Building Sparse Representations and Structure Determination on LS-SVM Substrates

Kristiaan Pelckmans [a] 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 86 58, 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/

Preprint submitted to Elsevier Science *9 July 2004*

Abstract

This paper studies a method to obtain sparseness and structure detection for a class of kernel machines related to Least Squares Support Vector Machines (LS-SVMs). The key method to derive such kernel machines is to adopt an hierarchical modeling strategy. Here, the first level consists of an LS-SVM substrate which is based upon an LS-SVM formulation with additive regularization trade-off. This regularization trade-off is tuned at higher levels such that sparse representations and/or structure detection are obtained. The conceptual levels are kept strictly separated by working with exact optimality conditions, while the hyper-parameters guide the interaction between the levels. From a computational point of view, all levels can be fused into a single convex optimization problem. Furthermore, the principle is applied in order to optimize the validation performance of the resulting kernel machine. Sparse representations as well as structure detection are obtained by using an L_1 regularization scheme and a measure of maximal variation respectively at a higher level. A number of case studies indicate the usefulness of these approaches both with respect to interpretability of the final model as well as for generalization performance.

Key words: Least Squares Support Vector Machines, regularization, structure detection, model selection, convex optimization

1 Introduction

The problem of inference of a model based on a finite set of observational data is nearly always ill-posed (Poggio *et al.*, 1985). To address this problem, typically a form of capacity control is introduced which is often expressed mathematically in the form of regularization (Tikhonov and Arsenin, 1977). Regularized cost functions have been applied successfully e.g. in splines, multilayer perceptrons, regularization networks (Poggio and Girosi, 1990), Support Vector Machines (SVM) and related methods (see e.g. (Hastie *et al.*, 2001)). SVM (Vapnik, 1998) is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation which has also led to many other recent developments in kernel based learning methods in general (Schölkopf and Smola, 2002). SVMs have been introduced within the context of statistical learning theory and structural risk minimization. In the methods one solves convex optimization problems, typically quadratic programs. Least Squares Support Vector Machines (LS-SVMs) (Suykens and Vandewalle, 1999; Saunders *et al.*, 1998) are reformulations to standard SVMs which lead to solving linear Karush-Kuhn-Tucker (KKT) systems for classification tasks as well as regression. Primal-dual LS-SVM formulations have also been given for kFDA, kPCA, kCCA, kPLS, recurrent networks and control (Suykens

et al., 2002) [1] . Recently, LS-SVM methods were studied in combination with additive models (Hastie and Tibshirani, 1990) resulting in so-called componentwise LS-SVMs (Pelckmans *et al.*, 2004*a*) which are suitable for component selection.

The relative importance between the *smoothness* of the solution (in a broad sense) and the norm of the residuals in the cost function involves a tuning parameter, usually called the regularization constant. The determination of regularization constants is important in order to achieve a good generalization performance with the trained model and is an important problem in statistics and learning theory (see e.g. (Hoerl *et al.*, 1975; MacKay, 1992; Hastie *et al.*, 2001; Schölkopf and Smola, 2002; Suykens *et al.*, 2003)). Several model selection criteria have been proposed in literature to tune this constant. Special attention was given in the machine learning community to cross-validation and leave-one-out based methods (Stone, 1974) and fast implementations were studied in the context of kernel machines, see e.g. (Cawley and Talbot, 2003). In this paper, the performance on an independent validation dataset is considered. The optimization of the regularization constant in LS-SVMs with respect to this criterion can be non-convex in general. In order to overcome this difficulty, a re-parameterization of the regularization trade-off has been recently introduced in (Pelckmans *et al.*, 2003) referred to as *additive regularization*. When applied to the LS-SVM formulation, this leads to LS-SVM substrates. In (Pelckmans *et al.*, 2003), it was illustrated how to employ these LS-SVM substrates to obtain models which were optimal in training and validation or cross-validation sense.

This paper investigates these methods towards hierarchical modeling based on optimization theory (Boyd and Vandenberghe, 2004). As in a Bayesian evidence framework (MacKay, 1992), different hierarchical levels can be considered. However, the proposed hierarchical kernel machines are not formulated within a Bayesian context and lead to convex optimization problems (while application of the Bayesian framework often result in non-convex optimization problems with many local minima). The LS-SVM substrate makes up the lowest level of the hierarchy whereas the tuning parameters occur in a particular suited manner and the kernel trick is applied. This paper shows how one can build additional levels resulting in sparse representations and structure detection, while fusing the resulting hierarchical kernel machine into a convex optimization problem. Sparse representations are obtained through the use of an L_1 based regularization scheme, while structure detection is expresed through a measure of maximal variation of a component. The hierarchical kernel model is finalized by a level which tunes the remaining hyper-parameters of the additional levels with a validation criterion.

This paper is organized as follows: Section 2 reviews the formulations of the LS-SVM regressor and its extensions towards componentwise LS-SVMs and LS-SVM

[1] The Internet portal for LS-SVM related research and software can be found at http://www.esat.kuleuven.ac.be/sista/lssvmlab

substrates. Section 3 employs these building blocks to obtain sparse LS-SVMs for sparseness in the support values as well as structure detection. Section 4 discusses how one can optimize the LS-SVM substrate or the sparse LS-SVM with respect to its corresponding regularization trade-offs. Section 5 presents numerical results on both artificial and benchmark datasets.

2 Model Training

Let $\{x_i, y_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ be the training data with inputs x_i and outputs y_i. Consider the regression model $y_i = f(x_i) + e_i$ where $x_1, \ldots, x_N$ are deterministic points (fixed design), $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown real-valued smooth function and $e_1, \ldots, e_N$ are uncorrelated random errors with $E[e_i] = 0$, $E[e_i^2] = \sigma_e^2 < \infty$. The n data points of the validation set are denoted as $\{x_j^v, y_j^v\}_{j=1}^n$. In the case of classification, $y_i, y_j^v \in \{-1, 1\}$ for all $i = 1, \ldots, N$ and $j = 1, \ldots, n$. Let y denote $(y_1, \ldots, y_N)^T \in \mathbb{R}^N$ and $y^v = (y_1^v, \ldots, y_n^v)^T \in \mathbb{R}^n$.

