hill.

Morozov, V.A. (1984). *Methods for Solving Incorrectly Posed Problems*. Springer-Verlag.

Müller, U.U., A. Schick and W. Wefelmeyer (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. *Statistics* **37**(3), 179–188.

Neumaier, A. (1998). Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review* **40**(3), 636–666.

Poggio, T. and F. Girosi (1990). Networks for approximation and learning. *Proceedings of the IEEE* **78**(9), 1481–1497.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. of Statist* (12), 1215–1230.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels*. MIT Press. Cambridge, MA.

Schwartz, G. (1979). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Sen, A. and M. Srivastava (1997). *Regression analysis, Theory, methods and applications*. Springer-Verlag.

Suykens, J.A.K., T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle (2002). *Least Squares Support Vector Machines*. World Scientific.

Tong, H. (1990). *Nonlinear Time Series Analysis: A Dynamic Approach*. Oxford University Press. Oxford.

Van Gestel, T., J.A.K. Suykens, D. Baestaens, A. Lambrechts, G. Lanckriet, B. Vandaele, B. De Moor and J. Vandewalle (2001). Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Transactions on Neural Networks (special issue on Neural Networks in Financial Engineering)* **12**(4), 809–821.

Vapnik, V.N. (1998). *Statistical Learning Theory*. John Wiley and Sons.

von Neumann, J. and I.N. Schoenberg (1941). Fourier integrals and metric geometry. *Transactions of the American Mathematical Society* (50), 226–251.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Caption of Table

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

Captions of Figures

Figure 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 $\Delta^2_{x,ij} = \|x_i - x_j\|_2^2$ and $\Delta^2_{y,ij} = \|y_i - y_j\|_2^2$ for $i < j = 1,\ldots,N$. The solid line represents the estimated differogram model; **(c)** All differences boxed using a log scale for $\Delta^2_{x,ij}$. The intercept of the curve crossing the Y-axis corresponds to twice the estimated noise variance $2\hat{\sigma}_e^2$.

Figure 2: Differogram of a nonlinear function. **(a)** Data are generated according to the nonlinear dataset described in (Wahba, 1990). with the noise standard deviation of 0.1 and $N = 100$. **(b)** Differogram cloud of all differences $\Delta^2_{x,ij} = \|x_i - x_j\|_2^2$ and $\Delta^2_{y,ij} = \|y_i - y_j\|_2^2$ 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 $\hat{\sigma}_e^2 = \hat{\Upsilon}(0) = 0.1086$.

Figure 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.

Figure 4: **(a)** Log transformation of the Lynx data; **(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.

	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 Dataset						
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

