Handling Missing Values in Support Vector Machine Classifiers

K. Pelckmans, J.A.K. Suykens, B. De Moor
Katholieke Universiteit Leuven
ESAT - SCD/SISTA
Kasteelpark 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 discusses the task of learning a classifier from observed data containing missing values amongst the inputs which are missing completely at random[1]. A non-parametric perspective is adopted by defining a modified risk taking into account the uncertainty of the predicted outputs when missing values are involved. It is shown that this approach generalizes the approach of mean imputation in the linear case and the resulting kernel machine reduces to the standard Support Vector Machine (SVM) when no input values are missing. Furthermore, the method is extended to the multivariate case of fitting additive models using componentwise kernel machines, and an efficient implementation is based on the Least Squares Support Vector Machine (LS-SVM) classifier formulation.**

I. Introduction

Missing data frequently occur in applied statistical data analysis. There are several reasons why the data may be missing (Rubin, 1976, 1987). They may be missing because equipment malfunctioned, observations become incomplete due to people becoming ill or observations which are not entered correctly. Here the data are missing completely at random (MCAR). The missing data for a random variable X are 'missing completely at random' if the probability of having a missing value for X is unrelated to the values of X itself or to any other variables in the data set. Often the data are not missing completely at random, but they may be classifiable as missing at random (MAR). The missing data for a random variable X are 'missing at random' if the probability of missing data on X is unrelated to the value of X, after controlling for other random variables in the analysis. MCAR is a special type of MAR. If the missing data are MCAR or MAR, the missingness is ignorable and we don't have to model the missingness property. If, on the other hand, data are not missing at random but are missing as a function of some other random variable, a complete treatment of missing data would have to include a model that accounts for missing data.

Three general methods have been mainly used for handling missing values in statistical analysis (Rubin, 1976, 1987). One is the so-called 'complete case analysis', which ignores the observations with missing values and bases the analysis on the complete case data. The disadvantages of this approach are the loss of efficiency due to discarding the incomplete observations and biases in estimates when data are missing in a systematic way. The second approach for handling missing values is the imputation method, which imputes values for the missing covariates and carries out the analysis as if the imputed values were observed data. This approach may reduce the bias of the complete case analysis but lead to additional bias in multivariate analysis if the imputation fails to control for all multivariate relationships. The third approach is to assume some models for the covariates with missing values and then use a maximum likelihood approach to obtain estimates for the models. Methods to handle missing values in non-parametric predictive settings do often rely on different multi-stage procedures or boil down to hard global optimization problems, see e.g. (Hastie et al., 2001) for references.

This paper proposes an alternative approach where no attempt is made to reconstruct the values which are missing, but only the impact of the missingness on the outcome and the expected risk is modeled explicitly. This strategy is in line with the previous result (Pelckmans et al., 2005a) where, however, a worst case approach was taken. The proposed approach is based on a number of insights into the problem, including (i) a global approach for handling missing values which can be reformulated into a one-step optimization problem is preferred; (ii) there is no need to recover the missing values, only the expected outcome of the observations containing missing values is relevant for prediction; (iii) the setting of additive models (Hastie and Tibshirani, 1990) and componentwise kernel machines (Pelckmans et al., 2005b) is preferred as it enables the modeling of the mechanism for handling missing values per variable; (iv) the methodology of primal-dual kernel machines (Vapnik, 1998; Suykens et al., 2002) can be employed to solve the problem efficiently. The cases of standard SVMs (Vapnik, 1998), componentwise SVMs (Pelckmans et al., 2005a) which is related to kernel ANOVA decompositions (Stitson et al., 1999), and componentwise LS-SVMs (Suykens and Vandewalle, 1999; Suykens et al., 2002; Pelckmans et al., 2005b) are elaborated. From a practical perspective, the method can be seen as a weighted version of SVMs and LS-SVMs (Suykens et al., 2002) based on an extended set of dummy variables and is strongly related to the method of sensitivity analysis frequently used for structure detection in

[1]An abbreviated version of some portions of this article appeared in (Pelckmans et al., 2005a) as part of the IJCNN 2005 proceedings, published under the IEEE copyright.

multi-layer perceptrons (see e.g. (Bishop, 1995)).

This paper is organized as follows. The following section discusses the approach taken towards handling missing values in risk based learning. Into section III, this approach is applied in order to build a learning machine for learning a classification rules from a finite set of observations extending the result of SVMs and LS-SVM classifiers. Section IV reports results obtained on a number of artificial as well as benchmark datasets.