2.1 Least Squares Support Vector Machines

The LS-SVM model is given as $f(x) = w^T\varphi(x)+b$ in the primal space where $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_h}$ denotes the potentially infinite ($n_h = \infty$) dimensional feature map. The regularized least squares cost function is given by (Saunders *et al.*, 1998; Suykens *et al.*, 2002)

$$\min_{w,b,e_i} \mathcal{J}_\gamma(w, e) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^N e_i^2 \text{ s.t. } w^T\varphi(x_i)+b+e_i = y_i, \quad \forall i = 1, ..., N. \quad (1)$$

Note that the regularization constant γ appears here as in classical Tikhonov regularization (Tikhonov and Arsenin, 1977). Let $e = (e_1, \ldots, e_N)^T \in \mathbb{R}^N$ be a vector of residuals. The Lagrangian of the constrained optimization problem becomes $\mathcal{L}_\gamma(w, b, e_i; \alpha_i) = 0.5w^T w + 0.5\gamma\sum_{i=1}^N e_i^2 - \sum_{i=1}^N \alpha_i(w^T x_i + b + e_i - y_i)$. By taking the conditions for optimality $\partial\mathcal{L}_\gamma/\partial\alpha_i = 0$, $\partial\mathcal{L}_\gamma/\partial b = 0$, $\partial\mathcal{L}_\gamma/\partial w = 0$, $\partial\mathcal{L}_\gamma/\partial e_i = 0$ and application of the kernel trick $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$ with a positive definite (Mercer) kernel K, one gets $e_i\gamma = \alpha_i$, $w = \sum_{i=1}^N \alpha_i\varphi(x_i)$, $\sum_{i=1}^N \alpha_i = 0$ and $w^T\varphi(x_i) + b + e_i = y_i$. The dual problem can be summarized as

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + I_N/\gamma \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \quad (2)$$

where $\Omega \in \mathbb{R}^{N\times N}$ with $\Omega_{ij} = K(x_i, x_j)$. The estimated function $\hat{f}$ can be evaluated at a new point x^* by $\hat{f}(x^*) = \sum_{i=1}^N \hat{\alpha}_i K(x_i, x^*) + \hat{b}$ where $\hat{\alpha}$ and $\hat{b}$ are the unique solution to (2).

Optimization of the regularization constant γ with respect to a validation performance in the regression case can be written as

$$\min_{\gamma} \sum_{j=1}^{n} (y_j^v - \hat{f}_\gamma(x_j^v))^2 = \sum_{j=1}^{n} \left(y_j^v - \left[\begin{array}{c|c} 1 & \\ \hline \Omega^v & \end{array}\right]^T \left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + I_N/\gamma \end{array}\right]^{-1} \left[\begin{array}{c|c} 0 & \\ \hline y & \end{array}\right] \right)^2, \tag{3}$$

where $\Omega^v \in \mathbb{R}^{N \times n}$ with $\Omega_{ij}^v = K(x_i, x_j^v)$ for all $i = 1, \ldots, N$ and $j = 1, \ldots, n$. This optimization problem in the hyper-parameter γ is usually non-convex. For the choice of the kernel $K(\cdot, \cdot)$, see e.g. (Genton, 2001; Chapelle *et al.*, 2002). Typical examples are the use of a linear kernel $K(x_i, x_j) = x_i^T x_j$ or the RBF kernel $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$ where σ denotes the bandwidth of the kernel.

A derivation of LS-SVMs was given originally for the classification task (Suykens and Vandewalle, 1999). The LS-SVM classifier $\mathrm{sign}(w^T\varphi(x)+b)$ is optimized with respect to

$$\min_{w,b,e_i} \mathcal{J}_\gamma(w, e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \text{ s.t. } y_i(w^T\varphi(x_i)+b) = 1 - e_i, \quad \forall i = 1, \ldots, N, \tag{4}$$

where $e = (e_1, \ldots, e_N)^T \in \mathbb{R}^N$ are slack-variables to tolerate misclassifications. Using a primal dual optimization interpretation, the unknowns $\hat{\alpha}, \hat{b}$ of the estimated classifier $\mathrm{sign}(\sum_{i=1}^N \hat{\alpha}_i y_i K(x_i, x) + \hat{b})$ are found by solving the dual set of linear equations

$$\left[\begin{array}{c|c} 0 & y^T \\ \hline y & \Omega^y + I_N/\gamma \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 1_N \end{array}\right], \tag{5}$$

where $\Omega^y \in \mathbb{R}^{N \times N}$ with $\Omega_{ij}^y = y_i y_j K(x_i, x_j)$ for all $i, j = 1, \ldots, N$. The remainder focuses on the regression case, although it is applicable just as well to the classification problem as illustrated in Section 5.

2.2 Componentwise LS-SVMs

It is often useful to consider less general classes of nonlinear models as the additive model class (Hastie and Tibshirani, 1990) in order to overcome the curse of dimensionality. Let a superscript l of the input data x denote the lth component (variable). Consider the following model

$$f(x) = \sum_{l=1}^{d} w_l^T \varphi_l(x^l) + b, \tag{6}$$

where $\varphi_l : \mathbb{R} \to \mathbb{R}^{d_f}$ is a possibly infinite dimensional mapping of the lth component. One considers the regularized least squares cost-function as (Pelckmans *et*

al., 2004*a*),

$$\min_{w_l,b,e_i} \mathcal{J}_\gamma(w_l,e) = \frac{1}{2}\sum_{l=1}^{d} w_l^T w_l + \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2$$
$$\text{s.t.} \quad \sum_{l=1}^{d} w_l^T \varphi_l(x_i^l) + b + e_i = y_i, \quad \forall i = 1,...,N. \quad (7)$$

Construction of the Lagrangian and taking the conditions for optimality as in the previous subsection results in the following linear system

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + I_N/\gamma \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \quad (8)$$

where $\Omega = \sum_{l=1}^{d} \Omega^l \in \mathbb{R}^{N\times N}$ and $\Omega_{ij}^l = K^l(x_i^l, x_j^l)$ is the kernel evaluated between the lth component of the ith and the jth data-point. It is interesting to compare this result with the iterative back-fitting procedure (Hastie and Tibshirani, 1990), the kernel ANOVA decomposition (Stitson *et al.*, 1999; Gunn and Kandola, 2002) and the splines technique for additive models MARS (Wahba, 1990; Hastie *et al.*, 2001). Note that the difference between (2) and (8) can be stated entirely in terms of the used kernels, though the starting point was a different model and optimality criterion.

2.3 Substrate LS-SVMs with additive regularization

An alternative way to parameterize the regularization trade-off associated with the model $f(x) = w^T\varphi(x) + b$ is by means of a tuning vector $c \in \mathbb{R}^N$ (Pelckmans *et al.*, 2003):

$$\min_{w,b,e_i} \mathcal{J}_c(w,e) = \frac{1}{2}w^T w + \frac{1}{2}\sum_{i=1}^{N}(e_i - c_i)^2 \ \text{ s.t. } \ w^T\varphi(x_i) + b + e_i = y_i, \quad \forall i = 1,\ldots,N, \quad (9)$$

where the elements of the vector c serve as tuning parameters, called additive regularization constants. Note that at this level, the tuning parameters c_i are fixed. An interpretation of these hyper-parameters is that they influence the distribution of the residuals. After constructing the Lagrangian with multipliers α and taking the conditions for optimality w.r.t. w, b, e_i, α_i (being $e_i = c_i + \alpha_i$, $w = \sum_{i=1}^{N} \alpha_i\varphi(x_i)$, $\sum_{i=1}^{N} \alpha_i = 0$ and $w^T\varphi(x_i) + b + e_i = y_i$), the following dual linear system is obtained

$$\left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega + I_N \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] + \left[\begin{array}{c} 0 \\ \hline c \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right], \quad (10)$$

and $\alpha + c = e$. Note that at this point the value of c is not considered as an unknown to the optimization problem: once c is fixed, the solution of α, b is uniquely determined.

