Building Sparse Representations and Structure Determination on LS-SVM Substrates

K. Pelckmans [a] J. A.K. Suykens [a] B. 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/

Abstract

This paper proposes a new 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 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 determined at higher levels such that sparse representations and/or structure detection are obtained. Using the necessary aand sufficient conditions for optimality given by the Karush-Kuhn-Tucker conditions, one can guide the interaction between different levels via a well-defined set of hyper-parameters. 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 respectively by using an L_1 regularization scheme and a measure of maximal variation at the second 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, Sparseness.

1 Introduction

The problem of inference of a model based on a finite set of observational data is often 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. So-called additive models consisting of a sum of lower dimensional nonlinearities per component (variable) have become one of the widely used non-parametric techniques as they offer a compromise between the somewhat conflicting requirements of flexibility, dimensionality and interpretability (see e.g. (Hastie *et al.*, 2001)).

The relative importance between the *smoothness* of the solution (as defined in different ways) 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 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 the 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 the following 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 (and even non-smooth) 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 trade-off*. 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 for the inference of model (parameters), regularization constants and model structure. 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). Figure 1 shows the hierarchical structure. The LS-SVM substrate constitutes the basis level of the hierarchy (*level 1*). This paper shows how one can consider additional levels resulting in sparse representations and structure detection by taking suitable cost-functions at level 2. By fusing all levels into one single problem, the resulting hierarchical kernel machine can be obtained by solving a convex optimization problem. Sparse representations are obtained through the use of an L_1 based regularization scheme at level 2 and are motivated by a sensitivity argument

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

from optimization theory. This hierarchy is visualized in Figure 2. Structure detection is expressed at the second level through a measure of maximal variation of a component, leading to the hierarchical kernel machine as represented by Figure 3. The hierarchical kernel model is finalized by an additional third level which tunes the remaining hyper-parameters of the second level 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. Section 3 explains the structure of hierarchical kernel machines based on LS-SVM substrates and employs this idea for obtaining structure detection and LS-SVMs with sparse support values. Section 4 discusses how one can optimize the LS-SVM substrate or the sparse LS-SVM with respect to the performance on a validation set. Section 5 presents numerical results on both artificial and benchmark datasets. This paper is an extended version of the ESANN 2004 contribution (Pelckmans *et al.*, 2004*b*), with a discussion of more efficient methods. A main improvement of this paper is the use of a three level hierarchy instead of a two level as in the previous papers on the subject.

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, $f : \mathbb{R}^d \to \mathbb{R}$ 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)

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

In the sequel we refer to this scheme as standard LS-SVM regressors with Tikhonov regularization. Let $e = (e_1, \ldots, e_N)^T \in \mathbb{R}^N$ denote a vector of residuals. The Lagrangian of the constrained optimization problem becomes $\mathcal{L}_\gamma(w, b, e_i; \alpha_i) =$

$0.5w^Tw + 0.5\gamma\sum_{i=1}^N e_i^2 - \sum_{i=1}^N \alpha_i(w^T\varphi(x_i) + b + e_i - y_i)$. 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 is summarized as

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

where $\Omega \in \mathbb{R}^{N\times N}$ with $\Omega_{ij} = K(x_i, x_j)$, $I_N \in \mathbb{R}^{N\times N}$ denotes the identity matrix and $1_N \in \mathbb{R}^N$ is a vector containing ones. 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

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

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 and $\|\cdot\|_2^2$ denotes the squared Euclidean norm.

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

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

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 mainly on the regression case, although it is applicable also 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 as they offer a compromise between the somewhat conflicting requirements of flexibility, dimensionality and interpretability (see e.g. (Hastie *et al.*, 2001)). This subsection reviews results obtained in (Pelckmans *et al.*, 2004*a*). Let a superscript (p) denote the pth component $x_i^{(p)} = (x_i^{p_1}, \ldots, x_i^{p_{D_p}})^T \in \mathbb{R}^{D_p}$ such that every datasample consists of the union of P different components $x_i = \bigcup\{x_i^{(1)}, \ldots, x_i^{(P)}\}$. In the simplest case $n_p = 1$ such that $x_i^{(p)} = x_i^p$. Consider the following model

$$f(x) = \sum_{p=1}^{P} w_p^T \varphi_p\left(x^{(p)}\right) + b, \tag{6}$$

where $\varphi_p : \mathbb{R} \to \mathbb{R}^{d_f}$ is a possibly infinite dimensional mapping of the pth component. One considers the regularized least squares cost-function as introduced in (Pelckmans *et al.*, 2004*a*),

$$\begin{aligned} \min_{w_p,b,e_i} \mathcal{J}_\gamma(w_p, e) = \frac{1}{2}\sum_{p=1}^{P} w_p^T w_p + \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2 \\ \text{s.t.} \quad \sum_{p=1}^{P} w_p^T \varphi_p(x_i^{(p)}) + b + e_i = y_i, \quad \forall i = 1, ..., N. \end{aligned} \tag{7}$$

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

$$\begin{cases} \left[\begin{array}{c|c} 0 & 1_N^T \\ \hline 1_N & \Omega^P + I_N/\gamma \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \\ \gamma e = \alpha \\ w_p^T \varphi_p\left(x_i^{(p)}\right) = \Omega_i^{(p)} \alpha, \ \ \forall i = 1, \ldots, N, \forall p = 1, \ldots, P, \end{cases} \tag{8}$$

where $\Omega^P = \sum_{p=1}^{P} \Omega^{(p)} \in \mathbb{R}^{N \times N}$ and $\Omega_{ij}^{(p)} = K_p\left(x_i^{(p)}, x_j^{(p)}\right)$ is the kernel evaluated between the pth component of the ith and the jth data-point. A qualitative comparison of the componentwise LS-SVM and 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) was given in (Pelckmans *et al.*, 2004*a*). 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.

3 Hierarchical Kernel Models for Structure Detection and Sparse Representations

3.1 Level 1: Substrate LS-SVMs with additive regularization trade-off

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) which penalizes the error variables in an additive way, i.e.

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

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. Motivations for this approach are seen in the corresponding problem of model selection, while additional interpretations of the additive scheme were given in (Pelckmans *et al.*, 2003) as the constants c_i are also directly influencing 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

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

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 α, b to (10) is uniquely determined. The determination of c is postponed to higher levels. Note at this stage that the extension of this derivation to the case of componentwise models follows straightforwardly.