II. Minimal risk modeling with missing values

A. Risk with missing values

Let $\ell : \mathbb{R} \to \mathbb{R}$ denote a loss function (as e.g. $\ell(e) = e^2$ or $\ell(e) = |e|$ for all $e \in \mathbb{R}$). Let (X, Y) denote a random vector, $X \in \mathbb{R}^D$ and $Y \in \mathbb{R}$. Let $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ denote the set of training samples with inputs $x_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. The global risk $\mathcal{R}(f)$ of a function $f : \mathbb{R}^D \to \mathbb{R}$ with respect to a fixed (but unknown) distribution P_{XY} is defined as follows (Vapnik, 1998; Bousquet et al., 2004)

$$\mathcal{R}(f) = \int \ell\left(y - f(x)\right) dP_{XY}(x, y). \tag{1}$$

Let $\mathcal{A} \subset \{1, \ldots, N\}$ denote the set with indices of the complete observations and $\overline{\mathcal{A}} = \{1, \ldots, N\} \backslash \mathcal{A}$ the indices with missing values. Let $|\mathcal{A}|$ denote the number of observed values and $|\overline{\mathcal{A}}| = N - |\mathcal{A}|$ the number of missing observations.

Assumption 1 *[Model for Missing Values] The following probabilistic model for the missing values is assumed. Let P_X denote the distribution of X and Then we define*

$$P_X^{(x_i)} \triangleq \begin{cases} \Delta_X^{(x_i)} & \text{if } i \in \mathcal{A} \\ P_X & \text{if } i \in \overline{\mathcal{A}}, \end{cases} \tag{2}$$

where $\Delta_X^{(x_i)}$ denotes the pointmass distribution at the point x_i defined as

$$\Delta_X^{(x_i)}(x) \triangleq \mathcal{I}(x \geq x_i) \quad \forall x \in \mathbb{R}^D, \tag{3}$$

where $\mathcal{I}(x \geq x_i)$ equals one if $x \geq x_i$ and zero elsewhere.

Remark that so far, an input of an observation is either complete or entirely missing. In many practical cases, observations are only partially missing. Section III will deal with the latter by adopting additive models and componentwise kernel machines. The empirical counterpart of the risk $\mathcal{R}(f)$ in (1) then becomes

$$\begin{aligned} \mathcal{R}_{emp}(f) &= \sum_{i=1}^{N} \int \ell\left(y_i - f(x)\right) dP_X^{(x_i)}(x) \\ &= \sum_{i \in \mathcal{A}} \ell\left(y_i - f(x_i)\right) + \sum_{i \in \overline{\mathcal{A}}} \int \ell\left(y_i - f(x)\right) dP_X(x), \end{aligned} \tag{4}$$

after application of the definition in (2) and using the property that integrating over a pointmass distribution equals an evaluation (Pestman, 1998). An unbiased estimate of $\mathcal{R}_{emp}$ can be obtained as follows following the theory of U-statistics (Hoeffding, 1961; Lee, 1990) as follows

$$\mathcal{R}_{emp}^*(f) = \sum_{i \in \mathcal{A}} \ell\left(y_i - f(x_i)\right) + \frac{1}{|\mathcal{A}|} \sum_{i \in \overline{\mathcal{A}}} \sum_{j \in \mathcal{A}} \ell\left(y_i - f(x_j)\right). \tag{5}$$

Note that in case no observations are missing, the risk $\mathcal{R}_{emp}^*$ reduces to the standard empirical risk

$$\mathcal{R}_{emp}(f) = \sum_{i=1}^{N} \ell\left(y_i - f(x_i)\right). \tag{6}$$

B. Mean imputation and minimal risk

Here we prove that the proposed empirical risk bounds the classical method of mean imputation in the case of the squared loss function.

Lemma 1 *Consider the squared loss $\ell = (\cdot)^2$. Define the risk after imputation of the mean $\overline{f} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} f(x_i)$:*

$$\overline{\mathcal{R}}_{emp}(f) = \sum_{i \in \mathcal{A}} \left(f(x_i) - y_i\right)^2 + \sum_{i \in \overline{\mathcal{A}}} \left(\overline{f} - y_i\right)^2. \tag{7}$$

Then the following inequality holds

$$\mathcal{R}_{emp}^*(f) \geq \overline{\mathcal{R}}_{emp}(f). \tag{8}$$

Proof: The first terms of both $\mathcal{R}_{emp}(f)$ and $\overline{\mathcal{R}}_{emp}(f)$ are equal, the second terms are related as follows

$$\begin{aligned} \sum_{j \in \mathcal{A}} \left(f(x_j) - y_i\right)^2 &= \sum_{j \in \mathcal{A}} \left(\left(f(x_j) - \overline{f}\right) - \left(\overline{f} - y_i\right)\right)^2 \\ &= \sum_{j \in \mathcal{A}} \left(\left(f(x_j) - \overline{f}\right)^2 + \left(\overline{f} - y_i\right)^2\right) \\ &\geq |\mathcal{A}| \left(\overline{f} - y_i\right)^2, \end{aligned} \tag{9}$$

from which the inequality follows. ∎

Corollary 1 *Consider the model class*

$$\mathcal{F} = \left\{ f : \mathbb{R}^D \to \mathbb{R} \;\middle|\; f(x, w) = w^T x,\ w \in \mathbb{R}^D \right\}, \tag{10}$$

such that the observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ satisfy $y_i = w^T x_i + e_i$. Then $\mathcal{R}_{emp}^*(w)$ is an upperbound to the standard risk $\mathcal{R}_{emp}(w)$ as in (6) using mean imputation $\overline{x} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} x_i$. of the missing values $\in \overline{\mathcal{A}}$.