Table 1

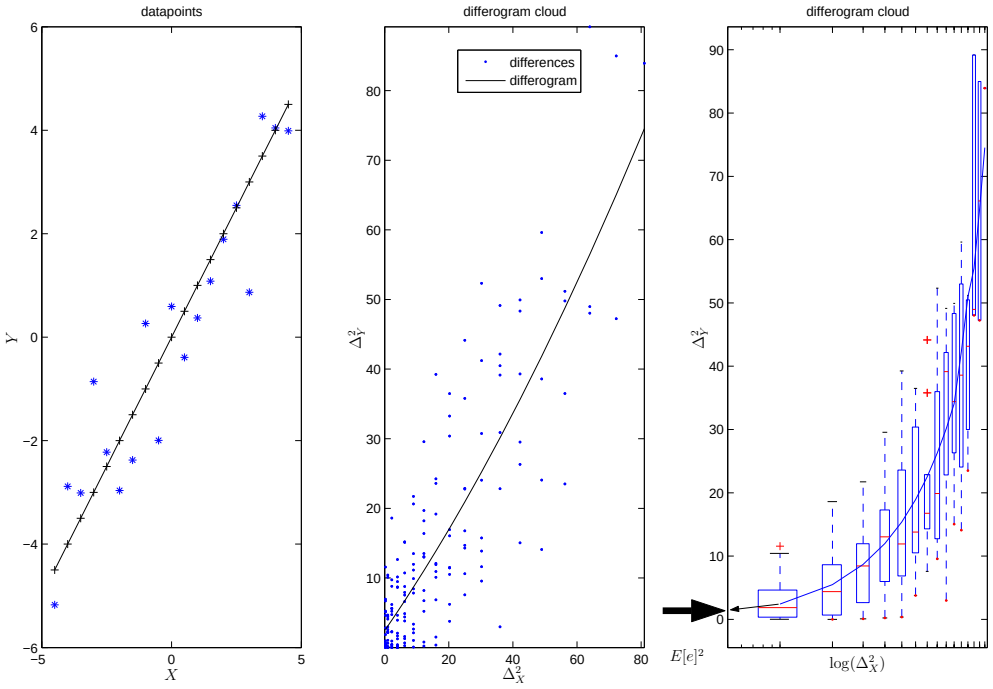
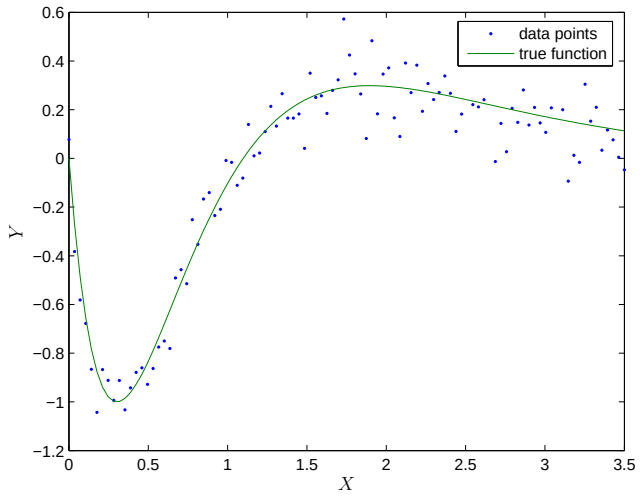
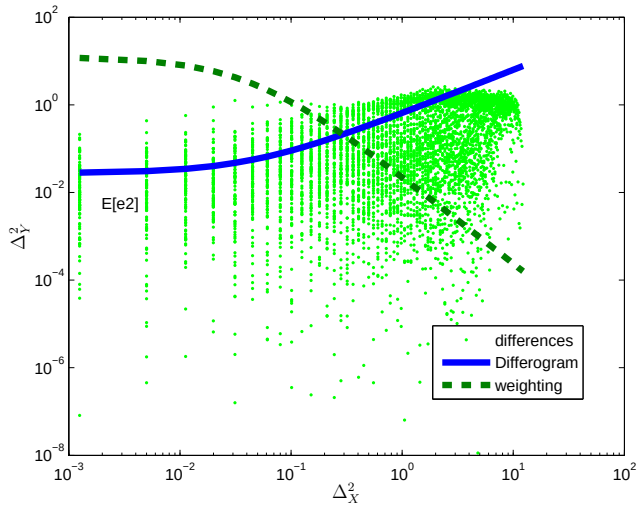

datapoints
differogram cloud
differences
differogram
differogram cloud
$E[\epsilon]^2$
$\log(\Delta_X^2)$

Fig. 1.

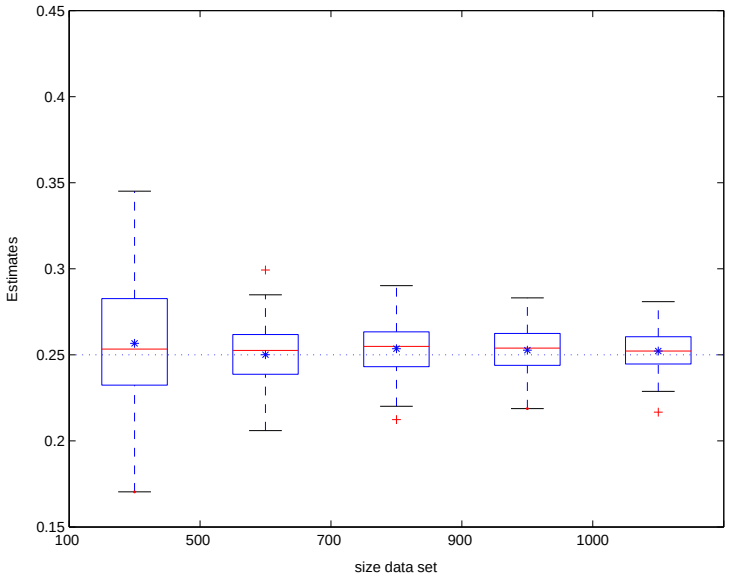

data points
true function
Y
X

(a)

