Maximal Variation and Missing Values for Componentwise Support Vector Machines

K. Pelckmans, J.A.K. Suykens, B. De Moor
ESAT - SCD/SISTA
Katholieke Universiteit Leuven
Kateelpark Arenberg 10, B-3001 Leuven, Belgium
E-mail: {kristiaan.pelckmans, johan.suykens}@esat.kuleuven.ac.be

J. De Brabanter
Hogeschool KaHo Sint-Lieven
(Associatie KULeuven)
Departement Industrieel Ingenieur
B-9000 Gent, Belgium

***Abstract*— This paper proposes primal-dual kernel machine classifiers based on worst-case analysis of a finite set of observations including missing values of the inputs. Key ingredients are the use of a componentwise Support Vector Machine (cSVM) and an empirical measure of maximal variation of the components to bound the influence of the component which can not be evaluated due to missing values. A regularization term based on the L_1 norm of the maximal variation is used to obtain a mechanism for structure detection in that context. An efficient implemtation using the hierarchical kernel machines framework is elaborated.**

I. Introduction

This work extends recent advances on primal-dual kernel machines for fitting additive models [11] where the primal-dual optimization point of view [2] (as exploited by SVMs [23] and LS-SVMs [20], [21]) is seen to provide an efficient implementation [14]. Although relations exist with results on ANOVA kernels [23], [10], the optimization framework established a solid foundation for extensions towards structure detection similar to LASSO [22] and bridge regression [5], [9], [1] in the context of regression as elaborated in [14], [16]. The key idea was to employ a measure of maximal variation (as defined in the sequel) for the goal of regularization. These componentwise primal-dual kernel machines enable to search through a variety of variables or combinations of these (referred to as components) while enforcing the contributions as much as possible to zero (structure detection). Moreover, the measure of maximal variation can be used to bound the influence of missing values of an observation in the model. As such, a worst-case approach to the problem of missing values can be incorporated during the training stage.

Dealing with missing values is a common problem in statistical analysis. The formal study of missing data began with [17], though ad hoc procedures for dealing with missing observations in standard models were widely used much earlier. A standard reference is [13]. A difference is made on the nature (cause) of the value missing. In this paper, we only consider values that are missing completely at random [18]. Classical methods for handling missing data are to discard the samples that contain missing values or to infer (impute) suitable values for the missing data. More recently general algorithms such as the Expectation-Maximization (EM) [4] and data imputation and augmentation procedures [18], combined with powerful computing resources have largely provided a solution to this aspect of the problem. The different methods proposed in the sequel rely on the learning algorithm itself to deal with missing values in its training phase. The same setting was adopted for the regression spline method MARS, the decision tree CART [6] and PRIM [7]. Black box methods such as neural networks and SVMs are quite useful in predictive settings but are considered less useful for handling data with missing values (See e.g. [12], Table 10.1.). With this paper, we propose a way to overcome this disadvantage by using the framework of SVMs and primal-dual kernel machines.

The paper is organized as follows. Section II provides the linear parametric derivation of SVMs using the measure of maximal variation and extensions to missing values. Section III describes the nonlinear (kernel) version and an hierarchical approach which is computational much more feasible. Section IV reports on empirical results illustrating the method.

II. Basic Ingredients

A. Additive Large Margin Classifiers

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{R}$ be the training data with inputs $x_i \in \mathbb{R}^D$ and outputs $y_i \in \{-1, 1\}$. Consider the classifier $\text{sign}[f(x_i)]$ where $x_1, \ldots, x_N$ are deterministic points and $f : \mathbb{R}^D \to \mathbb{R}$ is an unknown real-valued smooth mapping. Let $x_i^{(p)} = \left(x_i^{p_1}, \ldots, x_i^{p_{D_p}}\right)^T \in \mathbb{R}^{D_p}$ denote the pth component of the ith sample with $p = 1, \ldots, P$ (in the simplest case $n_p = 1$ such that $x_i^{(p)} = x_i^p$). Let this be denoted as $x_i \doteq \left(x_i^{(1)}, \ldots, x_i^{(P)}\right)$.

