VARIOGRAM BASED NOISE VARIANCE ESTIMATION AND ITS USE IN KERNEL BASED REGRESSION

Kristiaan Pelckmans, Jos De Brabanter, Johan A.K. Suykens and Bart De Moor
K.U. Leuven - ESAT - SCD/SISTA
Kasteelpark Arenberg 10, B-3001 Leuven - Heverlee (Belgium)
Phone: +32-16-32 85 40, Fax: +32-16-321970
E-mail: {kristiaan.pelckmans, johan.suykens}@esat.kuleuven.ac.be
Web: http://www.esat.kuleuven.ac.be/sista/lssvmlab/

Abstract. Model-free estimates of the noise variance are important for doing model selection and setting tuning parameters. In this paper a data representation is discussed which leads to such an estimator suitable for multi-dimensional input data. The visual representation - called the *differogram cloud* - is based on the 2-norm of the differences of the input- and output-data. A corrected way to estimate the variance of the noise on the output measurement and a (tuning-) parameter free version are derived. Connections with other existing variance estimators and numerical simulations indicate convergence of the estimators. As a special case, this paper focuses on model selection and tuning parameters of Least Squares Support Vector Machines [19].

***Keywords*: Non-parametric Noise Variance Estimator, Signal-to-noise ratio, Geostatistics, U-statistics, Complexity criteria, Least Squares Support Vector Machines**

INTRODUCTION

Noise variance and estimates of this quantity are strongly related with tuning parameters for different modeling techniques. This relation is reflected in a number of applications. Firstly, the variance of the noise plays an important role in various complexity criteria as Akaike's information criteria [1] and C_p-statistic [13] which can be used to select the appropriate model amongst a set (class) of models (see e.g. [20] and [19]). Another point is that the presented approximator gives rise to good initial starting values of the tuning parameters in ridge regression, (least squares) support vector machines, regularization networks and Nadaraya-Watson estimators. This reasoning can be extrapolated easily towards the general class of kernel based methods. Additional links between the noise variance, smoothing and regularization are given by the Bayesian framework [12, 19], Gaussian Processes [12, 19], statistical learning theory [3, 7], splines [21] and regularization theory [15, 7].

One way to avoid the paradox of noise variance estimation before effective modeling and estimating the residuals is to use a non-parametric estimator. The basic principle of the presented estimator is that repeated measurements of a noisy function for the same input reveal properties of the noise. This idea is generalized by relaxing the constraint of equality of inputs. It was already explored in different domains like statistics (as the U-statistic [8]) and geostatistics (as the variogram (an overview is given in [4])).

In the remainder of this section basic concepts from statistics and geostatistics, related methods and existing noise variance estimators are revised. In the second section, the differogram and corresponding noise variance estimators are derived. The third and the fourth section discus the importance of the estimator in the context of tuning parameters and present results of numerical simulations.

Point of view of statistics

Various proposals as to how the error variance might be estimated have been made. The first subclass contains variance estimators based on the residuals from an estimator of the regression itself. In the context of a linear model, one [9] suggested splitting the data into non-intersecting subsets and finding a more complicated model. Another approach is to estimate the regression function locally and to use the sum of squares of residuals about the obtained fit. Such an approach was suggested by [2, 3]. Non-parametric methods as splines were also used for this purpose [21]. The second subclass uses the differences between y-values for which the corresponding x-values are close [5]. These correspond to locally fitting constants or lines [17, 8]. These estimators are independent of a model fit, but the order of difference (the number of related outputs involved to calculate a local residual) has some influence. This paper extends some results from weighted second order based difference estimators [14].

The remainder of this paper operates under following assumptions: let $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ being observations of the input- and the output space. Consider the regression model $y_i = f(x_i) + e_i$ where $x_1, \ldots, x_N$ are deterministic points (with negligible measurement errors), $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown deterministic three times differentiable smooth function and $e_1, \ldots, e_N$ are independent and identically distributed (i.i.d.) random Gaussian errors satisfying the Gauss-Markov (G.-M.) conditions ($E[e_i] = 0$, $E[(e_i)^2] = \sigma_e^2 < \infty$ and $E[e_i e_j] = 0 \quad \forall i \neq j$).