The estimated function $\hat{f}$ can be evaluated at a new point x^* by $\hat{f}(x^*) = \hat{w}^T\varphi(x^*) + \hat{b} = \sum_{i=1}^{N} \hat{\alpha}_i K(x_i, x^*) + \hat{b}$ where $\hat{\alpha}$ and $\hat{b}$ are the solutions to (10). The residual $\hat{f}(x_j^v) - y_j^v$ is denoted as e_j^v such that one can write

$$y_j^v = w^T\varphi(x_j^v) + b + e_j^v = \sum_{i=1}^{N} \alpha_i K(x_i, x_j^v) + b + e_j^v, \quad \forall j = 1, \ldots, n. \tag{11}$$

By comparison of (2) and (10), LS-SVMs with Tikhonov regularization can be seen as a special case of LS-SVM substrates with the following additional constraint on the vectors α, c, γ

$$\gamma^{-1}\alpha = \alpha + c \quad \text{s.t.} \quad \gamma > 0. \tag{12}$$

This means that the solution to the LS-SVM substrates are also solutions to LS-SVMs whenever the support values α_i are proportional to the residuals $e_i = \alpha_i + c_i$ for all $i = 1, \ldots, N$. This type of collinearity or rank constraints is known to be very hard to cast as convex optimization problems (Boyd and Vandenberghe, 2004). Finally, note that the additive regularization scheme does not replace the tikhonov regularization scheme, but parameterizes the trade-off in a different way which allows for inference of primal-dual kernel machines based on alternative criteria (Pelckmans *et al.*, 2004*c*) as will be shown in the next sections.

3 LS-SVM Substrates for Structure Detection and Sparse Representations

Now, we show that when using LS-SVM substrates and componentwise LS-SVMs, sufficient tools are at our disposal in order to design a kernel machine which detects structure in the model components and/or results in sparseness in the support vectors while working with a least squares loss function.

3.1 Structure detection by sparse components

The formulation of componentwise LS-SVMs suggests the use of a dedicated regularization scheme which is often very useful in practice. In the case where the nonlinear function consists of a sum of (one-dimensional) components, one may ask oneself which components have a trivial contribution ($f^l(\cdot) = 0$) for prediction. Sparseness amongst the components is often referred to as structure detection. The described method is closely related to the kernel ANOVA decomposition (Stitson *et al.*, 1999) and the structure detection method of (Gunn and Kandola, 2002), however, this paper starts from a clear optimality principle.

To formalize this argument, the measure of maximal variation of components is introduced as follows:

$$t_l = \max_i |w_l^T \varphi(x_i^l)|, \quad \forall l = 1, \ldots, d. \tag{13}$$

Note that this measure is related to the measure of total variation, see e.g. (Rudin *et al.*, 1992). By using this measure as a regularization term, optimization problems are obtained which require much less (primal) variables and as such can handle much higher dimensions as the method employed in e.g. (Pelckmans *et al.*, 2004*b*). The kernel machine for structure detection minimizes the following criterion for a given tuning constant $\rho \in \mathbb{R}_0^+$:

$$\min_{e,\alpha,b,c} \|e\|_2^2 + \rho \sum_{l=1}^{d} t_l \text{ s.t. (10) holds and}$$
$$-t_l \leq w_l^T \varphi(x_i^l) \leq t_l, \ \forall i = 1, \ldots, N, \forall l = 1, \ldots, d. \tag{14}$$

It is known that the use of 1-norms may lead to sparse solution which are unnecessarily biased (Fan, 1997). To overcome this drawback, one has proposed the use of norms as the Smoothly Clipped Absolute Deviation (SCAD) penalty function as suggested by (Fan, 1997) and implemented in a primal-dual kernel machine in (Pelckmans *et al.*, 2004*a*). This paper will not pursue this issue as it leads to non-convex optimization criteria in general. Instead, the use of the 1-norm is studied to detect structure, while the final predictions should be made based on a classical model with the selected components (compare to basis pursuit, see e.g. (Chen *et al.*, 2001)).

3.2 Sparse LS-SVM substrates

Sparseness is often regarded as good practice in the machine learning community (Vapnik, 1998; von Luxburg *et al.*, 2004) as it gives an optimal and minimal representation of the solution (from the viewpoint of VC theory and compression) The primal-dual framework also provides another line of thought based on sensitivity analysis as explained in (Boyd and Vandenberghe, 2004). Consider the derivation of the LS-SVM (2) or the LS-SVM substrate (10). The optimal Lagrange multipliers $\hat{\alpha}$ contain information of how much the (dual) optimal solution changes when a constraint is perturbed. This perturbation is proportional to c as can be seen from the constraints in (9). As such, one can write

$$\hat{\alpha}_i = -\frac{\partial p^*}{\partial c_i} \quad \text{and} \quad p^* = \inf_{w,b} \mathcal{J}_c \quad \text{s.t. (9) holds,} \tag{15}$$

see (Boyd and Vandenberghe, 2004). In this respect, one can design a kernel machine that minimizes its own sensitivity to model misspecifications or atypical data

observations by minimizing an appropriate norm of the Lagrange multipliers. The 1-norm is considered

$$\min_{e,\alpha,b,c,e^v} \|e\|_2^2 + \zeta\|\alpha\|_1 \text{ s.t. (10) holds and } \alpha + c = e. \tag{16}$$

where $\zeta \in \mathbb{R}_0^+$ acts as a hyper-parameter. This criterion leads to sparseness (Vapnik, 1998) as already exploited under the name of ν-SVM (Chang and Lin, 2002). A similar principle is applied in the estimation of sparse parametric models known as basis pursuit (Chen *et al.*, 2001) or LASSO (Hastie *et al.*, 2001), but these approaches lack the above interpretation in terms of sensitivity and are treated at different hierarchical levels.

4 Fusion of training and Validation

4.1 Fusion of Tikhonov regularized LS-SVM with validation

We now study the interplay between exact training and optimizing the regularization trade-off with respect to a validation criterion. For this purpose, the fusion argument as introduced in (Pelckmans *et al.*, 2003) is briefly discussed in relation to regularization parameter tuning of the standard LS-SVM. The estimation of the LS-SVM regressor on the training data for a fixed value γ is given as (2)

$$\textbf{Level 1}: \ (\hat{w}, \hat{b}, \hat{e}) = \arg\min_{w,b,e} \mathcal{J}_\gamma(w, e) \quad \text{s.t. constraints (1) hold,} \tag{17}$$

which results into solving a set of linear equations (2) after elimination of w and e. Tuning the regularization parameter by using a validation criterion amounts to minimizing e.g. the following cost on a validation set

$$\textbf{Level 2}: \ \hat{\gamma} = \arg\min_{\gamma} \sum_{j=1}^{n} \left(f(x_j^v; \hat{\alpha}, \hat{b}) - y_j^v \right)^2 \quad \text{with} \quad (\hat{\alpha}, \hat{b}) = \arg\min_{\alpha,b} \mathcal{J}_\gamma, \tag{18}$$

satisfying again the constraints of (1) (compare with (3) in a LS-SVM context). Using the conditions for optimality (2), one can rewrite (18) as

$$\begin{aligned} \textbf{Fusion}: \ (\hat{\gamma}, \hat{\alpha}, \hat{b}) = \arg\min_{\gamma,\alpha,b} & \sum_{j=1}^{n} (e_j^v)^2 \\ \text{s.t.} \quad & \begin{cases} \sum_{i=1}^{N} \alpha_i K(x_i, x_j^v) + b + e_j^v = y_j^v, \ \forall j = 1, \ldots, n \\ \text{(2) holds,} \end{cases} \end{aligned} \tag{19}$$

which is referred to as *fusion* of training and validation. The resulting constrained optimization problem was noted to be non-convex as the optimal solutions w, b or dual variables α corresponding with all values $\gamma > 0$ describe a non-convex set.