Definition II.1 [Additive Classifier] *Let $x \in \mathbb{R}^D$ be a point with components $x \doteq \left(x^{(1)}, \ldots, x^{(P)}\right)$. Consider the classification rule in componentwise form [11]*

$$\text{sign}[f(x)] = \text{sign}\left[\sum_{p=1}^{P} f_p\left(x^{(p)}\right) + b\right], \tag{1}$$

with sufficiently smooth mappings $f_p : \mathbb{R}^{D_p} \to \mathbb{R}$ such that the decision boundary is described as in [23], [19]

$$\mathcal{H}_f = \left\{ x_0 \in \mathbb{R}^D \;\middle|\; \sum_{p=1}^{P} f_p(x_0^{(p)}) + b = 0, x_0 \in \mathbb{R}^P \right\}. \tag{2}$$

It is well known that the distance of any point x the hyperplane $\mathcal{H}_{f_p}$ is given as

$$d\left(x,\mathcal{H}_f\right)=\frac{|f(x)|}{\|f'(x)\|}\geq\frac{y_i\left(\sum_{p=1}^{P}f_p\left(x^{(p)}\right)+b\right)}{\sum_{p=1}^{P}\|f^{(p)'}(x^{(p)})\|}, \quad (3)$$

as $\left\|\sum_{p=1}^{P}f^{(p)'}\left(x^{(p)}\right)\right\|\leq\sum_{p=1}^{P}\left\|f^{(p)'}\left(x^{(p)}\right)\right\|$ due to the triangle inequality. The optimal separating hyperplane can be expressed as the model (3) solving

$$\max_{C\geq 0,f_p,b}C\quad\text{s.t.}\quad d(x_i,\mathcal{H}_{f_p})\geq C. \quad (4)$$

After the change of variables in the function f such that $C\sum_{p=1}^{P}\|f^{(p)'}\|=1$ and the application of the lower-bound (3), one obtains equivalently $\min_{f_p,b}\sum_{p=1}^{P}\|f^{(p)'}\|$ such that $y_i\left[\sum_{p=1}^{P}f_p(x_i^{(p)})+b\right]\geq 1$ for all $i=1,\ldots,N$. Now the size of the margin is given as $C=1/\sum_{p=1}^{P}\|f^{(p)'}\|$. In case the data points are not perfectly separable (by elements of the model class (3)), the Hinge loss function is often employed to admit overlapping distributions of the two classes (as motivated e.g. in the theory of SVMs), leading to the cost function

$$\mathcal{J}_\gamma(f)=\sum_{p=1}^{P}\|f^{(p)'}\|+\gamma\sum_{i=1}^{N}\left[1-y_i\left(\sum_{p=1}^{P}f_p\left(x_i^{(p)}\right)+b\right)\right]_+, \quad (5)$$

where the function $[\cdot]_+:\mathbb{R}\to\mathbb{R}^+$ is defined as $[z]_+=\max(z,0)$ and the constant $\gamma\geq 0$ acts as a hyper-parameter controlling the trade-off. This derivation follows along the same lines as the derivation of the standard SVM classifier when implementing a 2-norm for $\|\cdot\|$, see e.g. [23], [19]. Although the term $\sum_{p=1}^{P}\left\|f^{(p)'}\right\|$ is derived as the term maximizing the margin, one can view (5) also as a regularized cost criterion consisting of a fitting term (the Hinge loss) and a penalization term. The following sections proposes different appropriate regularization schemes in the context of additive models (5).

B. Regularization and structure detection

For illustrative reasons, let us focus for a moment on a simple linear model (belonging to the class of additive models):

$$f_a(x)=\sum_{p=1}^{P}a_px^p+b=a^Tx+b. \quad (6)$$

The maximal linear separating hyperplane is found as in (5) as

$$\mathcal{J}_\gamma(a,b)=\|a\|_2^2+\gamma\sum_{i=1}^{N}\left[1-y_i\left(a^Tx_i+b\right)\right]_+. \quad (7)$$

While Tikhonov regularization schemes using the 2-norm $\|a\|_2^2$ are commonly used in order to improve estimates, interest in L_1 based regularization schemes has emerged e.g. in LASSO estimators [22] and basis pursuit [8], [3] algorithms. The generalization to general Minkowski norms has been discussed

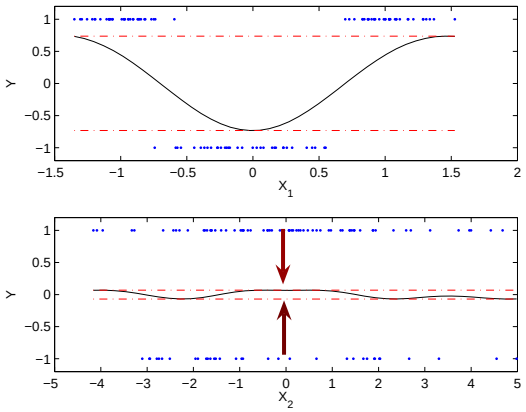

Fig. 1. *Contributions (solid line) of two components of componentwise SVM with corresponding maximal variations (dashed-dotted lines). The component on the top figure has the largest maximal variation. The component displayed on the lower panel has a contribution approaching zero (indicated by the black arrows).*

e.g. under the name of bridge regression [5], [9], [1]. In [16], the use of the following criterion is proposed:

Definition II.2 [Maximal Variation] *The maximal variation of a function $f_p:\mathbb{R}^{D_p}\to\mathbb{R}$ is defined as*

$$\mathcal{M}_p=\max_{x^{(p)}\in\mathbb{R}^{D_p}}\left|f_p\left(x^{(p)}\right)\right|, \quad (8)$$

for all $x^{(p)}\in\mathbb{R}^{D_p}$ belonging to the domain of f_p. The empirical maximal variation can be defined as

$$\hat{\mathcal{M}}_p=\max_i\left|f_p\left(x_i^{(p)}\right)\right|, \quad (9)$$

with $x_i^{(p)}$ denoting the pth component of the ith sample of the training set $\mathcal{D}$.

This definition corresponds with an L_∞ norm on the contributions. $\hat{\mathcal{M}}_p=\left\|\left(f_p\left(x_1^{(p)}\right),\ldots,f_p\left(x_N^{(p)}\right)\right)\right\|_\infty$.

A main advantage of this measure over classical schemes based on the norm of the parameters is that this measure is not directly expressed in terms of the parameter vector (which can be infinite dimensional in the case of kernel machines). Under the assumption of the linear model (6) and of proper normalization of the inputs (say, $L\leq x_p\leq L$), the following relation holds

$$|a_p|_1=\frac{1}{L}|La_p|_1=\frac{\mathcal{M}_p}{L}, \quad (10)$$

Let $L=1$ after proper rescaling of the input observations, then one can rewrite (5) as follows by replacing the maximal variations $\mathcal{M}_p$ by its empirical counterpart, it can be solved efficiently as

$$\min_{a,b,t}\mathcal{J}_\gamma(a,b,t)=\sum_{p=1}^{P}t_p+\gamma\sum_{i=1}^{N}\left[1-y_i(a^Tx_i+b)\right]_+$$
$$\text{s.t.}\quad -t_p\leq a_px_i^{(p)}\leq t_p,\quad\forall i,p, \quad (11)$$

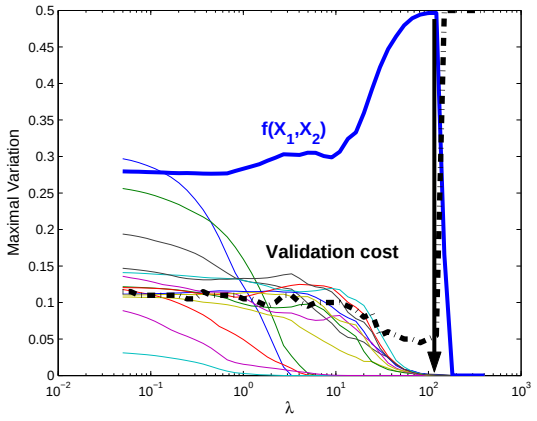

Fig. 2. *Structure diagram of extended Ripley dataset ($D = 5$) including all first to second order components. The diagram gives the evolution of $\mathcal{M}_p$ per component when varying λ. The bold face line indicates the contribution on the original variables X_1 and X_2. The structure is detected successfully at the optimal validation performance (arrow).*

which can be solved as a linear programming problem with $2P + 1$ unknowns and $2NP$ inequalities. Note that the optimization problem (11) has increased complexity with respect to the optimization problem corresponding with the LASSO formulation where the complexity is written in terms of $|a|_1$. However, the extension to the kernel version and the way to cope with the missing values will crucially depend on this measure of maximal variation. Furthermore, as the measure of maximal variation depends only on the predicted outputs and not on the parameterized mapping, the regularization scheme becomes independent of the normalization and dimensionality of the individual components.

Along the same lines, one can also motivate the use of the following combined regularization scheme where both the 2-norm case as the maximal variation scheme can be seen as special cases (by a suitable choice of the hyper-parameters γ and λ):

$$\min_{a,b,t} \mathcal{J}_{\gamma,\lambda} = a^T a + \lambda \sum_p t_p + \gamma \sum_{i=1}^{N} \left[1 - y_i(a^T x_i + b)\right]_+$$
$$\text{s.t.} \quad -t_p \leq a_p x_i^{(p)} \leq t_p, \quad \forall p, i. \qquad (12)$$

Although an extra hyper-parameter is to be determined, this criterion has increased flexibility over the criterion (11). Furthermore, this combined regularization results in sparse components (with zero empirical maximal variation) as well as smoothness (minimal 2-norm of the parameters). These different properties will be especially important in the nonlinear (kernel) setting as will be discussed in section III.