If the regression function f were known, then the errors $e_i = y_i - f(x_i)$ were observable and the sample variance based on the errors can be written as an U-statistic (see [10]): $\hat{\sigma}_e^2 = \frac{1}{N(N-1)} \sum_{i,j=1|i\neq j}^N \frac{1}{2}(e_i - e_j)^2$ and is asymptotically equivalent to the second sample moment $\frac{1}{N}\sum_{i=1}^N e_i^2$. Based on the fact that the regression function is unknown and motivated by the U-statistic of the sample variance, the covariance matched U-statistic is given by $\hat{U}_N\left[(y_i - y_j)^2\right] = \frac{1}{N(N-1)} \sum_{i,j=1|x_i=x_j}^N \frac{1}{2}(y_i - y_j)^2$. In the case of no repeated measurements, a pa-

rameterized weighting scheme $W_{kl}(S)$ was introduced [14]:

$$\hat{U}_{N,W}\left[(y_i - y_j)^2\right] = \frac{1}{N(N-1)} \sum_{i,j=1|i\neq j}^{N} \frac{1}{2}(y_i - y_j)^2 W_{ij}(S). \qquad (1)$$

The estimator U_N is related to difference based estimators for fixed design [17] and [8]. The parameter vector S is typically found by a resampling method.

Point of view of geostatistics

Geostatistics emerged in the 1980's as a hybrid discipline of mining engineering, geology, mathematics and statistics. Applications are also seen in many other disciplines as rainfall precipitation, atmospheric science and soil mapping [4]. Geostatistics considers spatial data $\{y = Z(x) : x \in \mathcal{D}\}$ with $\mathcal{D} \subset \mathbb{R}^d$. The process $Z(.)$ is typically decomposed in the effect of large scale variation (*spatial trend*) $E(Z(x_i))$, small-scale variation (*spatial correlation*) $E(Z(x_i), Z(x_j))$, micro-scale correlation $\lim_{||x_i - x_j||_2 \to 0} E(Z(x_i), Z(x_j))$ and measurement noise σ_e. For analysis of the spatial correlation in the data samples, often the variogram is used. The variogram $\Gamma(\cdot)$ is defined as $2\Gamma(x_i - x_j) = \mathrm{var}(y_i - y_j)$. The relation with the covariance (or auto-covariance) function cov is given as $\mathrm{var}(y_i - y_j) = \mathrm{var}(y_i) + \mathrm{var}(y_j) - 2\mathrm{cov}(y_i, y_j)$. The constraint for a valid variogram to be conditional negative definite relates the variogram with a distance measure [16]. *Kriging* [4, 19, 21] or least squares based spatial prediction makes inferences for unobserved values of the spatial process $Z(x^\star)$ based on the estimated variogram where $x^\star$ is a new point.

Context for this paper

This paper extends insights from statistics and geostatistics towards a machine learning context (see also [21, 4, 19, 20]). This results in a change of focus: firstly the structure of the noise under consideration is i.i.d.: spatial correlation is assumed to be part of the trend structure and micro-scale variations are neglected (motivation by Occam's razor). Secondly, the emphasis is here on the limit behavior of the $||y_i - y_j||_2^2$ when $||x_i - x_j||_2 \to 0$. The consequence is that the estimated differences of the outputs become in a large extent independent of the trend structure. Thirdly, the relation with the Taylor series development is used to derive properties of the true model without the need of a parametric modeling step. Finally the proposed method is naturally extendable towards multi-dimensional inputs as only the norms of the differences are considered. To stress this different context from geostatistics, the proposed representation is called a differogram.