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 unique solution to (10). The residuals $\hat{f}(x_j^v) - y_j^v$ from the evaluation on a validation dataset are denoted as e_j^v such that one can write

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

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 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 trade-off scheme does not replace the Tikhonov regularization scheme, but parameterizes the trade-off in a different way. This 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. The aim of this paper is to make use of this mechanism in order to obtain sparse models and performing structure detection.

3.2 Level 2: sparse hierarchical kernel machines

Sparseness is often regarded as good practice in the machine learning community (Vapnik, 1998; von Luxburg *et al.*, 2004) as it gives an optimal solution with a minimal representation (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 the corresponding constraints are perturbed. This perturbation is proportional to c as can be seen from the constraints in (9). As such, one can write

$$\hat{\alpha}_i = -\frac{\partial p^*}{\partial c_i} \quad \text{and} \quad p^* = \inf_{w,b} \mathcal{J}_c \quad \text{s.t. (9) holds,} \tag{13}$$

as explained e.g. in (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 on the Lagrange multipliers. The 1-norm is considered

$$\min_{e,\alpha,b,c} \|e\|_2^2 + \zeta \|\alpha\|_1 \quad \text{s.t. (10) holds and } \alpha + c = e. \tag{14}$$

where $0 < \zeta \in \mathbb{R}$ 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. The method in this paper is also treated at different hierarchical levels.

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 components, one may ask oneself which components have no contribution ($f_p(\cdot) = 0$) 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.

Previous results (Pelckmans *et al.*, 2004*a*) were obtained using a 1-norm of the contribution (i.e. $\sum_{i=1}^{N} |f_p(x_i^{(p)})|$) which is somewhat similar to the measure of total variation (Rudin *et al.*, 1992). This paper adopts a measure of maximal variation of components, defined as

$$t_p = \max_i \left| w_p^T \varphi_p \left(x_i^{(p)} \right) \right|, \quad \forall p = 1, \ldots, P. \tag{15}$$

Indeed, if t_p were zero, the corresponding component would not be able to contribute to the final model and $f_p(\cdot) = 0$. 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 than the method proposed in e.g. (Pelckmans *et al.*, 2004*b*). The kernel machine for structure detection minimizes the following criterion for a given tuning constant $0 < \rho \in \mathbb{R}$:

$$\min_{e,t_p,\alpha,b,c} \|e\|_2^2 + \rho \sum_{p=1}^{P} t_p \text{ s.t. (10) holds and} \\ -t_p 1_N \leq \Omega^{(p)} \alpha \leq 1_N t_p, \ \forall p = 1, \ldots, P, \tag{16}$$

where the contribution of the individual components are described as $w_p^T \varphi_p \left(x_i^{(p)} \right) = \Omega_i^{(p)} \alpha$. It is known that the use of 1-norms may lead to sparse solution which is 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 which have been implemented in a 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 in order to detect structure, while the final predictions can be made based on a standard model using only the selected components (compare to basis pursuit, see e.g. (Chen *et al.*, 2001)). In this paper, one works with a standard componentwise LS-SVM of Subsection 2.2 based on the selected structure.

4 Fusion of Validation with Previous Levels

The automatic tuning of hyper-parameters ζ or ρ with respect to the model selection criterion is highly desirable, especially in practice. The interplay between exact training and optimizing the regularization trade-off with respect to a validation criterion is studied. While fusion of a standard LS-SVM is often non-convex (Pelckmans *et al.*, 2003), the additive regularization trade-off scheme does circumvent this problem. Fusion of the validation part and the LS-SVM substrate was considered in (Pelckmans *et al.*, 2003)

$$\min_{\alpha,b,c,e,e^v} \|e\|_2^2 + \|e^v\|_2^2 \text{ s.t. (10) holds and } e = \alpha + c. \tag{17}$$

This criterion is motivated by the assumption that the training as well as the validation data are independently sampled from the same distribution. In order to confine the effective degrees of freedom of the space of possible c values, in (Pelckmans *et al.*, 2004*b*) the use of the following modified cost-function was proposed in order to obtain sparse solutions:

$$\min_{e,e^v,\alpha,c,b} \|e\|_2^2 + \|e^v\|_2^2 + \xi\|\alpha\|_1 \text{ s.t. (10) holds and } e = \alpha + c. \tag{18}$$

While this problem is well defined, one may object that a new hyper-parameter $0 < \xi \in \mathbb{R}$ has popped up. (Pelckmans *et al.*, 2004*b*) considered the problem at two levels. In this paper, a three level approach is introduced.

The following section extends these results by proposing a way to tune the hyper-parameter ρ in (16) and ζ in (14) of the second level with respect to a validation criterion using a third level of inference. This three level architecture constitutes the hierarchical kernel machine. Figure 1 gives a schematical representation of the conceptual idea of hierarchical kernel machines. The LS-SVM substrate constitute the first level, while the sparse LS-SVM and the LS-SVM for structure detection makes up the second level. The validation performance is used to tune the hyper-parameters ζ (or ρ) on a third level. For notational convenience, the bias term b is omitted from the derivations in the sequel.

4.1 Level 3: Fusion of sparse LS-SVMs with validation

Consider the conceptual three level hierarchical kernel machine for sparse LS-SVMs (see Figure 1). Figure 2 outlines the derivations of the hierarchical kernel machine for sparse representations and emphasizes the three level hierarchical structure.