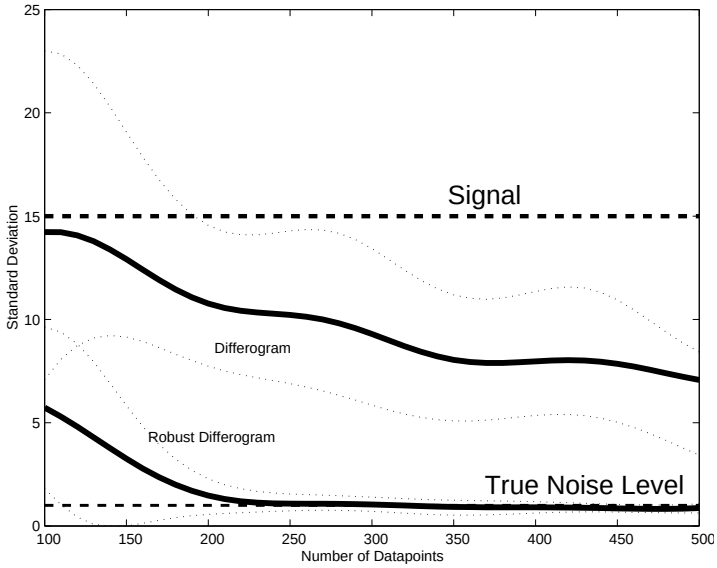

E[e2]
differences
Differogram
weighting
Δ_Y^2
Δ_X^2

(b)

Fig. 2.

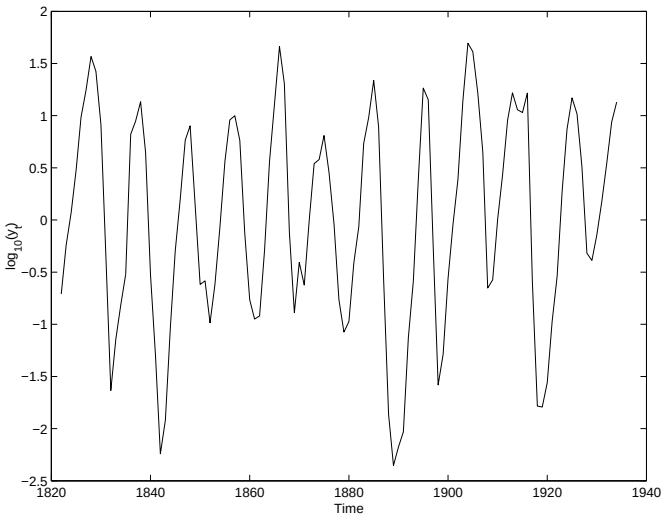

0.45
0.4
0.35
0.3
0.25
0.2
0.15
Estimates
100
500
700
900
1000
size data set

(a)

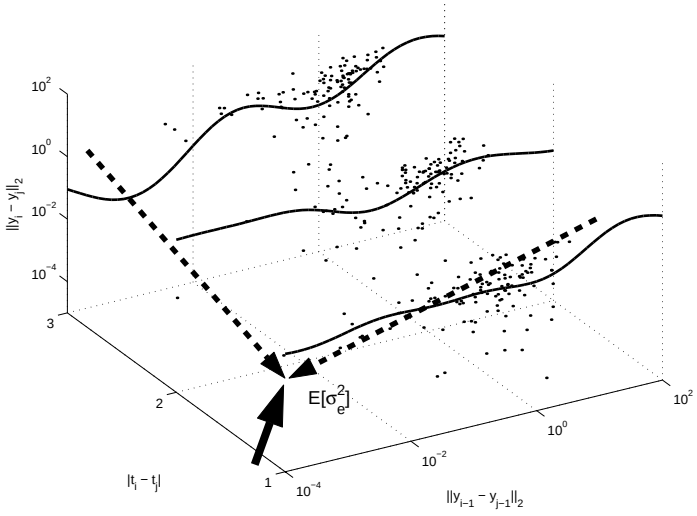

25
20
15
10
5
0
Standard Deviation
Signal
Differogram
Robust Differogram
True Noise Level
100
150
200
250
300
350
400
450
500
Number of Datapoints

(b)

Fig. 3.


2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
1820
1840
1860
1880
1900
1920
1940
Time
log10(yt)

(a)


10^2
10^0
10^−2
10^−4
||yi − yj||2
3
2
1
|ti − tj|
E[σe^2]
10^−4
10^−2
10^0
10^2
||yi−1 − yj−1||2

(b)

Fig. 4.