C. Missing data

Often it is highly desirable to include data samples $x \in \mathbb{R}^D$ which are only partially observed. Let for example the qth component of x (denoted as $x^{(q)}$ with $q \in \{1, \ldots, P\}$) be missing. Following from the additive model structure and the measure of maximal variation $\mathcal{M}_q$, one can bound the contribution of the missing variable $x^{(q)}$ in the additive model (3) by

$$-\mathcal{M}_q \leq f_q\left(x^{(q)}\right) \leq \mathcal{M}_q, \quad \forall q = 1, \ldots, P. \qquad (13)$$

Note at this stage that no assumptions are made on the missing values. One proceeds further under the assumption that the empirical maximal variation $\hat{\mathcal{M}}_p$ is not deviating too much from $\mathcal{M}_p$. A worst case approach to the problem of missing values is to incorporate the partially observed values and to substitute the missing parts with the worst case contribution motivating the following definition.

The classification risk functional was defined as follows [23]

$$\mathcal{R}(f, P_{XY}) = \int \left[-f(x)y\right]_+ dP_{XY}, \qquad (14)$$

where $[x]_+$ denotes the positive part $\max(x, 0)$ and P_{XY} denotes the joint distribution underlying the data. Its empirical counterpart is then defined as

$$\hat{\mathcal{R}}(w, \mathcal{D}) = \sum_{i=1}^{N} \left[-w^T \varphi(x_i) y_i\right]_+. \qquad (15)$$

The following counterpart is proposed in the case of missing values.

Definition II.3 [Worst-case Empirical Risk] *Let the function f be extended to $f^{\hat{\mathcal{M}}} : \mathbb{R}^D \to S \subset \mathbb{R}$ defined on the training data as follows*

$$f^{\hat{\mathcal{M}}}(x_i) = \left[\sum_{p \notin \mathcal{S}_i} w_p^T \varphi_p\left(x_i^{(p)}\right) - \sum_{p \in \mathcal{S}_i} \hat{\mathcal{M}}, \right.$$
$$\left. \sum_{p \notin \mathcal{S}_i} w_p^T \varphi_p\left(x_i^{(p)}\right) + \sum_{p \in \mathcal{S}_i} \hat{\mathcal{M}}\right]. \qquad (16)$$

The worst-case empirical counterpart may then be written as

$$\hat{\mathcal{R}}_{\hat{\mathcal{M}}}(w, \mathcal{D}) = \sum_{i=1}^{N} \max_{z \in f^{\hat{\mathcal{M}}}(x_i)} \left[-y_i z\right]_+. \qquad (17)$$

Formally, the distance from the margin can be bounded as follows

$$d(x_p^*, \mathcal{H}_{f_p}) \geq y_i \left(\frac{\sum_{p=1}^{P} f_p(x_*^{(p)}) + b}{\sum_{p=1}^{P} \|f^{(p)\prime}\left(x_*^{(p)}\right)\|}\right)$$
$$\geq \frac{y_i\left(\sum_{p \neq q} f_p(x_*^{(p)}) + b\right) - \mathcal{M}_q}{\sum_{p=1}^{P} \|f^{(p)\prime}\left(x_*^{(p)}\right)\|}.$$

While theoretical analysis of this worst-case scenario in the form of bounds on the deviation of $\mathcal{R}(f, P_{XY})$ and $\hat{\mathcal{R}}_{\hat{\mathcal{M}}}(w, \mathcal{D})$ remains to be done, we proceed here with the construction of a kernel machine implementing this formulation.

Let the set $\mathcal{S}_i$ denote indices of the missing observations of the P components corresponding to the ith data point. Given the empirical maximal variance measure as represented by

$t_p \geq 0$ for all $p = 1, \ldots, P$, the cost-function can be modified as follows

$$\min_{a,b,t} \mathcal{J}_{\gamma,\mathcal{S}}(a,b,t) = \sum_{p=1}^{P} t_p + \gamma \sum_{i=1}^{N} \xi_i \quad \text{s.t.}$$
$$\begin{cases} y_i \left(\sum_{p \notin \mathcal{S}_i} a_p x_i^p \right) - \sum_{q \in \mathcal{S}_i} t_q \geq 1 - \xi_i & \forall i = 1, \ldots, N \\ \xi_i \geq 0 & \forall i = 1, \ldots, N \\ -t_p \leq a_p x_i^p \leq t_p, & \forall i, p \mid x_i^p \in \mathcal{S}_i, \end{cases} \tag{18}$$

where $\xi_i \in \mathbb{R}_0^+$ are slack-variables for all $i = 1, \ldots, N$.