Consider the level 2 cost-function (14) where ζ acts as a hyper-parameter. Necessary and sufficient conditions for the global optimum of the cost-function (14) are

given by the Karush-Kuhn-Tucker conditions. The Lagrangian becomes

$$\mathcal{L}_\zeta(\alpha, a; \xi^+, \xi^-) = \frac{1}{2}\left\|\Omega\alpha - y\right\|_2^2 + \rho\sum_{i=1}^{N} a_i + \xi^{-T}\left(-a-\alpha\right) + \xi^{+T}\left(-a+\alpha\right), \quad (19)$$

with positive multipliers $\xi^+, \xi^- \in \mathbb{R}^N$. The corresponding Karush-Kuhn-Tucker conditions are necessary and sufficient as the primal problem is convex (Boyd and Vandenberghe, 2004), p.244, for the determination of the global optimum:

$$\begin{cases} \Omega^T\Omega\alpha - y^T\Omega = (\xi^- - \xi^+), & (a) \\ \zeta = 1_N^T(\xi^- + \xi^+) & (b) \\ \xi^+, \xi^- \geq 0, & (c) \\ -a \leq \alpha \leq a, & (d) \\ \xi_i^-(a_i + \alpha_i) = 0, \qquad \forall i = 1, \ldots, N & (e) \\ \xi_i^+(a_i - \alpha_i) = 0, \qquad \forall i = 1, \ldots, N & (f) \end{cases} \quad (20)$$

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-parameter ζ:

$$\min_{\zeta, a, \alpha, \xi^-, \xi^+} \mathcal{J}^v = \frac{1}{2}\|\Omega^v\alpha - y^v\|_2^2 \quad \text{s.t.} \quad (20) \text{ hold}, \quad (21)$$

Except for conditions (e) and (f), the problem (21) can be solved as a Quadratic Programming (QP) problem. We propose the use of the modified QP below instead, which results in the same optimal solution as (21):

$$\min_{\zeta, a, \alpha, \xi^-, \xi^+} \mathcal{J}^v = \|\Omega^v\alpha - y^v\|_2^2 + b^-\xi^{-T}\left(a+\alpha\right) + b^+\xi^{+T}\left(a-\alpha\right) \quad \text{s.t. conditions (20.abcd) holds.} \quad (22)$$

by proper selection of the weighting terms $b^+ > 0$ and $b^- > 0$ such that the Hessian of the QP remains positive semidefinite and the complementary slackness conditions (20.e) and (20.f) are satisfied.

4.2 Level 3: Fusion of structure detection with validation

A third level is added to the LS-SVM for structure detection in order to tune the hyper-parameter ζ of the second level. Figure 3 summarizes the derivation below and points out the hierarchical approach.

Let $t = (t_1, \ldots, t_P)^T \in \mathbb{R}^P$ be a vector of bounds on the maximal variation per component. Consider the optimization problem (16) where ρ acts as a hyper-parameter. One can eliminate e and c from this optimization problem leading to

$$\min_{t,\alpha} \mathcal{J}_\rho(\alpha, t) = \frac{1}{2}\|\Omega^P\alpha - y\|_2^2 + \rho\sum_{p=1}^{P} t_p \quad \text{s.t.} \quad -t_p 1_N \leq \Omega^{(p)}\alpha \leq t_p 1_N, \ \forall p = 1, \ldots, P. \tag{23}$$

The Lagrangian becomes

$$\begin{aligned}\mathcal{L}_\rho(\alpha, t; \xi^{+p}, \xi^{-p}) = \frac{1}{2}\left\|\Omega^P\alpha - y\right\|_2^2 + \rho\sum_{p=1}^{P} t_p \\ + \sum_{p=1}^{P} \xi^{-pT}\left(-t_p 1_N - \Omega^{(p)}\alpha\right) + \sum_{p=1}^{P} \xi^{+pT}\left(-t_p 1_N + \Omega^{(p)}\alpha\right),\end{aligned} \tag{24}$$

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

$$\begin{cases} \Omega^{PT}\Omega^P\alpha - y^T\Omega^P = \sum_{p=1}^{P}(\xi^{-p} - \xi^{+p}), & (a) \\ \rho = 1_N^T(\xi^{-p} + \xi^{+p}) & \forall p = 1, \ldots, P \ \ (b) \\ \xi^{+p}, \xi^{-p} \geq 0, & \forall p = 1, \ldots, P \ \ (c) \\ -t_p 1_N \leq \Omega^{(p)}\alpha \leq t_p 1_N, & \forall p = 1, \ldots, P \ \ (d) \\ \xi_i^{-p}(t_p + \Omega_i^{(p)}\alpha) = 0, \ \ \forall i = 1, \ldots, N, & \forall p = 1, \ldots, P \ \ (e) \\ \xi_i^{+p}(t_p - \Omega_i^{(p)}\alpha) = 0, \ \ \forall i = 1, \ldots, N, & \forall p = 1, \ldots, P \ \ (f) \end{cases} \tag{25}$$

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-parameter ρ:

$$\min_{\rho,t,\alpha,\xi^-,\xi^+} \mathcal{J}^v = \frac{1}{2}\|\Omega^{P,v}\alpha - y^v\|_2^2 \quad \text{s.t.} \quad (25) \text{ hold}, \tag{26}$$

where $\Omega^{P,v} \in \mathbb{R}^{n\times N} = \sum_{p=1}^{P}\Omega^{(p),v}$ and $\Omega_{ij}^{(p),v} = K^p\left(x_i^{(p)}, x_j^{(p),v}\right)$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, N$. Except for conditions (e) and (f), the problem (26) can be solved as a Quadratic Programming (QP) problem. We propose the use of the modified QP below instead, which results in the same optimal solution as (26):

$$\begin{aligned}\min_{\rho,t,\alpha,\xi^-,\xi^+} \mathcal{J}^v = \|\Omega^{P,v}\alpha - y^v\|_2^2 + \sum_{p=1}^{P} b_p^- \xi^{-pT}\left(t_p 1_N + \Omega_i^{(p)}\alpha\right) \\ + \sum_{p=1}^{P} b_p^+ \xi^{+pT}\left(t_p 1_N - \Omega_i^{(p)}\alpha\right) \quad \text{s.t. conditions (25.abcd) holds.}\end{aligned} \tag{27}$$

Crucial in this formulation is the observation that the complementary slackness terms $\xi^{-p^T}\left(t_p 1_N + \Omega_i^{(p)}\alpha\right)$ and $\xi^{+p^T}\left(t_p 1_N - \Omega_i^{(p)}\alpha\right)$ are bounded from below by zero as all cross-product terms are positive. Indeed if the minimum value 0 of the complementary slackness is attained in the global optimum, the solution to (27) would equal the solution to (26). For a proper choice of the weighting terms b_i, this will be attained. In order to keep the problem convex, the values of b_i should be chosen such that the Hessian of the quadratical programming problem remains positive definite. One can tune this weighting terms by checking the complementary slackness conditions (25.e) and (25.f) on the resulting solution. Similar as is the case for componentwise LS-SVM substrates, the model can be evaluated at new data points $x_* \in \mathbb{R}^d$ as

$$\hat{f}(x^*) = \hat{w}^T\varphi(x^*) = \sum_{i=1}^{N}\hat{\alpha}_i \sum_{t_p \neq 0} K^p\left(x_i^{(p)}, x_*^{(p)}\right), \tag{28}$$

where $\hat{\alpha}$ is the solution to (27).