DIFFEROGRAM AND ITS GRAPHICAL REPRESENTATION

Definition 1 *The differogram* $\Upsilon(\cdot)$ *is defined as*

$$2\Upsilon(||x_i - x_j||_2) = \mathrm{var}(y_i - y_j) \quad \textit{when} \quad ||x_i - x_j||_2 \to 0.$$

The differogram model with parameter vector S estimated from the data is denoted as $2\hat{\Upsilon}(||x_i - x_j||_2; S)$. The figure that visualizes these differences is called the differogram cloud.

We abbreviate $||x_i - x_j||_2$ and $||y_i - y_j||_2$ as resp. $||\Delta x_{ij}||_2$ and $||\Delta y_{ij}||_2$. The function $\Upsilon(\cdot)$ is called *isotropic* as it only depends on the norm of the distance between the input points. Figure 1 shows the differogram cloud for data generated with a linear function. The nugget effect is described [4] as the fact that despite

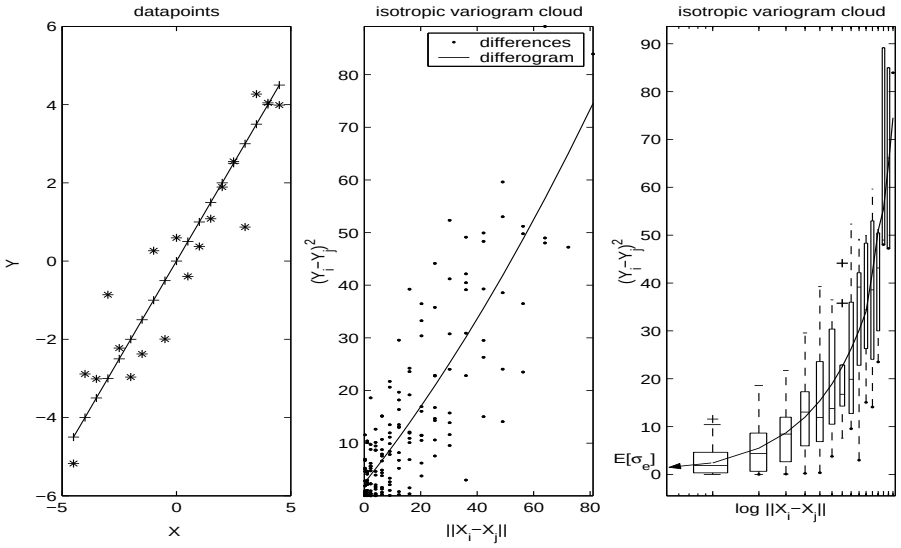


Figure 1: *The differogram cloud for 25 noisy data-points of a linear model is shown in the middle figure (and the boxplots of the data in log scale in the last one). The Taylor model contains estimators of the properties of the true model of the data: the intercept of the Taylor model estimates twice the variance of the noise, the slope approximates the squared slope of the underlying model and the smoothness of the differogram estimates the expected squared smoothness of the underlying model. Note that this equidistant one dimensional case leads to differences which can be grouped in bins very naturally.*

an L_2 continuous true function the variogram $2\Gamma(x_i - x_j) = \mathrm{var}(y_i - y_j)$ does not approach 0 when x_i is approaching x_j. To show the *nugget effect* is related to the noise variance, take the zero-th order Taylor expansion of x_i around point x_j: $T_0[f, x_j](x_i) = f(x_j) + \mathcal{O}^{(1)}$. It follows that (for a generalization, see Lemma 1)

$$\lim_{||x_i - x_j||_2 \to 0} \hat{U}_N\left[(y_i - y_j)^2\right] = \hat{U}_N\left[(f(x_j) + e_i - f(x_j) - e_j)^2\right] = \hat{U}_N\left[(e_i - e_j)^2\right] = \hat{\sigma}_e^2. \tag{2}$$

If micro-scale variations are neglected, an estimator is found for the variance of the noise. This effect relates the estimated variogram and the estimated differogram. However, the derivations of the differogram will focus completely on this extrapolation to zero, as the variogram considers the global covariance of the data. The number of differences can be reduced to $N(N-1)/2$ (as $||\Delta x_{ij}||_2^2$ equal $||\Delta x_{ji}||_2^2$).