When one adopts the additive regularization framework of Subsection 2.3, one obtains a convex optimization problem with unknowns c, α, e^v and b as (19) at the cost of over-parameterizing the trade-off. By combination of the (exact!) training conditions (10) and validation equalities (11), an under-determined linear system is obtained with unknowns α, b, c and e^v, summarized as follows

$$\left[\begin{array}{cc|cc} 0_N^T & 0_n^T & 0 & 1_N^T \\ \hline I_N & 0_{N\times n} & 1_N & \Omega + I_N \\ \hline 0_{n\times N} & I_n & 1_n & \Omega^v \end{array}\right] \left[\begin{array}{c} c \\ \hline e^v \\ \hline b \\ \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \\ \hline y^v \end{array}\right], \tag{20}$$

and $e = \alpha + c$. Different schemes for finding a 'best' among the many candidate solutions of the under-determined system (20) can be considered, e.g.

$$\min_{\alpha,b,c,e,e^v} \|e\|_2^2 + \|e^v\|_2^2 \text{ s.t. (20) holds and } e = \alpha + c. \tag{21}$$

This criterion is motivated by the assumption that the training as well as the validation criterion are independently sampled from the same distribution. The criterion (21) leads to a unique solution as obtained by solving this constrained linear least squares problem

$$\left\| \left[\begin{array}{c|c} 1_N & \Omega \\ \hline 1_N & \Omega^v \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] - \left[\begin{array}{c} y \\ \hline y^v \end{array}\right] \right\|_2^2 \quad \text{s.t.} \quad 1_N^T \alpha = 0. \tag{22}$$

However, straightforward application of the criterion (21) should be avoided whenever the number of training data exceeds the number of validation points as overfitting might occur on the validation data as shown in (Pelckmans *et al.*, 2003). One can overcome this problem by confining the space of possible c values (Pelckmans *et al.*, 2003) to an appropriately chosen subset. In (Pelckmans *et al.*, 2004*b*), the use of the following multi-criterion was advised:

$$\mathcal{J}(e, e^v; \alpha, c, b) = \|e\|_2^2 + \|e^v\|_2^2 + \xi\|\alpha\|_1 \text{ s.t. (20) holds and } e = \alpha + c. \tag{23}$$

While this problem is well defined, the reader may object that as a new hyperparameter $\xi \in \mathbb{R}_0^+$ has popped up.

This paper extends the results above (Pelckmans *et al.*, 2004*b*) by tuning the hyper-parameter ρ in (14) and ζ in (16) with respect to a validation criterion instead of the tuning parameter vector c. Only the case of fusion for structure detection as described in Subsection 3.1 is elaborated, the case of fusion towards finding sparseness follows along the same lines. Figure 1 gives a schematical representation of fusing the hierarchical setting of using a validation measure to tune the hyper-parameters of the LS-SVM substrate for structure detection. For notational convenience, the bias term b is omitted in the sequel. Let $t = (t_1, \ldots, t_d)^T \in \mathbb{R}^d$ be a vector of bounds on the maximal variation per component. Consider the optimization problem (14):

$$\min_{t,\alpha,c,e} \mathcal{J}_\rho(e,t) = \frac{1}{2}\|e\|_2^2 + \rho \sum_{l=1}^{d} t_l \quad \text{s.t.} \quad \begin{cases} (\Omega + I_N)\,\alpha + c = y, \\ \alpha + c = e, \\ -t_l 1_N \leq \Omega^l \alpha \leq t_l 1_N,\ \forall l = 1, \ldots, d, \end{cases} \tag{24}$$

where $\rho > 0$ acts as a hyper-parameters. One can eliminate e and c from (24) leading to

$$\min_{\alpha,t} \mathcal{J}_\rho(\alpha,t) = \frac{1}{2}\|\Omega\alpha - y\|_2^2 + \rho \sum_{l=1}^{d} t_l \quad \text{s.t.} \quad -t_l 1_N \leq \Omega^l \alpha \leq t_l 1_N,\ \ \forall l = 1, \ldots, d. \tag{25}$$

The Lagrangian becomes

$$\begin{aligned} \mathcal{L}_\rho(\alpha, t; \xi^{+l}, \xi^{-l}) = \frac{1}{2}\,\|\Omega\alpha - y\|_2^2 + \rho \sum_{l=1}^{d} t_l \\ + \sum_{l=1}^{d} \xi^{-l^T} \left(-t_l 1_N - \Omega^l \alpha\right) + \sum_{l=1}^{d} \xi^{+l^T} \left(-t_l 1_N + \Omega^l \alpha\right), \end{aligned} \tag{26}$$

with multipliers ξ^{+l} and $\xi^{-l} \in \mathbb{R}^{+,N}$ for all $l = 1, \ldots, d$. The corresponding Karush-Kuhn-Tucker conditions are necessary as well as sufficient as the primal problem is convex (Boyd and Vandenberghe, 2004), p.244, for the determination of

the global optimum:

$$\begin{cases} \Omega^T\Omega\alpha - y^T\Omega = \sum_{l=1}^{d}(\xi^{-l} - \xi^{+l}), & (a) \\ \rho = \sum_{l=1}^{d}(\xi^{-l} + \xi^{+l}) & (b) \\ \xi^{+l}, \xi^{-l} \geq 0, & \forall l = 1, \ldots, d \ \ (c) \\ -t_l 1_N \leq \Omega^l \alpha \leq t_l 1_N, & \forall l = 1, \ldots, d \ \ (d) \\ \xi_i^{-l}(t_l + \Omega_i^l \alpha) = 0, & \forall i = 1, \ldots, N, \ \forall l = 1, \ldots, d \ \ (e) \\ \xi_i^{+l}(t_l - \Omega_i^l \alpha) = 0, & \forall i = 1, \ldots, N, \ \forall l = 1, \ldots, d \ \ (f) \end{cases} \tag{27}$$

which are all linear (in)equalities except for the so-called complementary slackness conditions (e) and (f). Now consider the fusion of the validation criterion and this conditions with respect to the hyper-parameters t:

$$\min_{\rho,t,\alpha,\xi^-,\xi^+} \mathcal{J}^v(\rho; t, \alpha, \xi^-, \xi^+) = \frac{1}{2}\|\Omega^v\alpha - y^v\|_2^2 \quad \text{s.t.} \quad (27) \text{ hold}, \tag{28}$$

where $\Omega^v \in \mathbb{R}^{n \times N}$ equals $\Omega^v_{ji} = \sum_{l=1}^{d} K^l(x_i^l, x_j^{vl})$ for all $i = 1, \ldots, n$ or $j = 1, \ldots, N$. Except for conditions (e) and (f), the problem (28) can be solved as a Quadratic Programming (QP) problem. One can show that the modified QP below results in the same optimal solution as (28):

$$\begin{aligned} \min_{\rho;t,\alpha,\xi^-,\xi^+} \mathcal{J}^v(\rho; t, \alpha, \xi^-, \xi^+) = \|\Omega^{vd}\alpha - y^v\|_2^2 + \sum_{l=1}^{d} \xi^{-l^T}\left(t_l 1_N + \Omega_i^l \alpha\right) \\ + \sum_{l=1}^{d} \xi^{+l^T}\left(t_l 1_N - \Omega_i^l \alpha\right) \quad \text{s.t. conditions (27.abcd) holds.} \end{aligned} \tag{29}$$