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 $Y = \mathrm{sinc}(X) + e$ 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 (see Figure 4), the artificial Ripley dataset ($N = 250$, $d = 2$) (see Figure 5) and the PIMA dataset ($N = 468$, $d = 8$) from UCI at classification problems. The models resulting from sparse LS-SVM substrates were tested against the standard 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

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 from Subsection 3.3, 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 via a naive line-search as well as using the method which is described in Subsection 4.2.

Figure 6 and 7 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:

$$f(x) = 10 \sin(X^1) + 20 \ (X^2 - 0.5)^2 + 10 \ X^3 + 5 \ X^4 \tag{29}$$

such that $y_i = f(x_i) + e_i$ with $e_i \sim \mathcal{N}(0, 1)$ for all $i = 1, \ldots, 100$. Figure 7 gives the nontrivial components ($t_p > 0$) associated with the LS-SVM substrate with ρ optimized in validation sense. Figure 6 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 using the mechanism of maximal variation for detection of which input variables have a non-trivial contribution. 250 data-points 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 3.3 was used to construct the maximal variation evolution diagram, see Figure 8. Figure 9 displays the contributions of the individual components. 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 4.2. 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.

Testing was done by retraining a componentwise LS-SVM based on only the selected inputs. The resulting additive model increases in performance expressed in MSE on an independent test-set with 22%. The improvement is even more significant (32%) with respect to a standard 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 exploiting the 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 separated 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 proposed method both with respect to interpretability as well as generalization performance.

Acknowledgments Research supported by Research Council KUL: GOA-Mefisto 666, GOA AMBioRICS, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects, G.0240.99 (multilinear algebra), G.0197.02 (power islands), G.0141.03 (identification and cryptography), G.0491.03 (control for intensive care glycemia), G.0120.03 (QIT), G.0452.04 (new quantum algorithms), G.0407.02 (support vector machines), G.0499.04 (robust statistics), G.0211.05 (nonlinear identification), G.0080.01 (collective behaviour) AWI: Bil. Int. Collaboration Hungary/ Poland; research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants,GBOU (McKnow); Belgian Federal Science Policy Office: IUAP P5/22 (‘Dynamical Systems and Control: Computation, Identification and Modelling’, 2002-2006),IUAP V; PODO-II (CP/40: TMS and Sustainability); EU: FP5-Quprodis; ERNSI; Eureka 2063-IMPACT; Eureka 2419-FliTE; Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard.

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 additive 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. Chapter in *Support Vetor Machines: Theory and Applications*, (ed. Wang L.), Springer.

Pelckmans, K., J.A.K. Suykens and B. De Moor (2003). Additive regularization trade-off: 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 12th 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. *Proceedings of the IEEE* **78**(9), 1481–1497.

Poggio, T., V. Torre and C. Koch (1985). Computational vision and regularization theory. *Nature* **317**(6035), 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., Horvath G., Basu S., Micchelli C. and Vandewalle J. (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 design of hierarchical kernel machines. 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: Schematical representation of the hierarchical kernel machine for sparse representations. From a conceptual point of view, inference is done at different levels and interaction is guided via a set of hyper-parameters. The first level constitutes of an LS-SVM substrate. On the second level, inference of the c_i is defined in terms of a cost function inducing sparseness, while ζ is optimized on a third level using a validation criterion.

Figure 3: Schematical representation of the hierarchical kernel machine for structure detection. On the second level, inference of the c_i is expressed in terms of a least squares cost function with a minimal amount of maximal variation, while ρ is optimized on a third level using a validation criterion.

Figure 4: Comparison of the SVM, LS-SVM and sparse LS-SVM substrate of subsection 3.2 on the Motorcycle regression dataset. One sees the difference in selected support vectors of **(a)** a standard SVM and **(b)** a sparse hierarchical kernel machine.

Figure 5: 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 **(a)** a standard SVM and **(b)** a sparse hierarchical kernel machine. The support vectors of the former concentrate around the margin while the sparse hierarchical kernel machine will provide a more global support.

Figure 6: Results of structure detection on 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 arrow indicates a value ρ corresponding with a minimal cross-validation performance. Note that for the corresponding value of ρ, the underlying structure is indeed detected successfully.

Figure 7: Results of structure detection on an artificial dataset as used in (Vapnik, 1998). The estimated componentwise LS-SVM (7) with $\rho = 300$ as tuned by cross-validation. All inputs except the four first are irrelevant to the problem. The structure detection algorithm detects the irrelevant inputs by measuring a maximal variation of zero (as indicated by the black arrows).

Figure 8: Results of structure detection on the Boston housing dataset consisting of 250 training, 100 validation and 156 randomly selected testing 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 9: Results of structure detection on the Boston housing dataset consisting of 250 training, 100 validation and 156 randomly selected testing samples. The contributions of the variables which have a non-zero maximal variation are shown. The fusion argument as described in Subsection 4.2 was used to tune the parameter ρ.