Proof: The proof follows readily from Lemma 1 and the equality

$$\overline{y} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} w^T x_i = w^T \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} x_i = w^T \overline{x},$$

where $\overline{x}$ is defined as the empirical mean of the input. ∎

Both results establish a result with the technique of mean imputation (Rubin, 1987). In the case of nonlinear models, however, imputation should rather be based on the average response $\overline{f}$ instead of the input $\overline{x}$.

C. Risk for additive models with missing variables

Additive models are defined as follows (Hastie and Tibshirani, 1990):

Definition 1 *[Additive Models] Let an input vector $x \in \mathbb{R}^D$ consists of Q components of dimension D_q for $q = 1, \ldots, Q$, denoted as $x_i = \left(x_i^{(1)}, \ldots, x_i^{(Q)}\right)$ with $x_i^{(q)} \in \mathbb{R}^{D_q}$. (in the simplest case $n_q = 1$, we denote $x_i^{(q)} = x_i^q$). The class of additive models using these components is defined as*

$$\mathcal{F}^Q = \left\{ f : \mathbb{R}^D \to \mathbb{R} \;\middle|\; f(x) = \sum_{q=1}^{Q} f_q\left(x^{(q)}\right) + b, \right.$$
$$\left. f_q : \mathbb{R}^{D_q} \to \mathbb{R}, b \in \mathbb{R}, \forall x = \left(x^{(1)}, \ldots, x^{(Q)}\right) \in \mathbb{R}^D \right\}. \quad (11)$$

Let furthermore X^q denote the random variable (vector) corresponding to the q-th component for all $q = 1, \ldots, Q$.

Let the sets $\mathcal{A}_q$ and $\mathcal{B}_i$ be defined as follows

$$\mathcal{A}_q = \left\{ i \in \{1, \ldots, N\} \;\middle|\; x_i^{(q)} \quad \text{observed} \quad \right\}, \qquad \forall q = 1, .., Q$$
$$\mathcal{B}_i = \left\{ q \in \{1, \ldots, Q\} \;\middle|\; x_i^{(q)} \quad \text{observed} \quad \right\}, \qquad \forall i = 1, .., N, \quad (12)$$

and let $\overline{\mathcal{A}}_q = \{1, \ldots, N\} \backslash \mathcal{A}_q$ and $\overline{\mathcal{B}}_i = \{1, \ldots, Q\} \backslash \mathcal{B}_i$. In the case of this class of models, one may refine the probabilistic model for missing values to a mechanism which handles the missingness per component.

Assumption 2 *[Model for Missing Values with Additive Models] The probabilistic model for the missing values of the q-th component is given as follows*

$$P_{X^q}^{(x_i)} \triangleq \begin{cases} \Delta_{X^q}^{(x_i)} & \text{if} \;\; i \in \mathcal{A}_q \\ P_{X^q} & \text{if} \;\; i \in \overline{\mathcal{A}}_q, \end{cases} \quad (13)$$

where $\Delta_{X^q}^{(x_i)}$ denotes the pointmass distribution at the point $x_i^{(q)}$ defined as

$$\Delta_{X^q}^{(x_i)}(x) \triangleq \mathcal{I}\left(x^{(q)} \geq x_i^{(q)}\right) \quad \forall x^{(q)} \in \mathbb{R}^{D_q}, \quad (14)$$

where $\mathcal{I}(z \geq z_i)$ equals one if $z \geq z_i$ and zero elsewhere. Under the assumption the variables $X^1, \ldots, X^Q$ are independent, the probabilistic model for the complete observation becomes

$$P_X^{(x_i)} = \prod_{q=1}^{Q} P_{X^q}^{(x_i)} \quad \forall x_i \in \mathcal{D}. \quad (15)$$

Given the empirical risk function $\mathcal{R}_{emp}(f)$ as defined in (4), the risk or the additive model then becomes

$$\mathcal{R}_{emp}(f) = \sum_{i=1}^{N} \int \ell\left(y_i - f(x)\right) dP_X(x) =$$
$$\sum_{i=1}^{N} \int \ell\left(\sum_{q=1}^{Q} f_q\left(x^{(q)}\right) + b - y_i\right) dP_{X^1}(x^{(1)}) \ldots dP_{X^Q}(x^{(Q)}).$$