Crucial in this formulation is the observation that the complementary slackness terms $\xi^{-l^T}\left(t_l 1_N + \Omega_i^l \alpha\right)$ and $\xi^{+l^T}\left(t_l 1_N - \Omega_i^l \alpha\right)$ are bounded from below by zero as all cross-product terms are positive. By checking the complementary slackness conditions (27.ef) again on the resulting solution, one can verify that the problem is indeed solved accurately. As for componentwise LS-SVM substrates, the model can be evaluated in new data points $x_* \in \mathbb{R}^d$ as

$$\hat{f}(x^*) = \hat{w}^T\varphi(x^*) = \sum_{i=1}^{N} \hat{\alpha}_i \sum_{t_l \neq 0} K^l(x_i^l, x_*^l), \tag{30}$$

where $\hat{\alpha}$ is the solution to (29).

5 Experiments

5.1 Sparseness

The performance of the proposed sparse LS-SVM substrate was measured on a number of regression and classification datasets, respectively an artificial dataset sinc (generated as $X = \mathrm{sinc}(X) + e$ with $e \sim \mathcal{N}(0, 0.1)$ and $N = 100$, $d = 1$) and the motorcycle dataset (Eubank, 1999) ($N = 100$, $d = 1$) for regression, the artificial Ripley dataset ($N = 250$, $d = 2$) and the PIMA dataset ($N = 468$, $d = 8$) from UCI for classification. The models resulting from sparse LS-SVM substrates were tested against the classical SVMs and LS-SVMs where the kernel parameters and the other tuning-parameters (respectively C, ϵ for the SVM, γ for the LS-SVM and ξ for sparse LS-SVM substrates) were obtained from 10-fold cross-validation (see Table 1).

5.2 Structure detection

In order to test the structure detection mechanism, an artificial example is taken from (Vapnik, 1998) and the Boston housing dataset from the UCI benchmark repository was used for analyzing the practical relevance of the structure detection mechanism. This subsection considers the formulation as elaborated in Subsection 3.1, where sparseness amongst the components is obtained by use of the sum of maximal variation. The performance on a validation set was used to tune the parameter ρ both manually as well as described in Subsection 4.3.

Figure 4 and 5 shows results obtained on an artificial dataset consisting of 100 samples and dimension 25, uniformly sampled from the interval $[0, 1]^{25}$. The underlying function takes the following form:

$$f(x) = 10 \sin(X^1) + 20 \, (X^2 - 0.5)^2 + 10 \, X^3 + 5 \, X^4 \tag{31}$$

such that $y_i = f(x_i) + e_i$ with $e_i \sim \mathcal{N}(0, 1)$ for all $i = 1, \ldots, 100$. Figure 5 gives the nontrivial components ($t_l > 0$) associated with the LS-SVM substrate with ρ optimized in validation sense. Figure 4 presents the evolution of values of t when ρ is increased from 1 to 1000 in a maximal variation evolution diagram (similarly as used for LASSO (Hastie *et al.*, 2001)).

The Boston housing dataset was taken from the UCI benchmark repository. This dataset concerns the housing values in suburbs of Boston. The dependent continuous variable expresses the median value of owner-occupied homes. From 13 given inputs, an additive model was build up where the mechanism of maximal variation was used to detect which input variables have a non-trivial contribution. 250

datapoints were used for training purposes and 100 were randomly selected for validation. The analysis works with standardized data (zero mean and unit variance), while results are expressed in the original scale. The structure detection algorithm as proposed in Subsection 3.1 was used to construct the maximal variation evolution diagram, see Figure 6. The performance on the validation dataset was used to tune the kernel parameter and ρ. The latter was determined both manually (by a line-search) as automatically by fusion as described in Subsection 4.3. For the optimal parameter ρ, the following inputs have a maximal variation of zero:

1 CRIM: per capita crime rate by town,
2 ZN: proportion of residential land zoned for lots over $25,000$ sq.ft.,
4 CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise),
10 TAX: full-value property-tax rate per $10,000$,
12 B: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks.

When omitting these variables, the additive model increases in performance on an independent testset with 22%. The improvement is even more significant (32%) with respect to a general nonlinear LS-SVM model with an RBF-kernel.

6 Conclusions

This paper discussed an hierarchical method to build kernel machines on LS-SVM substrates resulting in sparseness and/or structure detection. The hierarchical modeling strategy is enabled by the use of additive regularization and the exploitation of necessary and sufficient KKT conditions, while interactions between the levels are guided by a proper set of hyper-parameters. Higher levels are based on the one hand on L_1 regularization and a measure of maximal variation, and on the other hand on maximization of the validation performance. While the resulting hierarchical kernel machine has properly separated the conceptual levels, the machine can be fused into a single convex optimization problem resulting in the training solution and the hyper-parameters at once. A number of experiments illustrate the use of the elaborated method both with respect to interpretability as well as generalization performance.

Acknowledgments. This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven. It is supported by grants from several funding agencies and sources: Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD / postdoc grants, projects G.0407.02, G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), Eureka-Impact (MPC-control), Eureka-FLiTE (flutter modeling), several PhD grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP/40: Sustainibility effects of Traffic Management Systems); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS and BDM are an associate and full professor with K.U.Leuven Belgium, respectively.

References

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Cawley, G.C. and N.L.C. Talbot (2003). Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognition* **36**(11), 2585–2592.

Chang, C.C. and C.J. Lin (2002). Training nu-support vector regression: theory and algorithms. *Neural Computation* **14**(8), 1959–77.

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

Chen, S.S., D.L. Donoho and M.A. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM Review* **43**(1), 129–159.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. Vol. 157. Marcel Dekker, New York.

Fan, J. (1997). Comments on wavelets in statistics: A review. *Journal of the Italian Statistical Association* (6), 131–138.

Genton, M.G. (2001). Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research* **2**, 299–312.

Gunn, S. R. and J. S. Kandola (2002). Structural modelling with sparse kernels. *Machine Learning* **48**(1), 137–163.

Hastie, T. and R. Tibshirani (1990). *Generalized addidive models*. London: Chapman and Hall.

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

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.