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

Table 1

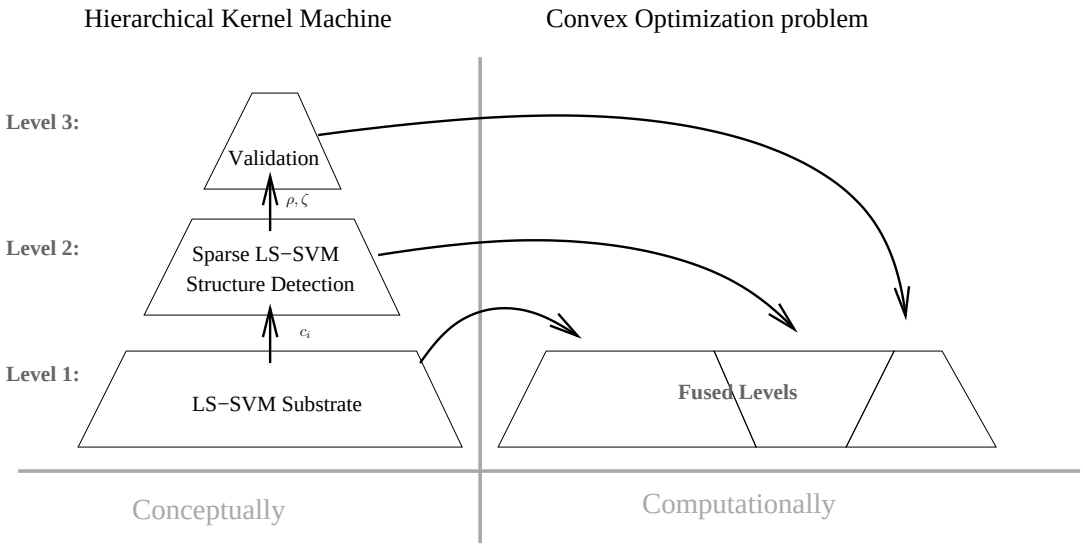
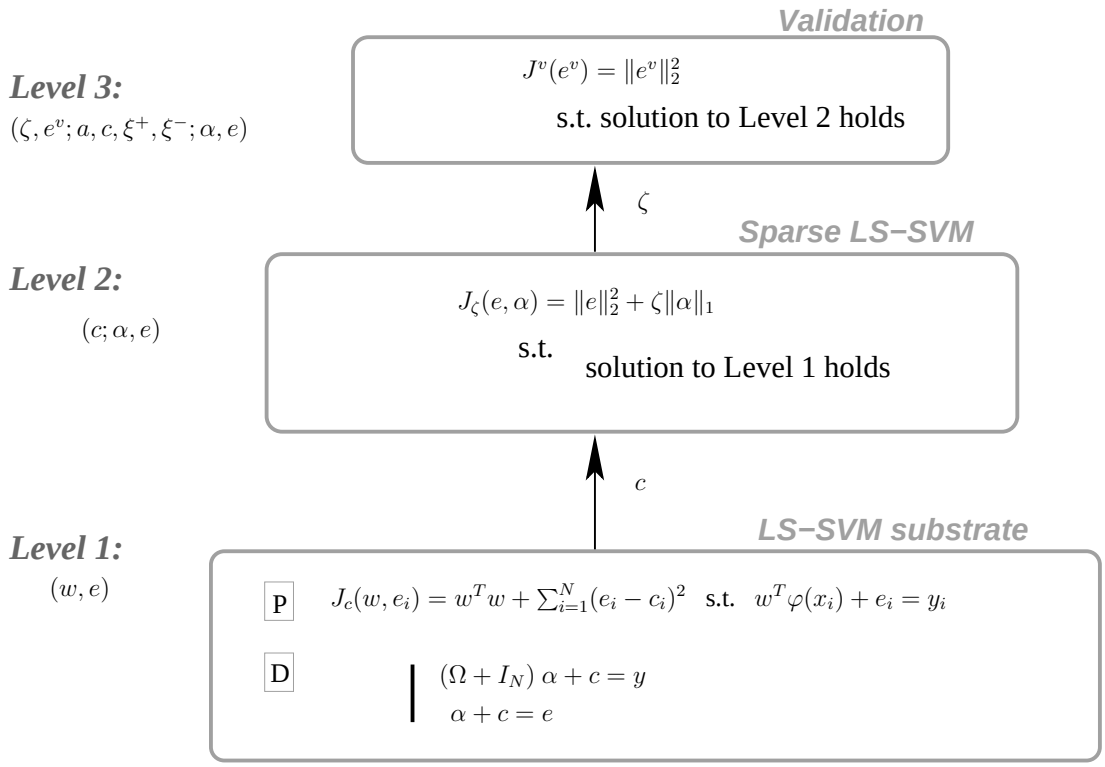

Hierarchical Kernel Machine
Convex Optimization problem
Level 3:
Validation
ρ, ζ
Level 2:
Sparse LS−SVM
Structure Detection
c_i
Level 1:
LS−SVM Substrate
Fused Levels
Conceptually
Computationally

Fig. 1.

Conceptually: hierarchical kernel machines for sparse LS−SVMs



Computationally

(Convex Optimization Problem obtained after Fusion)

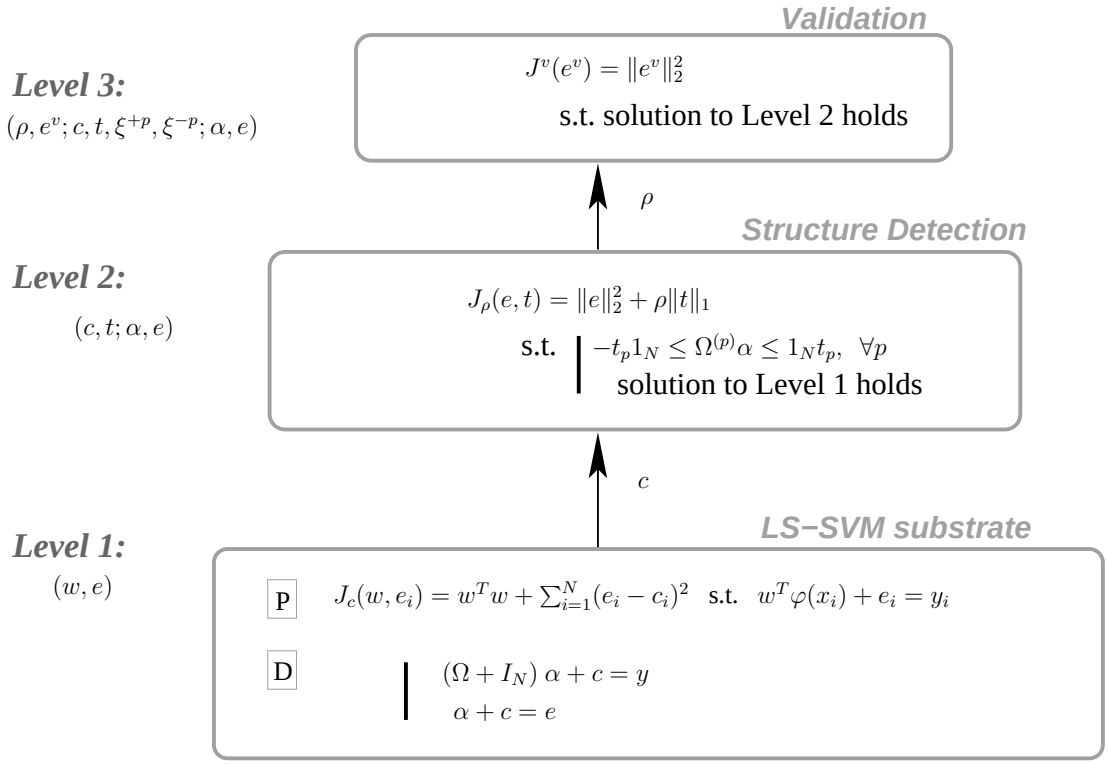
$$\min_{\zeta,\xi^+,\xi^-,a;\alpha} \|\Omega\alpha - y^v\|^2 + b^-\xi^{-T}(a - \alpha) + b^+\xi^{+T}(a + \alpha)$$