In the one-dimensional equidistant case, the differences $||\Delta x_{ij}||_2^2$ will have at most $N-1$ different values and as such they can be ordered (denoted as $||\Delta x_{(n)}||_2^2$ with $||\Delta x_{(n)}||_2^2 \leq ||\Delta x_{(m)}||_2^2$ if $n < m$, with $n, m = 1, \ldots, N-1$). As such, a straightforward approximation of the limit in (2) can be found:

$$\hat{\sigma}_e^{nugget} = \frac{1}{N-1} \sum_{i,j=1 \mid ||x_i - x_j||_2^2 = ||\Delta x_{(1)}||_2^2}^{N} \Big[\frac{1}{2}||\Delta y_{ij}||_2^2\Big]. \tag{3}$$

An improved estimator of the noise variance

The differogram and the Taylor series expansion are related in following result:

Lemma 1 (Differogram) *Given the data $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ generated by a model $y_i = f(x_i) + e_i$, its corresponding differences $\{||\Delta x_{ij}||_2^2, ||\Delta y_{ij}||_2^2\}_{i,j=1}^N \subset \mathbb{R}^+ \times \mathbb{R}^+$, the noise terms e_i Gaussian distributed and satisfying G.-M., f a P times continuously differentiable function with $P \in \{0, 1, 2\}$ the order of the Taylor series under consideration and isotropic f (its covariance structure is only dependent on the norm of the difference vector), then*

$$\hat{\sigma}_e^2 = \hat{U}_N\big[(e_i - e_j)^2\big] = \lim_{||x_i - x_j||_2 \to 0} \Big[\hat{U}_N\big[||\Delta y_{ij}||_2^2\big] - \sum_{p=1}^{P} (\frac{s_p}{p!})^2 ||\Delta x_{ij}||_2^{2p}\Big]$$

with $P \leq s_p^2 = E[||\nabla^{(p)} f||_2^2],\ \forall p = 1, \ldots, P$.

A sketch of the proof is given in Appendix A. Consequently, a natural parametric model for $\Upsilon(\cdot)$ close to the origin is $2\Upsilon(||\Delta x||_2; S) = 2\Upsilon(||\Delta x||_2; s_0, s_1, s_2) = s_0^2 + s_1^2||\Delta x||_2^2 + s_2^2||\Delta x||_2^4$ where $||\Delta x||_2$ denotes the norm of the difference between any two input points x_i and x_j. Remark that the conditional negative definiteness [4] is not required for this model as we only bother about the differogram approaching zero. A constrained least squares criterion

$$\min_{s_0^2, s_1^2, s_2^2} \lim_{||x_i - x_j||_2 \to 0} \sum_{i,j=1 | i \neq j}^{N} \Big[(s_0^2 + s_1^2||\Delta x_{ij}||_2^2 + s_2^2||\Delta x_{ij}||_2^4) - ||\Delta y_{ij}||_2^2\Big]^2 \tag{4}$$

subject to $s_0^2, s_1^2, s_2^2 \geq 0$, can be used to estimate the model parameters. This optimization problem can be solved by quadratic programming.

For example, in the equidistant one-dimensional case, minimally (and optimally) the differences at $||\Delta x_{(n)}||_2$ for $n = 1, 2, 3$ are to be used to compute the optimal model parameters $\hat{s}_0, \hat{s}_1, \hat{s}_2$. By combination of Lemma 1 and (4) an improved estimator is obtained:

$$\hat{\sigma}_e^{corrected} = \sqrt{\hat{s}_0^2/2}. \tag{5}$$

A fast way to handle non-equidistant data is to group the differences into $N-1$ equidistant bins and to proceed as if these contain the equidistant differences

$||\Delta x_{(n)}||_2$. Figure 3.a shows the convergence of this simple estimator (similar to (4)) for data in a 2 dimensional input space.