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

Pelckmans, K., I. Goethals, J. De Brabanter, J.A.K. Suykens and B. De Moor (2004*a*). Componentwise least squares support vector machines. *(Submitted for Publication) Internal Report 04-75, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2003). Additive regularization: Fusion of training and validation levels in kernel methods. *(Submitted for Publication) Internal Report 03-184, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2004*b*). Sparse LS-SVMs using additive regularization with a penalized validation criterion. In: *Proceedings of the 12e European Symposium on Artificial Neural Networks*. Vol. 12. pp. 435–440.

Pelckmans, K., M. Espinoza, J. De Brabanter, J.A.K. Suykens and B. De Moor (2004*c*). Primal-dual monotone kernel machines. *(Submitted for Publication) Internal Report 04-108, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*.

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

Poggio, T., V. Torre and C. Koch (1985). Computational vision and regularization theory. *Nature* **317**(26035), 314–9.

Rudin, L., S.J. Osher and E. Fatemi (1992). Nonlinear total variation based noise removal algorithms. *Physica D* **60**, 259–268.

Saunders, C., A. Gammerman and V. Vovk (1998). Ridge regression learning algorithm in dual variables. In: *Proc. of the 15th Int. Conf. on Machine learning(ICML'98)*. Morgan Kaufmann. pp. 515–521.

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

Stitson, M., A. Gammerman, V. Vapnik, V. Vovk, C. Watkins and J. Weston (1999). *Advanced in Kernel methods: Support Vector Learning*. Chap. Support vector regression with ANOVA decomposition kernels. The MIT Press. Cambridge Massachusetts.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistics Society Series* **B**(36), 111–147.

Suykens, J. A. K. and J. Vandewalle (1999). Least squares support vector machine classifiers. *Neural Processing Letters* **9**(3), 293–300.

Suykens, J.A.K., G., Horvath, S., Basu, C., Micchelli and J., Vandewalle, Eds.) (2003). *Advances in Learning Theory: Methods, Models and Applications*. Vol. 190 of *NATO Science Series III: Computer & Systems Sciences*. IOS Press Amsterdam.

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

Tikhonov, A. N. and V. Y. Arsenin (1977). *Solution of Ill-Posed Problems*. Winston. Washington DC.

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

von Luxburg, U., O. Bousquet and B. Schölkopf (2004). A compression approach to support vector model selection. *Journal of Machine Learning Research* (5), 293–323.

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

Caption of Table

Table 1: Performances of SVMs, LS-SVMs and the sparse LS-SVM substrates of Subsection 3.2 expressed in Mean Squared Error (MSE) on a test set in the case of regression or Percentage Correctly Classified (PCC) in the case of classification. Sparseness is expressed in percentage of support vectors w.r.t. number of training data. The kernel machines were tuned for the kernel parameter and the respective hyper-parameters C, ϵ; γ and ζ with 10-fold cross-validation. These results indicate that sparse LS-SVM substrates are at least comparable in generalization performance with existing methods, but are often more effective in achieving sparseness.

Captions of Figures

Figure 1: Schematical representation of the hierarchical kernel machine. From a conceptual point of view, one builds upon the LS-SVM substrate which takes care of the classical ingredients of kernel machines. Hyper-parameters guide the interaction between different levels while KKT conditions enforce the individual levels to be defined properly. From a computational point of view, the whole can be fused into one convex optimization problem.

Figure 2: Comparison of the SVM, LS-SVM and sparse LS-SVM substrate of subsection 3.2 on the Motorcycle regression dataset. One can see the difference in selected support vectors of the SVM **(a)** against the ones chosen by the sparse LS-SVM **(b)**.

Figure 3: Comparison of the SVM, LS-SVM and sparse LS-SVM substrate of subsection 3.2 on the ripley classification dataset. One can see the difference in selected support vectors of the SVM **(a)** against the ones chosen by the sparse LS-SVM **(b)**: the former concentrate around the margin, while the sparse LS-SVM substrate will provide a more global support.

Figure 4: Results of structure detection on the an artificial dataset as used in (Vapnik, 1998), consisting of 100 data samples generated by four componentwise non-zero functions of the first 4 inputs and 21 irrelevant inputs and perturbed by i.i.d. unit variance Gaussian noise. This diagram shows the evolution of the maximal variations per component when increasing the hyper-parameter ρ from 1 to 10000. The black indicates a value ρ were the underlying structure is indeed detected successfully while the generalization performance is optimal.

Figure 5: Results of structure detection on the an artificial dataset as used in (Vapnik, 1998). The estimated componentwise LS-SVM (7) with $\rho = 300$ as tuned by cross-validation. All left inputs except the first four have a zero maximal variation as indicated in the 2 lowest sub-plots.

Figure 6: Results of structure detection on the Boston housing dataset consisting of 250 training, 100 validation and 156 randomly selected training samples. The evolution of the maximal variation of the variables when increasing ρ. The arrow indicates the choice of ρ made by the fusion argument minimizing the validation performance (solid line).

Figure 7: Results of structure detection on the Boston housing dataset consisting of 250 training, 100 validation and 156 randomly selected training samples. The contributions of the variables which have a non-zero maximal variation. The fusion argument as described in Subsection 4.3 was used to tune the parameter ρ.

	SVM		LS-SVM	Sparse LS-SVM substr.	
	Perf	**Sparse**	**Perf**	**Perf**	**Sparse**
Sinc	0.0052	68%	0.0045	0.0034	9%
Motorcycle	516.41	83%	444.64	469.93	11%
Ripley	90.10%	33.60%	90.40%	90.50%	4.80%
Pima	73.33%	43%	72.33%	74%	9%