III. Primal-Dual Kernel Machines

A. Componentwise Primal-Dual Kernel Classifiers

Consider the model

$$f(x) = \sum_{p=1}^{P} w_p^T \varphi_p \left(x^{(p)} \right) + b, \tag{19}$$

where $\varphi_p(\cdot) : \mathbb{R}^{D_p} \to \mathbb{R}^{n_h}$ denote the potentially infinite dimensional feature map for all $p = 1, \ldots, P$. The following regularized cost-function is considered:

$$\min_{w,\xi,t} \mathcal{J}_{C,\lambda}(w,\xi,t) = \lambda \sum_{p=1}^{P} t_p + \frac{1}{2} \sum_{p=1}^{P} w_p^T w_p + C \sum_{i=1}^{N} \xi_i,$$
$$\text{s.t.} \quad \begin{cases} y_i \left(\sum_{q \notin \mathcal{S}_i} w_p^T \varphi_p \left(x_i^{(p)} \right) + b \right) \\ \quad - \sum_{q \in \mathcal{S}_i} t_q \geq 1 - \xi_i & \forall i = 1, \ldots, N \\ \xi_i \geq 0 & \forall i = 1, \ldots, N \\ -t_p \leq w_p^T \varphi_p \left(x_i^{(p)} \right) \leq t_p, & \forall i, p \mid p \in \mathcal{S}_i. \end{cases} \tag{20}$$

The dual problem is given in the following Lemma.

Lemma III.1 [Componentwise Primal-Dual Kernel Machine for missing values] *The dual solution to problem (20) becomes*

$$\max_{\alpha_i, \rho_{ip}^+, \rho_{ip}^-} -\frac{1}{2} \sum_{i,j=1}^{N} \alpha_{y,i}^{(p)} \alpha_{y,j}^{(p)} \tilde{\Omega}_{ij}^P + \sum_{i=1}^{N} \alpha_i$$
$$\text{s.t.} \quad \begin{cases} \alpha_{y,i}^{(p)} = \alpha_i y_i + \rho_{ip}^+ - \rho_{ip}^- & \forall i \mid p \notin \mathcal{S}_i \\ \alpha_{y,i}^{(p)} = \alpha_i y_i & \forall i \mid p \in \mathcal{S}_i \\ \sum_{i=1}^{N} y_i \alpha_i = 0 & \\ \lambda = \sum_{i \mid p \notin \mathcal{S}_i} (\rho_{ip}^+ + \rho_{ip}^-) - \sum_{i \mid p \in \mathcal{S}_i} \alpha_i & \forall p \\ 0 \leq \alpha_i \leq C & \forall i = 1, \ldots, N \\ \rho_{ip}^+, \rho_{ip}^- \geq 0, & \forall i, \ \forall p \in \mathcal{S}_i, \end{cases} \tag{21}$$

where $\tilde{\Omega}_{ij}^P = \sum_{p=1}^{P} \tilde{K}_p \left(x_i^{(p)}, x_j^{(p)} \right)$ *for all* $i, j = 1, \ldots, N$ *and where* $\tilde{K}_p \left(x_i^{(p)}, x_j^{(p)} \right) = K_p \left(x_i^{(p)}, x_j^{(p)} \right)$ *if* $x_i^{(p)}$ *nor* $x_j^{(p)}$ *are missing and zero otherwise. The resulting nonlinear classifier evaluated on a new data point* $x_* = \left(x_*^{(1)}, \ldots, x_*^{(P)} \right)$ *takes the form*

$$\text{sign} \left[\sum_{p=1}^{P} \sum_{i=1}^{N} \alpha_i^{(p)} K_p \left(x_i^{(p)}, x_*^{(p)} \right) + b \right], \tag{22}$$

where $\hat{\alpha}_i^{(p)}$ *for all* $i = 1, \ldots, N$ *and* $p = 1, \ldots, P$ *are the unique solutions to (21)*

Proof: The dual solution is given after construction of the Lagrangian

$$\mathcal{L}_{C,\lambda}(w_p, \xi_i, t_p; \alpha_i, \nu_i, \rho_{ip}^+, \rho_{ip}^-) = \mathcal{J}_{C,\lambda}(w_p, \xi_i, t_p) - \sum_{i=1}^{N} \nu_i \xi_i$$
$$- \sum_{i=1}^{N} \alpha_i \left[y_i \left(\sum_{p \notin \mathcal{S}_i} w_p \varphi_p \left(x_i^{(p)} \right) - \sum_{p \notin \mathcal{S}_i} t_p + b \right) - 1 + \xi_i \right]$$
$$- \sum_{i,p} \rho_{ip}^+ \left(t_p + w_p^T \varphi_p \left(x_i^{(p)} \right) \right) - \sum_{i,p} \rho_{ip}^- \left(t_p - w_p^T \varphi_p \left(x_i^{(p)} \right) \right), \tag{23}$$

with positive multipliers $0 \leq \alpha_i, \nu_i, \rho_{ip}^+, \rho_{ip}^-$. The solution is given by the saddle point of the Lagrangian [2]

$$\max_{\alpha_i, \nu_i, \rho_{ip}^+, \rho_{ip}^-} \min_{w_p, b, \xi_i, t_p} \mathcal{L}_{C,\lambda}. \tag{24}$$