A more general way to deal with data continuously distributed over the input-space is to define a smoothly decreasing weighting function. (To avoid confusion with SVM and related kernel methods terminology, we shall not use the term *kernel* as frequently used in statistics). Appendix B and [4, 21] motivates mimimizing following cost function w.r.t. S:

$$\mathcal{J}_{wls}(S) = \sum_{i,j=1|i<j}^{N} \left[\frac{\lim_{\epsilon \to 0} 2\hat{\Upsilon}(\epsilon; S)}{2\hat{\Upsilon}(||\Delta x_{ij}||_2; S)}\right]^2 \left[||\Delta y_{ij}||_2^2 - 2\hat{\Upsilon}(||\Delta x_{ij}||_2; S)\right]^2. \quad (6)$$

When the model of (4) is optimized w.r.t. this cost function, a (tuning-) parameter free and bias corrected estimator is obtained:

$$\hat{\sigma}_e^{wls} = \sqrt{\hat{s}_0^2/2}. \quad (7)$$

To solve the non-convex optimization problem of (6), an iterative procedure is presented (as commonly applied in weighted least squares methods):

1. initialize $S = (s_0^2,\ s_1^2,\ s_2^2)^T$ randomly ,
2. compute the weighting function $W(||\Delta x_{ij}||_2; S)$ from S at the previous step,
3. given the new weighting function $W(\cdot)$, optimize S with weighted least squares,
4. redo step 2 and 3 until convergence.

The case of multi-dimensional data becomes similar with non-equidistant input data when $d \geq 2$. Numerical simulations (see Figure 3.b) illustrate that caution should be taken as the accuracy of the estimator decreases proportional with the dimension d.

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 [1], the Bayesian Information Criterion [18] and C_p statistic [13] take the form of a prediction error which consist of the sum of a training set error (e.g. the residual sum of squares) and a complexity term. In general:

$$J(S) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i; S))^2 + \lambda \left(Q_N(\hat{f})\right) \hat{\sigma}_e^2, \quad (8)$$

(see [6]). The complexity term $Q_N(\hat{f})$ represents a penalty term which grows as the number of free parameters (in the linear case) or the effective number of parameters (in the nonlinear case [21, 19]) 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. $J_{RR,\gamma}(w,b) = (1/2)w^T w + (1/2)\gamma \sum_{i=1}^{N}(y_i - w^T x_i - b)^2$. Using the Bayesian interpretation [12, 19] of ridge regression and under i.i.d. Gaussian assumptions, the posterior can be written as $p(w, b \mid x_i, y_i, \mu, \zeta) \propto exp(-\zeta(wx_i + b - y_i)^2)exp(-\mu(w^T w))$, the estimate of the noise variance $\zeta = 1/\hat{\sigma}_e^2$ and the expected variance of the first derivative $\mu = 1/\sigma_{w'}^2 = E\big[(\nabla f)^2\big]$ (see (4)) 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 $\hat{\gamma} = \hat{s}_1^2/\hat{\sigma}_e^2$.

   Another proposed guess for the regularization constant $\hat{\gamma}$ in ridge regression is given by [11]: $\hat{\gamma} = \hat{w}_{LS}^T \hat{w}_{LS}/(\hat{\sigma}_e d)$ where $\hat{\sigma}_e$ is the estimated variance of the noise, d is the number of the free parameters and $\hat{w}_{LS}$ are the estimated parameters of the ordinary least squares problem. This guesses can also be used to set the regularization constant in the parameteric step in fixed size LS-SVM [19].

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. The derived plugin estimator becomes $h_{opt} = C\hat{\sigma}^2 N^{-\frac{1}{5}}$ where $C \approx 6\sqrt{\pi}/25$.