Table 1

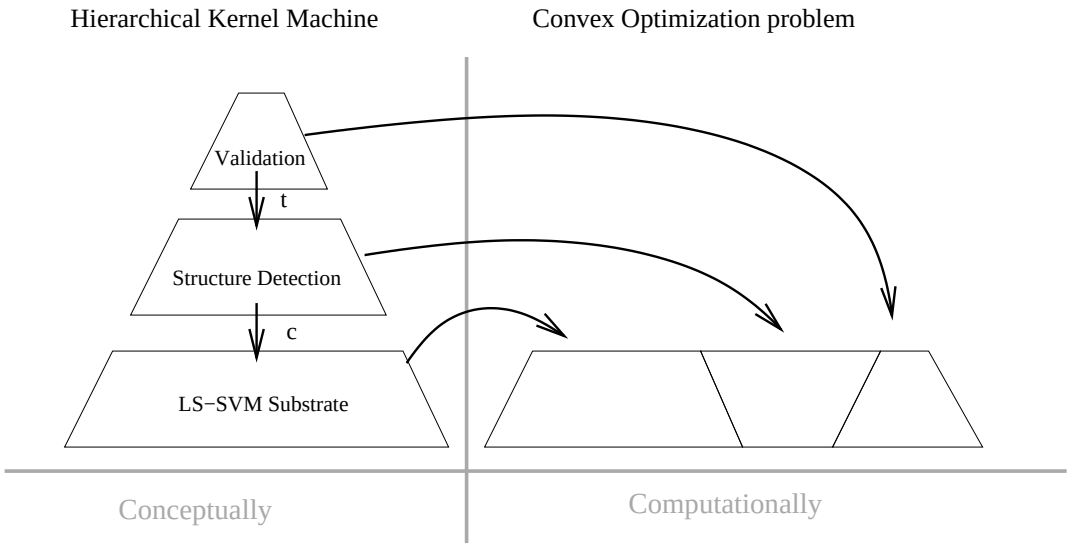
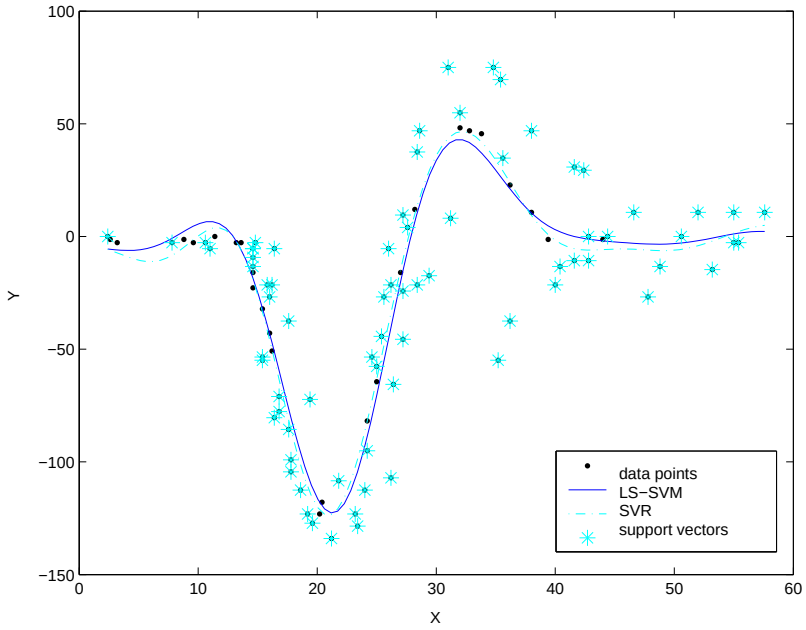
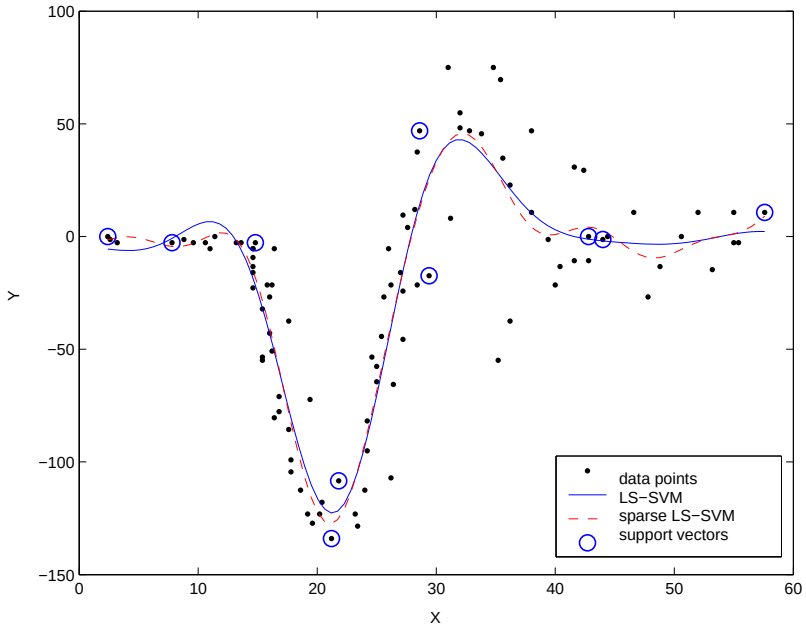

Hierarchical Kernel Machine
Convex Optimization problem
Validation
t
Structure Detection
c
LS−SVM Substrate
Conceptually
Computationally

Fig. 1.

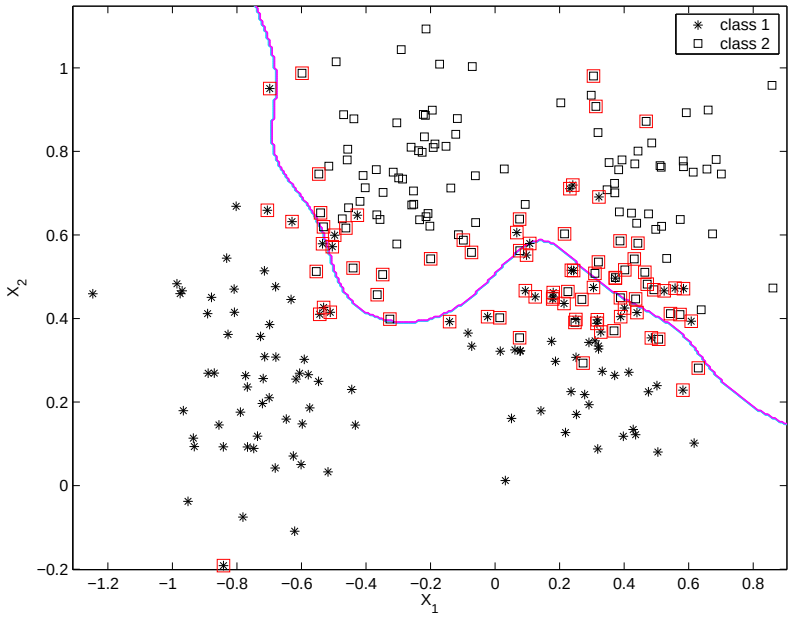


(a) Motorcycle: SVM

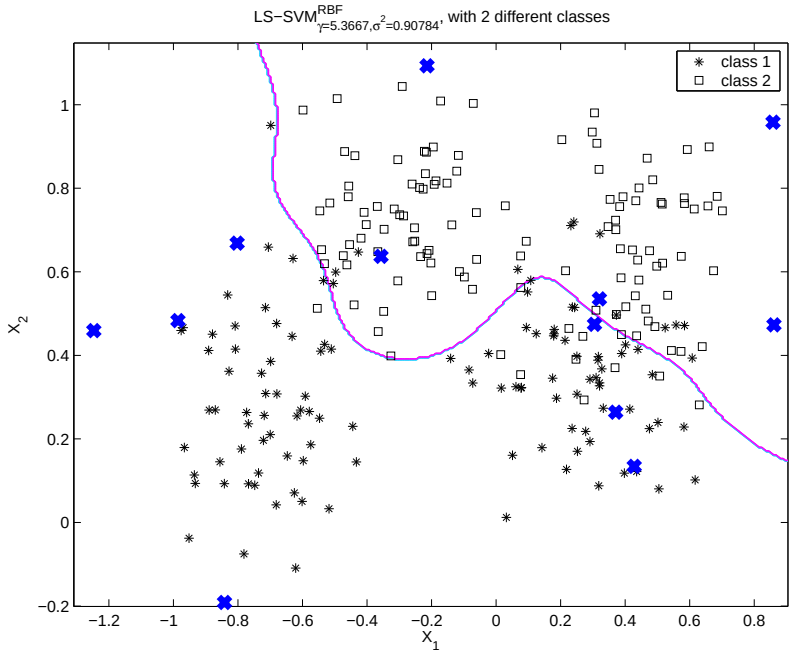


(b) Motorcycle: sparse LS-SVM substrate

Fig. 2.