By taking the first order conditions $\frac{\partial \mathcal{L}_{\gamma,\lambda}}{\partial w_p} = 0$, $\frac{\partial \mathcal{L}_{\gamma,\lambda}}{\partial b} = 0$, $\frac{\partial \mathcal{L}_{\gamma,\lambda}}{\partial \xi_i} = 0$ and $\frac{\partial \mathcal{L}_{\gamma,\lambda}}{\partial t_p} = 0$, one obtains the (in)equalities $w_p = \sum_{i \mid p \in \mathcal{S}_i} \left(\alpha_i y_i + \rho_{ip}^+ - \rho_{ip}^- \right) \varphi_p(x_i^{(p)})$, $\sum_{i=1}^{N} \alpha_i y_i = 0$, $0 \leq \alpha_i \leq C$ and $\lambda = -\sum_{i \notin \mathcal{S}_i} \alpha_i + \sum_{i \in \mathcal{S}_i} \left(\rho_{ip}^+ + \rho_{ip}^- \right)$. By application of the kernel trick $K_p(x_i, x_j) = \varphi_p(x_i^{(p)})^T \varphi_p(x_j^{(p)})$, the solution to (20) is found by solving the dual problem The primal variables b, ξ_i and t_p can be recovered from the complementary slackness conditions. ∎

An important difference of this result with derivations of componentwise LS-SVMs [14] is seen in that the sets of unknowns $\alpha^{(p)}$ are not equal for all different components.

The main drawback of this approach is the huge number of Lagrange multipliers ($N(2P+1)$) which occur in the dual optimization problem. Note that this number can be reduced by only including those constraints of maximal variation belonging to different input values $\left(x_i^{(p)} \neq x_j^{(p)} \right)$. This is especially useful in case a number of components consisting of categorical or binary values.

B. Componentwise LS-SVMs with additive regularization trade-off

To circumvent the main drawback of the previous derivation, an alternative optimization approach is described. Instead of looking directly for the optimum to the penalized criterion (20), a less complex derivation is employed in order to construct a flexible classifier in the form of an LS-SVM classifier with additive regularization trade-off [15] and its

extension to componentwise models [14]. The more complex loss function (20) is then used as a model selection criterion. This hierarchical modeling strategy was proposed in [16].

The componentwise LS-SVM classifiers with additive regularization takes the form

$$\text{sign}\left[\sum_{p=1}^{P} w_p^T \varphi_p\left(x^{(p)}\right) + b\right]. \tag{25}$$

The cost function exist of a 2-norm regularization term as classical and a squared loss of the slack variables (corresponding to nonlinear regularized Fisher Discriminant Analysis). The trade-off between both terms is made in an additive way using the vector of constants $c = (c_1, \ldots, c_N) \in \mathbb{R}^N$ as follows [20], [15]

$$\begin{aligned} &\min_{w_p, b, \xi_i} \mathcal{J}_c(w_p, \xi_i) = \frac{1}{2}\sum_{p=1}^{P} w_p^T w_p + \frac{1}{2}\sum_{i=1}^{N}(\xi_i - c_i)^2 \\ &\text{s.t.} \quad y_i(\sum_{p=1}^{P} w_p^T \varphi_p\left(x_i^{(p)}\right) + b) = 1 + \xi_i \quad \forall i = 1, \ldots, N. \end{aligned} \tag{26}$$

Lemma III.2 [Componentwise LS-SVM classifier with AReg] *Necessary and sufficient conditions are given as*

$$\begin{cases} \sum_{i=1}^{N} \alpha_i y_i = 0 & \forall i = 1, \ldots, N \\ \xi_i = \alpha_i + c_i & \forall i = 1, \ldots, N \\ y_i\left(\sum_{j=1}^{N} \alpha_j y_j \Omega_{ij}^P + b\right) = 1 - \xi_i & \forall i = 1, \ldots, N \end{cases} \tag{27}$$

The model can be evaluated on a new data point $x_* = \left(x_*^{(1)}, \ldots, x_*^{(P)}\right)$ *as*

$$\text{sign}\left[\sum_{i=1}^{N} \hat{\alpha}_i y_i \sum_{p=1}^{P} K_p\left(x_i^{(p)}, x^{p*}\right) + \hat{b}\right], \tag{28}$$

where $\hat{\alpha}_i$ *and* $\hat{b}$ *are the unique Lagrange multipliers solving (27)*