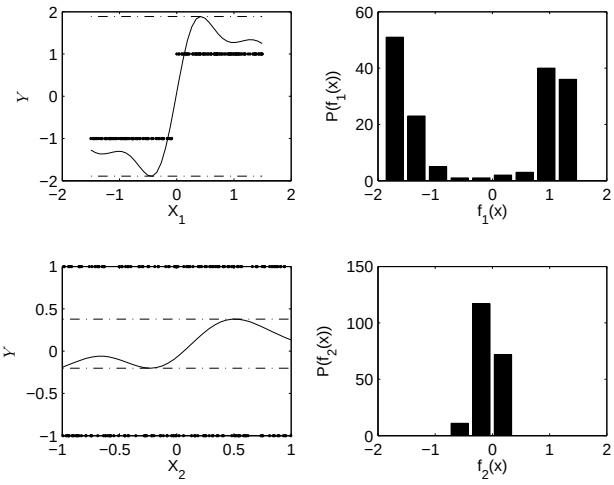


Fig. 1. *Illustration of the mechanism in the case of componentwise SVMs with empirical risk $\mathcal{R}_{emp}^*$ as described in Subsection II.C. Consider the bivariate function $y = f_1(x^1) + f_2(x^2)$ with samples given as the dots at locations $\{-1, 1\}$. The left panels show the contribution associated with the two variables X^1 and X^2 (solid line) and the samples with respect to the corresponding input variables. By inspection of the range of both functions, one may conclude that the first component is more relevant to the problem at hand. The two right panels give the empirical density of the values $f_1(X^1)$ and $f_2(X^2)$ respectively. This empirical estimate is then used to marginalize the influence of the missing variables from the risk.*

In order to cope with the notational incovenience due to the different dependent summands, the following index sets $\mathcal{U} \subset \mathbb{N}^Q$ and $\mathcal{V} \subset \mathbb{N}^Q$ are defined.

$$\mathcal{U}_i = \left\{ (j_1, \ldots, j_Q) \;\middle|\; \right.$$
$$\left. j_q = i \;\text{ if }\; q \in \mathcal{B}_i \;\text{ or }\; j_q = l, \forall l \in \mathcal{A}_q \;\text{ if }\; q \in \overline{\mathcal{B}}_i \right\}, \quad (16)$$

which reduces to the singleton $\{(i, \ldots, i)\}$ if the i-th sample is fully observed. Let $n_{\mathcal{U}}$ equals $\sum_{i=1}^{N} |\mathcal{U}_i|$. Consider e.g. the following dataset $\mathcal{D} = \left\{ \left(x_1^{(1)}, x_1^{(2)}\right), \left(x_2^{(1)}, x_2^{(2)}\right), \left(x_3^{(1)}, \boxed{?}\right) \right\}$ where the second variable of the third observation is missing. Then the sets $\mathcal{U}_i$ become $\mathcal{U}_1 = \{(1,1)\}$, $\mathcal{U}_2 = \{(2,2)\}$, $\mathcal{U}_1 = \{(3,1),(3,2)\}$ and $n_{\mathcal{U}} = 4$.

The empirical risk becomes in general

$$\mathcal{R}_{emp}^{Q,*}(f) =$$
$$\sum_{i=1}^{N} \frac{1}{|\mathcal{U}_i|} \sum_{(j_1, \ldots, j_Q) \in \mathcal{U}_i} \ell\left(\sum_{q=1}^{Q} f_q\left(x_{j_q}^{(q)}\right) + b - y_i\right), \quad (17)$$

where $x_{j_q}^{(q)}$ denotes the q-th component of the j_q-th observation. This expression will be employed to build a componentwise primal-dual kernel machine handling missing values in the next section.

D. Worst case approach using maximal variation

For completeness, the derivation of the worst case approach towards handling missing values is summarized based on (Pelckmans et al., 2005a). Consider again the additive models as defined in Definition 1. In (Pelckmans et al., 2005c), the use of the following criterion was proposed:

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

$$\mathcal{M}_q = \sup_{x^{(q)} \sim P_{X^q}} \left| f_q\left(x^{(q)}\right)\right| \tag{18}$$

for all $x^{(q)} \in \mathbb{R}^{D_q}$ sampled from the distribution P_{X^q} corresponding to the q-th component. The empirical maximal variation can be defined as

$$\hat{\mathcal{M}}_q = \max_{x^{(q)} \in \mathcal{D}_N} \left| f_q\left(x_i^{(q)}\right)\right|, \tag{19}$$

with $x^{(q)}$ denoting the q-th component of a sample of the training set $\mathcal{D}$.

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) and it was employed successfully in (Pelckmans et al., 2005c) in order to build a non-parametric counterpart to the linear LASSO estimator (Tibshirani, 1996) for structure detection. The following counterpart was proposed in the case of missing values.