(a) Ripley dataset: SVM



(b) Ripley dataset: sparse LS-SVM substrate

Fig. 3.

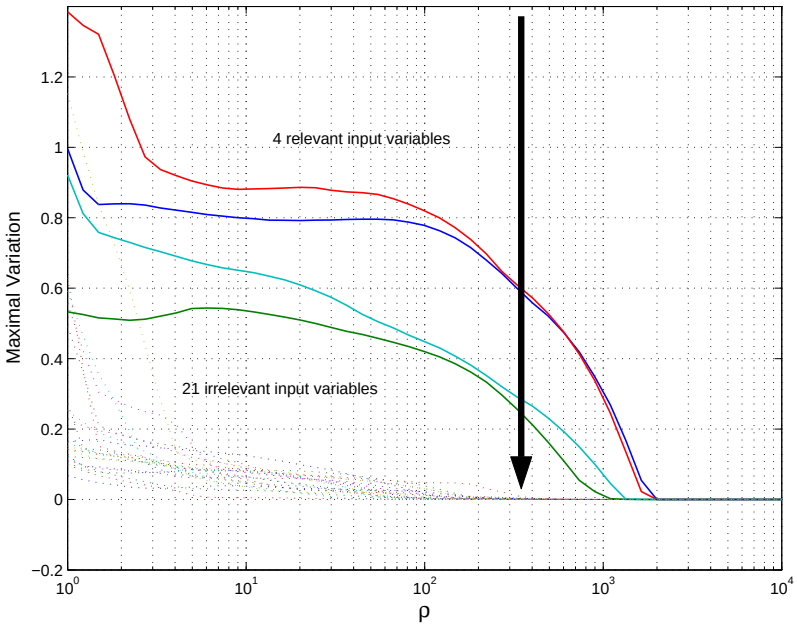

4 relevant input variables
21 irrelevant input variables
Maximal Variation
ρ
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
10^0
10^1
10^2
10^3
10^4

Fig. 4.

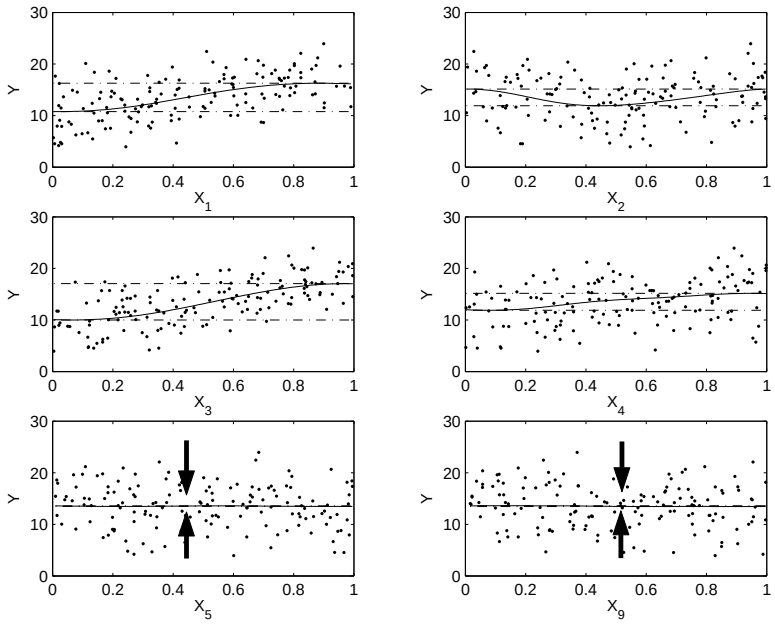


Fig. 5.

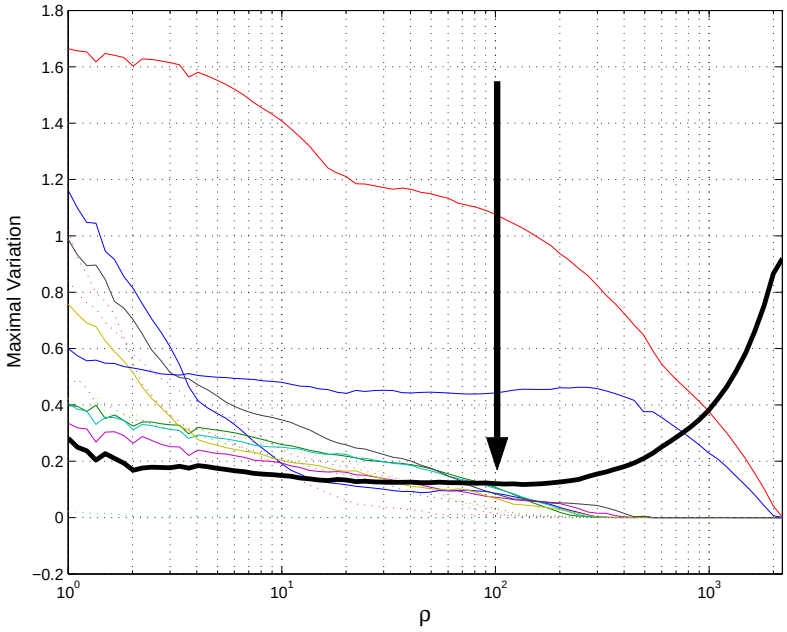


Fig. 6.

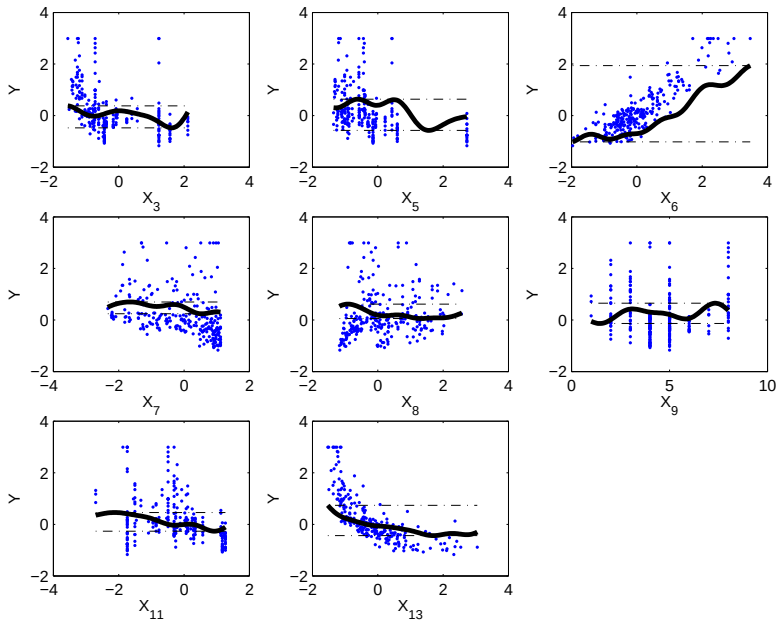


Fig. 7.