s.t.

$$\Omega^{PT}\Omega^P\alpha - y^T\Omega^P = (\xi^- - \xi^+)$$
$$\xi^-, \xi^+ \geq 0$$
$$-a \leq \alpha \leq a$$
$$\zeta = 1_N^T(\xi^- + \xi^+)$$

Fig. 2.

Conceptually (Hierarchical kernel machine for Structure Detection)



Computationally

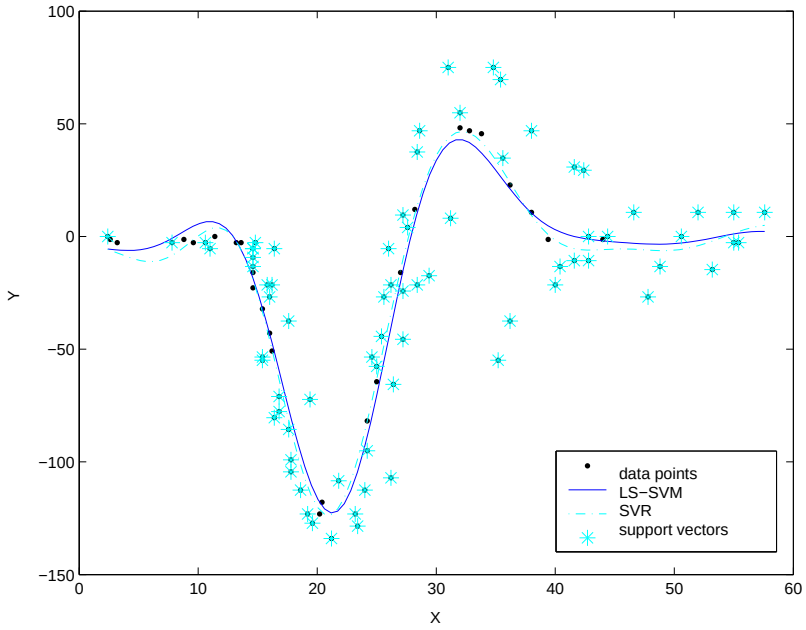
(Convex Optimization Problem obtained after Fusion)

$$\min_{\rho,\xi^+,\xi^-,t,\alpha} \|\Omega^{P,v}\alpha - y^v\|^2 + \sum_{p=1}^{P}\left[b_p^- \xi^{-pT}\left(t_p 1_N - \Omega^{(p)}\alpha\right) + b_p^+ \xi^{+pT}\left(t_p 1_N + \Omega^{(p)}\alpha\right)\right]$$

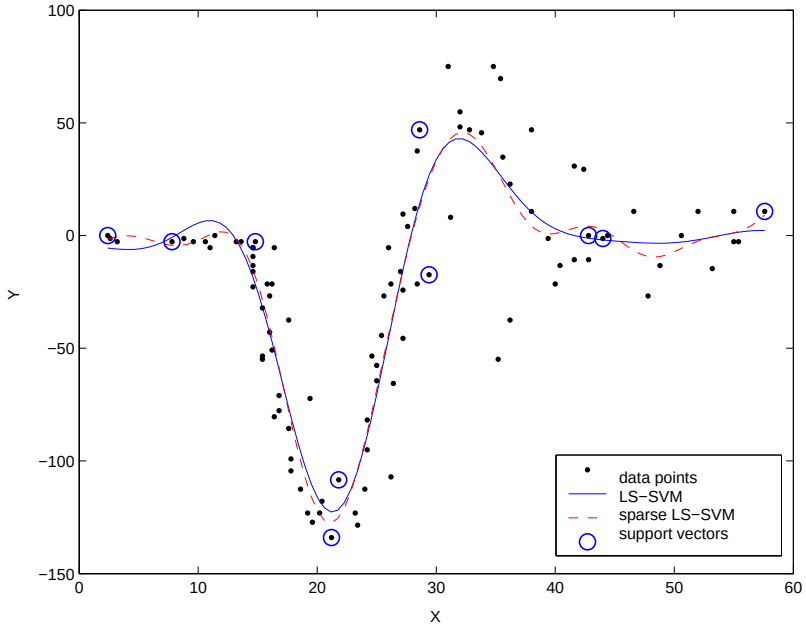
s.t.

$$\begin{aligned}
&\Omega^{PT}\Omega^P\alpha - y^T\Omega^P = \sum_{p=1}^{P}(\xi^{-p} - \xi^{+p}) \\
&\xi^{-p}, \xi^{+p} \geq 0, \ \forall p \\
&-t_p 1_N \leq \Omega^{(p)}\alpha \leq 1_N t_p, \ \forall p \\
&\rho = 1_N^T(\xi^{-p} + \xi^{+p}), \ \forall p
\end{aligned}$$

Fig. 3.