Proof: The Lagrangian becomes $\mathcal{L}_c(w_p, b, \xi_i; \alpha_i) = \frac{1}{2}\sum_{p=1}^{P} w_p^T w_p + \frac{1}{2}\sum_{i=1}^{N}(\xi_i - c_i)^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(\sum_{p=1}^{P} w_p^T \varphi_p\left(x_i^{(p)}\right) + b) - 1 + \xi_i\right]$ with multipliers $\alpha_i \in \mathbb{R}$. Then the Karush-Kuhn-Tucker (KKT) conditions for optimality $\frac{\partial \mathcal{L}_c}{\partial w_p} = 0$, $\frac{\partial \mathcal{L}_c}{\partial b} = 0$, $\frac{\partial \mathcal{L}_c}{\partial \xi_p} = 0$ and $\frac{\partial \mathcal{L}_c}{\partial \alpha_i} = 0$ for the training can be written as follows (for all $i = 1, \ldots, N$ and $p = 1, \ldots, P$) $w_p = \sum_{i=1}^{N} \alpha_i y_i \varphi_p(x_i^{(p)})$, $\sum_{i=1}^{N} \alpha_i y_i = 0$, $\xi_i = \alpha_i + c_i$ and $y_i(\sum_{p=1}^{P} w_p^T \varphi_p\left(x_i^{(p)}\right) + b) = 1 - \xi_i$. After elimination of the possibly infinite dimensional vectors w_p, one obtains the necessary and sufficient training conditions given in (27). ■

C. Hierarchical Kernel Machines

For model (25) and loss function (26) with hyperparameters $c \in \mathbb{R}^N$, the optimal solution is found by solving the set of linear equations (27). Given this highly flexible kernel machine classifier, one can consider the problem of how to select the model most satisfying a more complex criterion on the level of model selection. Inspired by (24), consider the criterion which should be minimized over the training variables w_p, b, α_i, ξ_i, the hyper-parameters c and the new variable t_p: $\mathcal{J}_{\gamma,\lambda}^H(t_p, w_p, b) = \lambda\sum_{p=1}^{P} t_p + \sum_{p=1}^{P} w_p^T w_p + \gamma\sum_{i=1}^{N}\left[1 - y_i\left(\sum_{p=1}^{P} w_p \varphi_p(x_i^{(p)}) + b\right)\right]_+$ such that the training equations of (27) hold exactly and t_p is the maximal variation of the pth component. Following from this procedure, training level (27) and model selection are separated exactly conceptually by exploiting the KKT conditions, while computationally, both follow from a single convex problem.

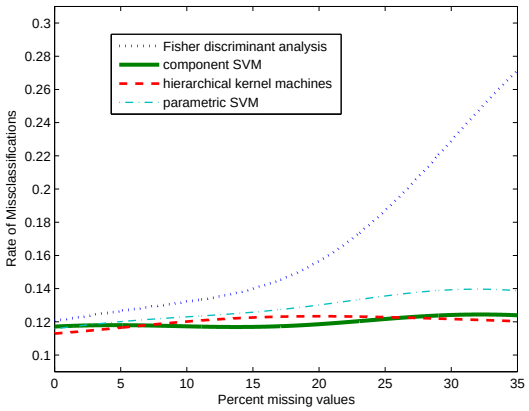

Fig. 3. *Misclassification rate of the extended Ripley dataset in function of the percentage of missing values. Notice that the worst-case analysis is not breaking down when the percentage of missing values is growing.*

After elimination of the variables w_p, ξ_i and c_i, one obtains the problem.

$$\min_{t_p, \alpha_i, b, \epsilon_i} \mathcal{J}_{\gamma,\lambda}^H(\alpha, t_p, \epsilon_i) = \lambda\sum_{p=1}^{P} t_p + \sum_{p=1}^{P} \alpha^T \Omega_y^p \alpha + \gamma\sum_{i=1}^{N} \epsilon_i$$
$$\text{s.t.} \begin{cases} \Omega_{y'}^P \alpha + bY \geq 1 - \epsilon_i \\ \epsilon \geq 0 \\ -1_N t_p \leq \Omega_y^{(p)} \alpha \leq t_p 1_N \\ \sum_{i=1}^{N} \alpha_i y_i = 0, \end{cases} \tag{29}$$

where $\Omega_y^{(p)} \in \mathbb{R}^{N \times N}$, $\Omega_{y,ij}^{(p)} = y_i y_j K_p(x_i^{(p)}, x_j^{(p)})$ and $\Omega_y^P = \sum_{p=1}^{P} \Omega_y^{(p)}$. The resulting tuned model can be evaluated as in the previous subsection. The derivation based on the modified Hinge loss (18) for handling the missing values follows along the same lines.