4. We note that $\hat{\sigma}_e^2$ also plays an important role in setting the tuning parameters of SVMs (see [20, 3]).

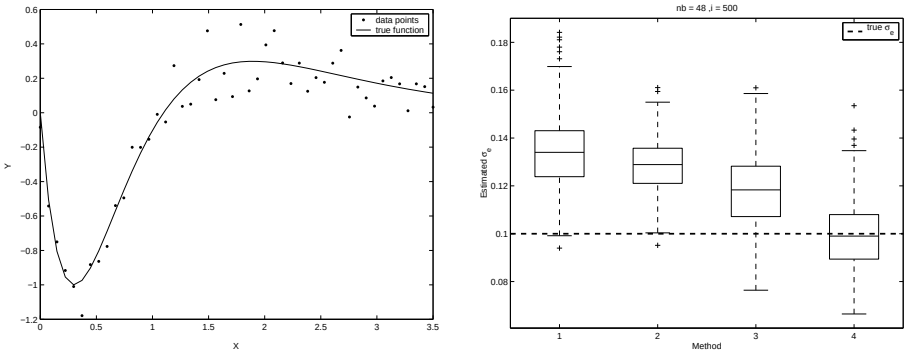
SIMULATION RESULTS

A numerical comparison was made between the presented noise variance estimators. Figure 2 compares the estimators corresponding with (3), (5), (7) and [8] on a classical dataset, described in [21]. Numerical results suggest better performances of (7) when the number of data is low and the signal-to-noise ratio is large. The third order estimator [8] gives better results but is restricted to one dimensional inputs. Figure 3 compares these estimators on multi-dimensional data.

The LS-SVM models with optimal tuning parameters obtained from complexity criteria were compared with classical resampling methods in terms of computation time and accuracy on an the data set generated by $y_i = \mathrm{sinc}(||x_i||_2^2) + e_i$ with $e_i \sim \mathcal{N}(0, 0.1)$ for $i = 1, \ldots, 48$. The LS-SVM model tuned with the C_p complexity criterion [13] performs 25 and 5 times as fast as resp. leave-one-out cross validation (CV) and 10-fold CV, while the generalization performance is not significantly worse.

CONCLUSIONS

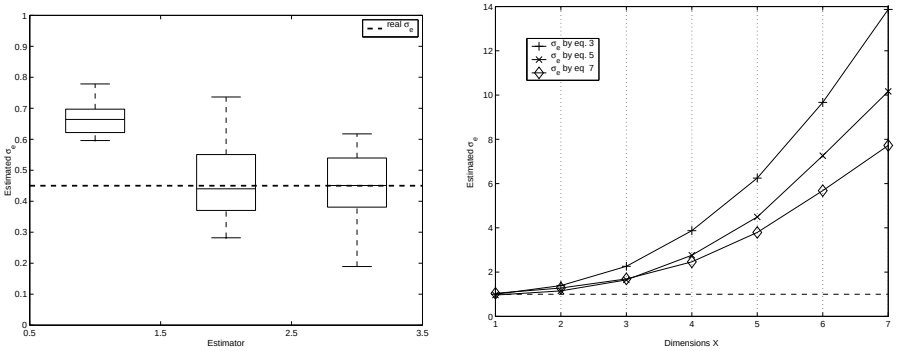
In this paper we presented a graphical data representation suited for exploring properties of the measurements before actual modeling based on connections with U-



(a) Dataset (b) Boxplots of estimates

Figure 2: *Results of a Monte Carlo simulation on the dataset (described by [21]).* $N = 48$ *data-points were taken, with true* $\sigma_e = 0.1$*. The used methods corresponds respectively with (3), (5), (7) and the third order estimator described in [8].*

statistics and geostatistics. The main result is a corrected and tuning parameter free noise variance estimator based on an isotropic (radial) weighting function. Current work on this framework include extensions to spatio-temporal, non-stationary and heteroscedastic data, robustification against non-Gaussian noise and outliers, handling large datasets, extensions towards other distance measures and connections



(a) $2D$, $N = 150$ (b) Increasing Dimensions,$N = 45$

Figure 3: **(a)** *Results on the convergence of the noise variance estimator for a 2 dimensional function with random Gaussian noise with* $\sigma_e = 0.45$*. The functions (*$N = 100$*) in this experiment were generated as in [4], with standard deviations of the output around 1.5. The used methods corresponds respectively with (3), (4) and (6).* **(b)** *Results of the same estimators when the dimension of a dataset is increased (*$N = 45$*). The true* $\sigma_e = 1$ *and the range of the output measurements* $\sigma_Y = 30$*. One can see the quality of the estimates for* σ_e *for different Taylor expansions.*

with bandwidth estimation in the context of kernel machines.