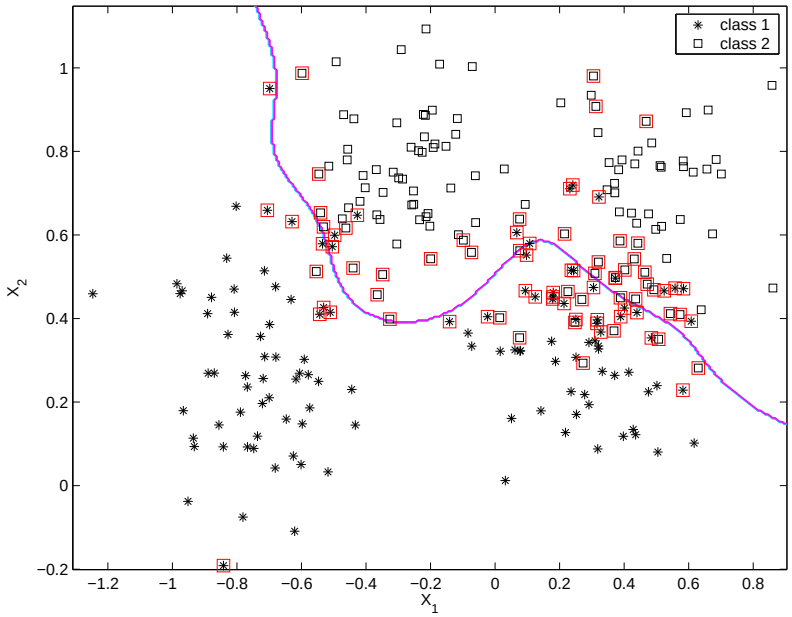


(a) Motorcycle: SVM

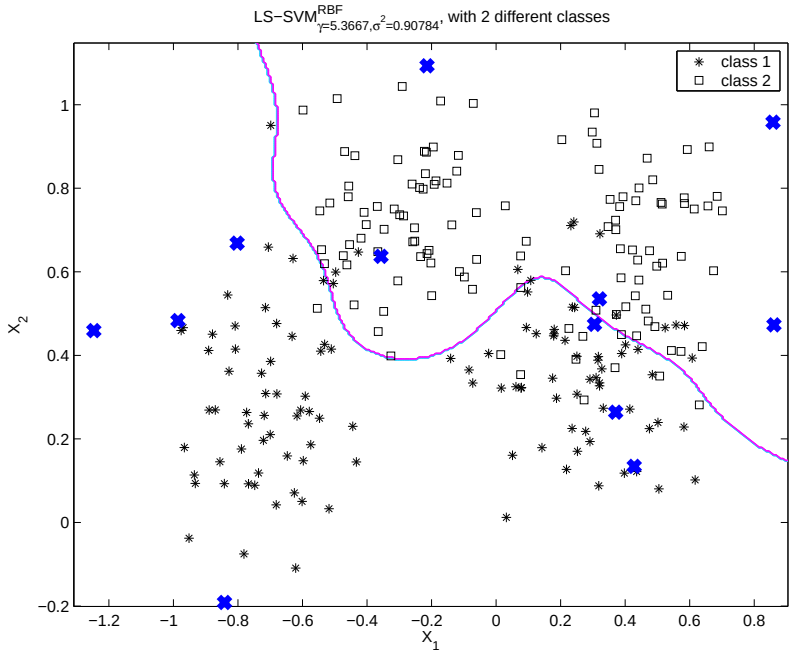


(b) Motorcycle: sparse LS-SVM substrate

Fig. 4.



(a) Ripley dataset: SVM



(b) Ripley dataset: sparse LS-SVM substrate

Fig. 5.

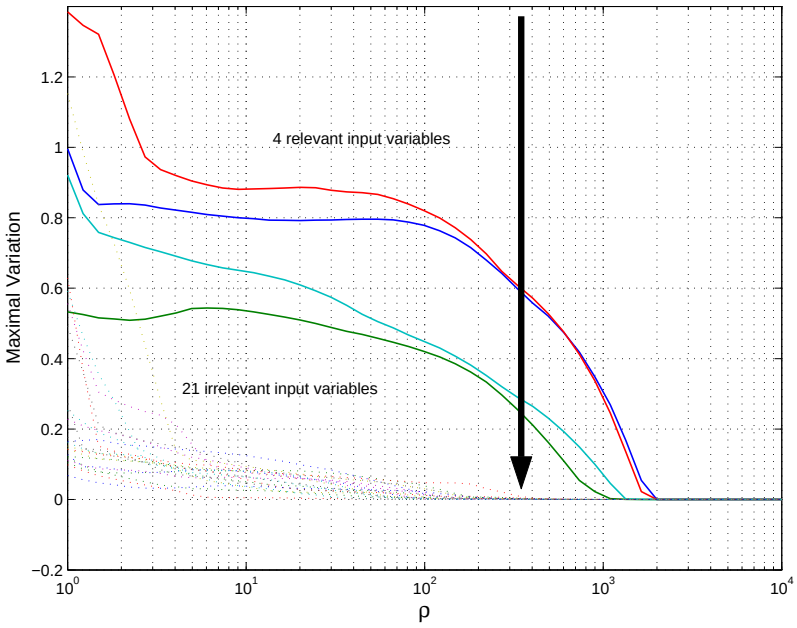

4 relevant input variables
21 irrelevant input variables
Maximal Variation
ρ
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
10^0
10^1
10^2
10^3
10^4

Fig. 6.