IV. Illustrative Examples

A data set was designed in order to quantify the improvements and the difference of the proposed (linear and kernel) componentwise SVM classificators over standard techniques in the case of missing data and multiple irrelevant inputs. The Ripley dataset ($n = 150$, $d = 2$, binary labels) was extended with three extra (irrelevant) inputs drawn from a normal distribution ($\mathcal{N}(0, 1)$). Figure 2 shows results from

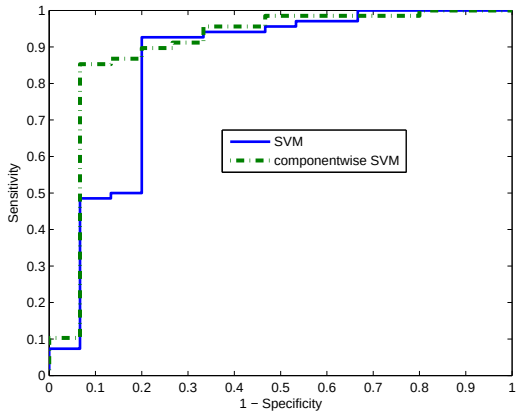

Fig. 4. *ROC curves on the testset of the UCI hepatitis dataset using an SVM with RBF kernel with imputation of missing values and componentwise SVM employing the measure of maximal variation employing the proposed method for handling missing values. The latter consists of 25 nonsparse out of the approximatively 400 components.*

the componentwise SVM with maximal variation (21) on this set where the evolution of the maximal variation $\hat{\mathcal{M}}_p$ per component is displayed against the hyper-parameter λ. The component consisting of inputs X_1 and X_2 is detected correctly by the hyper-parameter optimizing the validation performance. In a second experiment, a portion of the data was marked as missing data. The performance on a disjoint validation set consisting of 100 points was used to tune hyper-parameters, while the final classifier was trained on all 250 samples. The performance on a fresh test set of size 1000 was used to quantify the generalization performance. For the purpose of comparison, the results of linear Fisher discriminant analysis were computed which cope with the missing values by omitting the corresponding samples, while the other approaches follow the derivations of Subsection 2.3. Figure (3) shows the estimated generalization performance in function of the percentage of missing values.

As a second case, one considered the UCI hepatitis dataset ($n = 80, d = 19$) with approximately 50% of the samples containing at least one missing value. A standard SVM with RBF kernel and the componentwise SVM considering up to second order components were compared. The former replaces the missing values with the sample median of the corresponding variable while the latter follows the described worst-case approach. The respective hyper-parameters were tuned using leave-one-out cross-validation. Figure 4 displays the receiver operating characteristic (ROC) curve of both classifiers on a testset of size 55. As the componentwise only employed 25 nonsparse components out of the 380 components up to second order ($D_p \leq 2$), the proposed method outperformed the SVM both in interpretability as generalization performance.

V. Conclusions

This paper extends results from componentwise primal-dual kernel machines towards a setting of classification and structure detection in the presence of missing values. Linear as well as dual kernel versions and a fast implementation are derived.

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

References

[1] A. Antoniadis and J. Fan. Regularized wavelet approximations (with discussion). *Jour. of the Am. Stat. Ass.*, 96:939–967, 2001.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

[4] A.P Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Jour. of the Royal Stat. Soc. series B*, 39:1–38, 1977.

[5] L.E. Frank and J.H. Friedman. A statistical view of some chemometric regression tools. *Technometrics*, (35):109–148, 1993.

[6] J. Friedman. Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19(1):1–141, 1991.

[7] J. Friedman and N. Fisher. Bump hunting in high dimensional data. *Statistics and Computing*, 9:123–143, 1999.

[8] J. H. Friedmann and W. Stuetzle. Projection pursuit regression. *Jour. of the Am. Stat. Assoc.*, 76:817–823, 1981.

[9] W.J. Fu. Penalized regression: the bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, (7):397–416, 1998.

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

[11] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.

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

[13] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 1987.

[14] K. Pelckmans, I. Goethals, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Componentwise least squares support vector machines. *in "Support Vector Machines: Theory and Applications", ed. L. Wang*, 2005. Springer, in press.

[15] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Additive regularization: Fusion of training and validation levels in kernel methods. *Internal Report 03-184, ESAT-SISTA, K.U.Leuven, Belgium, submitted*, 2003.

[16] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Building sparse representations and structure determination on LS-SVM substrates. *Neurocomputing, in Press*, 2004.

[17] D.B. Rubin. Inference and missing data (with discussion). *Biometrika*, 63:581–592, 1976.

[18] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

[19] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

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

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

[22] R.J. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, (58):267–288, 1996.

[23] V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.