Acknowledgements: Research supported by *Research Council KUL*: GOA-Mefisto 666, IDO (IOTA oncology, genetic networks), several PhD/postdoc & fellow grants; *Flemish Government*: FWO: PhD & postdoc grants, G.0407.02 (support vector machines), projects G.0115.01 (microarrays/oncology), G.0240.99 (multilinear algebra), G.0080.01 (collective intelligence), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0197.02 (power islands), G.0141.03 (identification and cryptography), G.0491.03 (control for intensive care glycemia), G.0120.03 (QIT), research communities (ICCoS, ANMMM); AWI: Bil. Int. Collaboration Hungary, Poland, South Africa; IWT: PhD Grants, STWW-Genprom (gene promotor prediction), GBOU-McKnow (knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors); Soft4s (soft-sensors) *Belgian Federal Government*: DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006); PODO-II (CP/40: TMS and sustainibility); *EU*: CAGE; ERNSI; Eureka 2063-IMPACT; Eureka 2419-FliTE; *Contract Research/agreements*: Data4s, Electrabel, Elia, LMS, IPCOS, VIB; JS is a professor at K.U.Leuven Belgium and a postdoctoral researcher with FWO Flanders. BDM is a full professor at K.U.Leuven Belgium.

REFERENCES

[1] H. Akaike, "Statistical predictor identification," **Ann. Inst. Statist. Math.**, vol. 22, pp. 203–217, 1973.

[2] L. Breiman and W. Meisel, "General estimates of the intrinsic variability of data in nonlinear regression models," **J. Am. Statist. Assoc.**, , no. 71, pp. 301–307, 1976.

[3] V. Cherkassky, "Practical Selection of SVM Parameters and Noise Estimation for SVM Regression," **Neurocomputing, Special Issue on SVM**, submitted, 2002.

[4] N. Cressie, **Statistics for spatial data**, Wiley, 1993.

[5] C. Danniel and F. Wood, **Fitting equations to data**, New York: Wiley, 1971.

[6] J. De Brabanter, K. Pelckmans, J. Suykens, J. Vandewalle and 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.

[7] T. Evgeniou, M. Pontil and T. Poggio, "Regularization Networks and Support Vector Machines," **Advances in Computational Mathematics**, vol. 13, no. 1, pp. 1–50, 2000.

[8] T. Gasser, L. Sroka and C. Jennen-Steinmetz, "Residual variance and residual pattern in nonlinear regression," **Biometrika**, vol. 73, pp. 625–633, 1986.

[9] J. Green, "Testing departure from a regression, without using replication," **Technometrics**, vol. 13, pp. 609–615, 1971.

[10] W. Hoeffding, **The Strong Law of Large Numbers for U-Statistics.**, Univ. North Carolina Inst. Statistics Mimeo Series, No. 302, 1961.

[11] A. Hoerl, R. Kennard and K. Baldwin, "Ridge Regression: Some simulations," **Communications in Statistics, Part A - Theory and Methods**, vol. 4, pp. 105– 123, 1975.

[12] D. MacKay, "The evidence framework applied to classification networks," **Neural Computation**, vol. 4, pp. 698–714, 1992.

[13] C. Mallows, "Some comments on Cp," **Technometrics**, vol. 15, pp. 661–675, 1973.

[14] U. Müller, A. Schick and W. Wefelmeyer, "Estimating the error variance in nonparametric regression by a covariate-matched U-statistic," **Statistics**, 2003 (to appear).