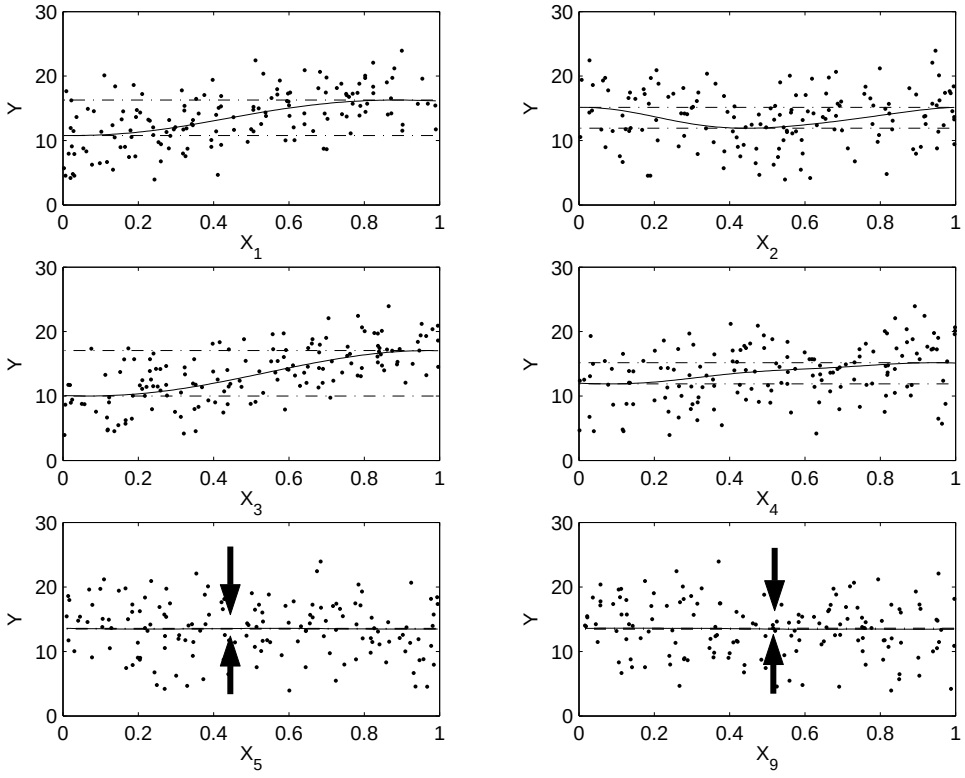


Fig. 7.

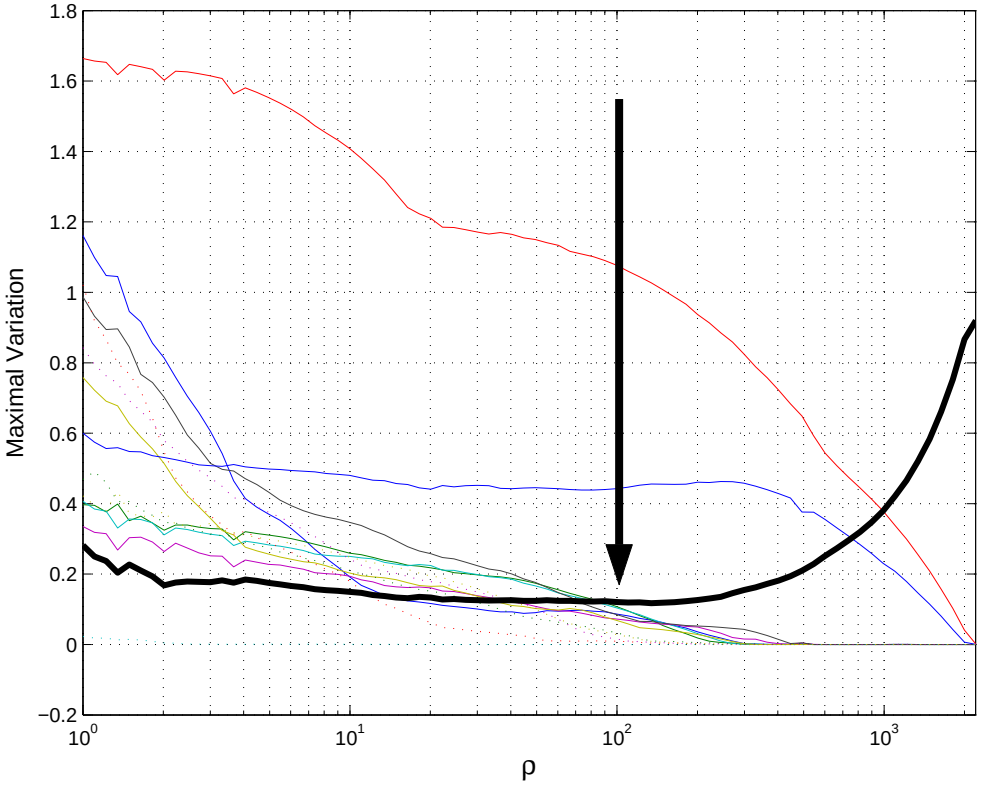


Fig. 8.

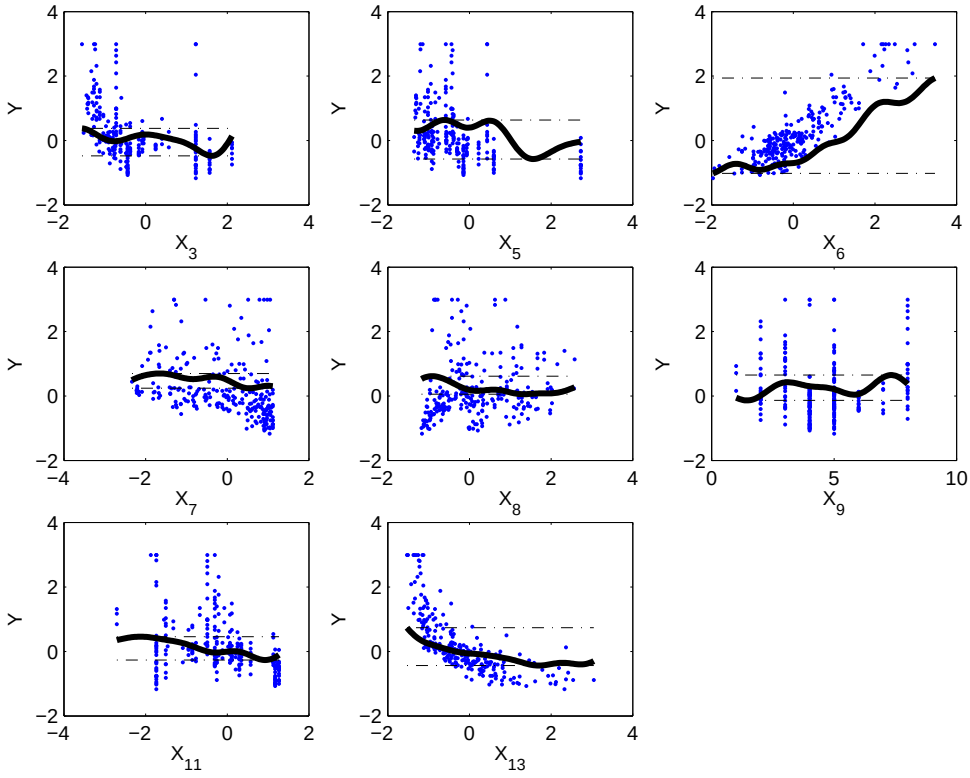


Fig. 9.