Definition 3 *[Worst-case Empirical Risk] Let an interval $m_i^f \subset \mathbb{R}$ be associated to each data-sample defined as follows*

$$\begin{cases} x_i \to m_i^f = \sum_{q=1}^{Q} f_q\left(x_i^{(q)}\right) & \text{if } \ i \in \mathcal{A} \\ x_i \to m_i^f = \left[-\sum_{q=1}^{Q} \mathcal{M}_q, \ \ \sum_{q=1}^{Q} \mathcal{M}_q\right] & \text{if } \ i \in \overline{\mathcal{A}} \\ x_i \to m_i^f = \Big[\sum_{q\in\mathcal{B}_i} f_q\left(x_i^{(q)}\right) - \sum_{p\in\overline{\mathcal{B}}_i} \mathcal{M}_p, \\ \qquad\qquad \sum_{q\in\mathcal{B}_i} f_q\left(x_i^{(q)}\right) + \sum_{p\in\overline{\mathcal{B}}_i} \mathcal{M}_p\Big], & \text{otherwise,} \end{cases} \tag{20}$$

such that complete observations are mapped onto a singleton $f(x)$ and an interval of possible outcomes is associated when missing entries are encountered. The worst-case empirical counterpart to the empirical risk $\mathcal{R}_{emp}$ as defined in (4) becomes

$$\mathcal{R}_{emp}^{\hat{\mathcal{M}}}(f) = \sum_{i=1}^{N} \max_{z \in m_i^f} \ell\left(y_i - z\right). \tag{21}$$

A modification to the componentwise SVM based on this worst case risk is studied in (Pelckmans et al., 2005a) and will be used in the experiments for comparison.

III. Primal Dual Kernel Machines

A. SVM classifiers handling missing values

Let us consider the case of general models at first. Consider the classifiers of the form

$$f_w(x) = \text{sign}\left[w^T \varphi\left(x\right) + b\right], \tag{22}$$

where $w \in \mathbb{R}^{D_\varphi}$ and D_φ is the dimension of feature space which is possibly infinite. Let $\varphi : \mathbb{R}^D \to \mathbb{R}^{D_\varphi}$ be a fixed but unknown mapping of the input data to a feature space. Consider the maximal margin classifier where the risk to violating the margin is to be minimized with risk function

$$\begin{aligned} &\mathcal{R}_{emp}^*(f_w) \\ &= \sum_{i\in\mathcal{A}} \left[1 - y_i(f_w(i))\right]_+ + \frac{1}{|\mathcal{A}|} \sum_{i\in\overline{\mathcal{A}}} \sum_{j\in\mathcal{A}} \left[1 - y_i\left(f_w(x_j)\right)\right]_+, \end{aligned} \tag{23}$$

where the function $[\cdot]_+ : \mathbb{R} \to \mathbb{R}^+$ is defined as $[z]_+ = \max(z, 0)$ for all $z \in \mathbb{R}$. The maximization of the margin while minimizing the risk $\mathcal{R}_{emp}^*(f_w)$ using elements of the model class (22) results in the following primal optimization problem which is to be solved with respect to ξ, w_p and b:

$$\min_{w,b,\xi} \mathcal{J}_{\mathcal{A}}(w,\xi) = \frac{1}{2} w^T w + C\left(\sum_{i\in\mathcal{A}} \xi_i + \frac{1}{|\mathcal{A}|} \sum_{i\in\overline{\mathcal{A}}} \sum_{j\in\mathcal{A}} \xi_{ij}\right)$$

s.t.

$$\begin{cases} 1 - \xi_i \geq y_i\left(w^T\varphi(x_i) + b\right) & \forall i \in \mathcal{A} \\ 1 - \xi_{ij} \geq y_i\left(w^T\varphi(x_j) + b\right) & \forall i \in \overline{\mathcal{A}}, j \in \mathcal{A} \\ \xi_i, \xi_{ij} \geq 0 \qquad \forall i = 1, \ldots, N, \ \forall j \in \mathcal{A}. \end{cases} \tag{24}$$

This problem can be rewritten in a substantially lower number of unknowns when at least one missing value occurs. Note that many of the individual constraints of (24) are equal whenever y_i and x_i are the same in $y_i\left(w^T\varphi(x_j) + b\right)$.

$$\begin{cases} 1 - \xi_i \geq y_i\left(w^T\varphi(x_i) + b\right) \\ 1 - \xi_{ki} \geq y_k\left(w^T\varphi(x_i) + b\right) \\ y_i = y_k = 1 \end{cases} \to \xi_i^+ \triangleq \xi_i = \xi_{ki}, \tag{25}$$

and similar for ξ_i^- which equals ξ_i and ξ_{ki} whenever $y_i = y_k = -1$ for all $i \in \mathcal{A}$. Let $\overline{\mathcal{A}}_+$ denote the indices of the samples which contain missing variables and have outputs equal to 1 and $\overline{\mathcal{A}}_-$ the set with outputs $y = -1$. Let $|\overline{\mathcal{A}}|$ denote the cardinality of the set $\overline{\mathcal{A}}$. One rewrites then

$$\min_{w,b,\xi} \mathcal{J}_{\mathcal{A}}^*(w,\xi^+,\xi^-) = \frac{1}{2} w^T w + C \sum_{i\in\mathcal{A}} \left(n_i^+ \xi_i^+ + n_i^- \xi_i^-\right)$$

$$\text{s.t.} \quad \begin{cases} 1 - \xi_i^- \geq -\left(w^T\varphi(x_i) + b\right) & \forall i \in \mathcal{A} \\ 1 - \xi_i^+ \geq \ \left(w^T\varphi(x_j) + b\right) & \forall i \in \mathcal{A} \\ \xi_i^-, \xi_i^+ \geq 0 & \forall i \in \mathcal{A}, \end{cases} \tag{26}$$

where $n_i^+ = \mathcal{I}(y_i > 0) + \frac{|\overline{\mathcal{A}}_+|}{|\mathcal{A}|}$ and $n_i^- = \mathcal{I}(y_i < 0) + \frac{|\overline{\mathcal{A}}_-|}{|\mathcal{A}|}$ are positive numbers.