[15] A. Neumaier, "Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization," **SIAM Review**, vol. 40, no. 3, pp. 636–666, 1998.

[16] M. Neumann, "Fully data-driven nonparametric variance estimators," **Statistics**, vol. 25, pp. 189–212, 1994.

[17] J. Rice, "Bandwidth choice for nonparametric regression," **Ann. of Statist**, , no. 12,

pp. 1215–1230, 1984.
[18] G. Schwartz, "Estimating the dimension of a model," **Ann. of Statist.**, vol. 6, pp. 461–464, 1979.
[19] J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle, **Least Squares Support Vector Machines**, World Scientific, 2002.
[20] V. Vapnik, **Statistical Learning Theory**, Wiley and Sons, 1998.
[21] G. Wahba, **Spline models for observational data**, SIAM, 1990.

APPENDIX A: PROOF OF LEMMA 1

Consider here the most general case $P = 2$. The Taylor series expansion of order P of the unknown true function f close to the point x_j is given by $T_2[f, x_j](x) = \lim_{x \to x_j} \Big[f(x_j) + (x - x_j)^T \nabla f(x_j) + (x - x_j)^T \frac{\nabla^2 f(x_j)}{2!} (x - x_j) + \mathcal{O}^{(3)} \Big]$. Neglection of the term $\mathcal{O}^{(3)}$ and application of this in the U-statistic:

$$\hat{U}_N \big[(y_i - y_j)^2 \big] = \lim_{||x_i - x_j||_2 \to 0} \hat{U}_N \Big[\big(T_2[f, x_j](x_i) + e_i - T_2[f, x_j](x_j) - e_j \big)^2 \Big]$$

$$= \lim_{||x_i - x_j||_2 \to 0} \hat{U}_N \Big[\big((e_i - e_j) + (x_i - x_j)^T \nabla f(x_j) + (x_i - x_j)^T \frac{\nabla^2 f(x_j)}{2!} (x_i - x_j) \big)^2 \Big].$$

By working out the square term a series of cross-terms appear. Those including $(e_i - e_j)$ annihilate under the G.-M. conditions. The other, namely $[(x_i - x_j)^T \nabla f(x_j)][1/2!(x - x_j)^T \nabla^2 f(x_j)(x_i - x_j)]$ is dependent on the direction of the difference vector. But as the function is isotropic under assumption, this term should equal zero. This argument concludes the proof.

APPENDIX B: WEIGHTED LEAST SQUARES SMOOTHING

A motivation is given for the weighted least squares cost-function (6) is given. The variance of the estimated differogram is approximated as the squared $\frac{2\Upsilon(||\Delta x_{ij}||_2; S)}{\lim_{\epsilon \to 0} 2\Upsilon(\epsilon; S)}$ (as argued in [4] and motivated by the derivations of generalized cross-validation [21]). As such $W(||\Delta x_{ij}||_2; S) = \big[\frac{\lim_{\epsilon \to 0} 2\Upsilon(\epsilon; S)}{2\Upsilon(||\Delta x_{ij}||_2; S)} \big]^2$ is an appropriate weighting term to be used. Remark that this weighting function has the desirable property $W(0; S) = 1$.

Note that the Gaussian radial basis weighting function is a special of this weighting function where $S = (s_0^2, s_1^2, s_2^2)^T$: the second order Taylor expansion of the *inverse* of this weighting function becomes

$$T[\exp(\frac{||\Delta x||_2^2}{h^2})](0) = 1 + \frac{||\Delta x||_2^2}{h^2} + \frac{1}{2!}(\frac{||\Delta x||_2^2}{h^2})^2 + \mathcal{O}^{(3)}, \qquad (9)$$

so that it is equal with (6) if $h^2 = \frac{\hat{s}_0^2}{\hat{s}_1^2}$ and $h^4 = \frac{\hat{s}_0^2}{2\hat{s}_2^2}$.