Lemma 2 *[Primal-Dual Characterization, I] Let π be a transformation of the indices such that π maps the set of indices $\{1, \ldots, |\mathcal{A}|\}$ onto an enumeration of all samples with completely observed inputs. The dual problem to (26) takes*

the following form

$$\max_{\alpha} \mathcal{J}_C^D(\alpha) =$$
$$-\frac{1}{2}\left({\alpha^+}^T\Omega\alpha^+ - 2{\alpha^+}^T\Omega\alpha^- + {\alpha^-}^T\Omega\alpha^-\right)+1_{|\mathcal{A}|}^T\alpha^+ + 1_{|\mathcal{A}|}^T\alpha^-$$
$$\text{s.t.} \quad \begin{cases} 1_{|\mathcal{A}|}\alpha - 1_{|\mathcal{A}|}\alpha^- = 0 & \\ 0 \leq \alpha_i^+ \leq n_i^+ C & \forall i \in \mathcal{A} \\ 0 \leq \alpha_i^- \leq n_i^- C & \forall i \in \mathcal{A}, \end{cases} \quad (27)$$

where $\Omega \in \mathbb{R}^{2|\mathcal{A}|\times 2|\mathcal{A}|}$ is defined as $\Omega_{kl} = K(x_{\pi(k)}, x_{\pi(l)})$ for all $k, l = 1, \ldots, |\mathcal{A}|$. The estimate can be evaluated in a new data-point $x_* \in \mathbb{R}^D$ as follows

$$\hat{y}_* = \text{sign}\left[\sum_{i\in|\mathcal{A}|}(\hat{\alpha}_i^+ - \hat{\alpha}_i^-)K(x_{\pi(i)}, x_*) + \hat{b}\right], \quad (28)$$

where $\hat{\alpha}$ is the solution to (27) and $\hat{b}$ follows from the complementary slackness conditions.

Proof: Let the positive vectors $\alpha^+ \in \mathbb{R}^{+,|\mathcal{A}|}$, $\alpha^- \in \mathbb{R}^{+,|\mathcal{A}|}$, $\nu^+ \in \mathbb{R}^{+,|\mathcal{A}|}$ and $\nu^- \in \mathbb{R}^{+,|\mathcal{A}|}$ contain the Lagrange multipliers of the constrained optimization problem (26). The Lagrangian of the constrained optimization problem becomes

$$\begin{aligned}\mathcal{L}_C(w, b, \xi; \alpha^+, \alpha^-, \nu^+, \nu^-) &= \mathcal{J}_C^+(w, \xi^+, \xi^-) \\ &- \sum_{i\in\mathcal{A}}\nu_i^+(\xi_i^+) - \sum_{i\in\mathcal{A}}\alpha_i^+\left(\left(w^T\varphi(x_i)+b\right) - 1 + \xi_i^+\right) \\ &- \sum_{i\in\mathcal{A}}\alpha_i^-\left(-\left(w^T\varphi(x_i)+b\right) - 1 + \xi_i^-\right) - \sum_{i\in\mathcal{A}}\nu_i^-(\xi_i^-),\end{aligned} \quad (29)$$

such that $\alpha_i^+, \nu_i^+, \alpha_i^-, \alpha_i^+ \geq 0$ for all $i = 1, \ldots, |\mathcal{A}|$. Then from taking the first order conditions for optimality over the primal variables (saddle point of the Lagrangian), one obtains

$$\begin{cases} w = \sum_{i\in\mathcal{A}}\left(\alpha_i^+ - \alpha_i^-\right)\varphi(x_i) & & (a) \\ 0 = \sum_{i\in\mathcal{A}}\left(\alpha_i^+ - \alpha_i^-\right) & & (b) \\ Cn_i^+ = \alpha_i^+ + \nu_i^+ & \forall i \in \mathcal{A} & (c) \\ Cn_i^- = \alpha_i^- + \nu_i^- & \forall i \in \mathcal{A} & (d). \end{cases} \quad (30)$$

The dual problem then follows by maximization over α^+, α^-, see e.g. (Boyd and Vandenberghe, 2004; Cristianini and Shawe-Taylor, 2000; Suykens et al., 2002). ■

From the expression (27), the following result follows:

Corollary 2 *The Support Vector Machine for handling missing values reduces to the standard support vector machine in case no values are missing.*

Proof: From the definition of n_i^+ and n_i^- it follows that only one of them can be equal to one in the case of no missing values, while the other equals zero. From the conditions (30.cd), equivalence with the standard SVM follows, see e.g. (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000; Suykens et al., 2002). ■

B. Componentwise SVMs handling missing values

The paradigm of additive models is employed to handle multivariate data where only some of the variables are missing at a time. Additive classifiers are then defined as follows. Let $x \in \mathbb{R}^D$ be a point with components $x = \left(x^{(1)}, \ldots, x^{(Q)}\right)$. Consider the classification rule in componentwise form (Hastie and Tibshirani, 1990)

$$\text{sign}[f(x)] = \text{sign}\left[\sum_{q=1}^{Q} f_q\left(x^{(q)}\right) + b\right], \quad (31)$$

with sufficiently smooth mappings $f_q : \mathbb{R}^{D_q} \to \mathbb{R}$ such that the decision boundary is described as in (Vapnik, 1998; Schölkopf and Smola, 2002)

$$\mathcal{H}_f = \left\{x_0 \in \mathbb{R}^D \;\middle|\; \sum_{q=1}^{Q} f_q\left(x_0^{(q)}\right) + b = 0\right\}. \quad (32)$$

The primal-dual characterization provides an efficient implementation of the estimation procedure for fitting such models to the observations. Consider additive classifiers of the form

$$\text{sign}[f_w(x)] = \text{sign}\left[\sum_{q=1}^{Q} w_q^T\varphi_q\left(x^{(q)}\right) + b\right], \quad (33)$$

with φ_q for all $q = 1, \ldots, Q$ fixed but unknown mappings from the q-th component $x^{(q)}$ to an element in a corresponding feature space $\varphi_q\left(x^{(q)}\right)$ belonging to a space $\mathbb{R}^{D_{\varphi_q}}$ which is possibly infinite. The derivation of the algorithm for additive models incorporating the missing values goes along the same lines as in Lemma 2 but involves a heavier notation. Let $\xi_{i,u_i} \in \mathbb{R}^+$ denote slack variables for all $i = 1, \ldots, N$ and $\forall u_i \in \mathcal{U}_i$. Then the primal optimization problem can be written as follows

$$\mathcal{J}_{\mathcal{A}}^Q(w_q, \xi) = \frac{1}{2}\sum_{q=1}^{Q} w_q^T w_q + C\sum_{i=1}^{N}\frac{1}{|\mathcal{U}_i|}\sum_{u_i\in\mathcal{U}_i}\xi_{i,u_i} \quad \text{s.t.}$$
$$\begin{cases} 1 - \xi_{i,u_i} \geq y_i\left(\sum_{q=1}^{Q} w_q^T\varphi_q\left(x_{j_q}^{(q)}\right) + b\right) \\ \qquad\qquad \forall i = 1, \ldots, N, \quad \forall u_i = (j_1, \ldots, j_Q) \in \mathcal{U}_i \\ \xi_{i,u_i} \geq 0 \qquad\qquad \forall i = 1, \ldots, N, \quad \forall u_i \in \mathcal{U}_i \end{cases} \quad (34)$$

which ought to be minimized over the primal variables w_q, b and $\xi_{i,}$ for all $q = 1, \ldots, Q$, $i = 1, \ldots, N$ and $u_i \in \mathcal{U}_i$ respectively. Let $u_{i,q}$ denote the q-th element of the vector u_i.

Lemma 3 *[Primal-Dual Characterization, II] The dual problem to (34) becomes*

$$\max_{\alpha} J_{\mathcal{A}}^{Q,D}(\alpha) = -\frac{1}{2}\alpha^T\Omega_{\mathcal{U}}^Q\alpha + 1_{n_{\mathcal{U}}}^T\alpha$$
$$\text{s.t.} \quad \begin{cases} 0 \leq \alpha_{i,u_i} \leq \sum_{u_i\in\mathcal{U}_i}\dfrac{C}{|\mathcal{U}_i|} \quad \forall i = 1, \ldots, N, \forall u_i \in \mathcal{U}_i \\ \sum_{i=1}^{N}\sum_{u_i\in\mathcal{U}_i}\alpha_{i,u_i} = 0. \end{cases} \quad (35)$$

Let the matrix $\Omega_{\mathcal{U}}^Q \in \mathbb{R}^{n_{\mathcal{U}} \times n_{\mathcal{U}}}$ *be defined such that* $\Omega_{\mathcal{U},u_i,u_j}^Q = \sum_{q=1}^Q y_i y_j K_q \left(x_{u_{i,q}}^{(q)}, x_{u_{j,q}}^{(q)}\right)$ *for all* $i,j = 1,\ldots,N$, $u_i \in \mathcal{U}_i$. *The estimate can be evaluated in a new point* $x_* = \left(x_*^{(1)},\ldots,x_*^{(Q)}\right)$ *as follows*

$$\sum_{i=1}^N y_i \sum_{u_i \in \mathcal{U}_i} \hat{\alpha}_{i,u_i} \sum_{q=1}^Q K_q \left(x_*^{(q)}, x_{u_{i,q}}^{(q)}\right) + \hat{b}, \tag{36}$$

where $\hat{\alpha}$ *and* $\hat{b}$ *are the solution to (35).*

Proof: The Lagrangian of the primal problem (34) becomes

$$\mathcal{L}(w_Q,\xi,b;\alpha,\nu) = \mathcal{J}_C^Q(w,\xi) - \sum_{i=1}^N \sum_{u_i\in\mathcal{U}_i} \nu_{i,u_i}\xi_{i,u_i}$$
$$-\sum_{i=1}^N \sum_{u_i\in\mathcal{U}_i} \alpha_{i,u_i}\left(y_i\left(\sum_{q=1}^Q w_q^T\varphi_q\left(x_{u_{i,q}}^{(q)}\right)+b\right)-1+\xi_{i,u_i}\right), \tag{37}$$

where α is a vector containing the positive Lagrange multipliers $\alpha_{i,u_i} \geq 0$ and where ν is a vector containing the positive Lagrange multipliers $\nu_{i,u_i} \geq 0$. The first order conditions for minimization with respect to the primal variables become

$$\begin{cases} w_q = \sum_{i=1}^N \sum_{u_i\in\mathcal{U}_i} \alpha_{i,u_i} y_i \varphi_q\left(x_{u_{i,q}}^{(q)}\right) & \forall q = 1,\ldots,Q \\ 0 \leq \alpha_{i,u_i} \leq \frac{C}{|\mathcal{U}_i|} & \forall i = 1,\ldots,N, \forall u_i \in \mathcal{U}_i \\ \sum_{i=1}^N \sum_{u_i\in\mathcal{U}_i} \alpha_{i,u_i} y_i = 0. \end{cases} \tag{38}$$

Substitution of this equalities into the Lagrangian and maximizing the expression over the dual varables leads to the dual problem (35). ∎

Again this derivation reduces to a componentwise SVM in the case no missing values are encountered.

C. Componentwise LS-SVMs for classification

A formulation based on the derivation of LS-SVM classifiers is considered resulting into a dual problem which one can solve much more efficiently by adoption of a least squares criterion and by substitution of the inequalities by equalities (Saunders et al., 1998; Suykens and Vandewalle, 1999; Suykens et al., 2002; Pelckmans et al., 2005b). The combinatorial increase in the number of terms can be avoided using the following formulation. The modified primal cost-function of the LS-SVM becomes

$$\min_{w_q,b,z_i} \mathcal{J}_\gamma^Q(w_q,z_i) = \frac{1}{2}\sum_{q=1}^Q w_q^T w_q +$$
$$\frac{\gamma}{2}\sum_{i=1}^N \frac{1}{|\mathcal{U}_i|}\sum_{u_i\in\mathcal{U}_i}\left(y_i\left(\sum_{q=1}^Q z_{u_{i,q}}^q + b\right)-1\right)^2$$
$$\text{s.t.} \quad w_q^T\varphi\left(x_i^{(q)}\right) = z_i^q \quad \forall q = 1,\ldots,Q, \forall i \in \mathcal{A}_q, \tag{39}$$

where $z_i^q = f^q\left(x_i^{(q)}\right) \in \mathbb{R}$ denotes the contribution of the q-th component of the i-th data point. This problem has a dual characterization with complexity independent of the number of terms in the primal cost-function. For notational convenience, define the following sets $\mathcal{V}_{iq} \in \mathbb{N}^Q$ and $\mathcal{V}_q \in \mathbb{N}^Q$. Let $\mathcal{V}_{iq}$ denote a set of vectors of Q indices for all $q = 1,\ldots,Q$ as follows

$$\mathcal{V}_{iq} = \left\{ v_k = (j_1,\ldots,j_Q) \;\middle|\; v_k \in \mathcal{U}_k, \forall k = 1,..,N \;\text{ s.t. }\; j_q = i \right\}. \tag{40}$$

Let $n_{iq} \in \mathbb{R}$ be defined as $n_{iq} = \sum_{v_k\in\mathcal{V}_{iq}} \frac{1}{|\mathcal{U}_k|}$ for all $i = 1,\ldots,N, q = 1,\ldots,Q$ and $d_{iq}^y = \sum_{v_k\in\mathcal{V}_{iq}} \frac{1}{|\mathcal{U}_k|} y_k$ for all $i = 1,\ldots,N, q = 1,\ldots,Q$. and let $n \in \mathbb{R}^{n_{\mathcal{U}}}$ and $d^y \in \mathbb{R}^{n_{\mathcal{U}}}$ be vectors enumerating the elements n_{iq} and d_{iq} respectively.

Lemma 4 *[Primal-Dual Characterization, III] Let* $n_\alpha = \sum_{q=1}^Q |\mathcal{A}_q|$ *denote the number of non-missing values. The dual solution to (39) is found as the solution to the set of linear equations*

$$\left[\begin{array}{c|c} 0 & d^T \\ \hline d & \Omega_{\mathcal{V}}^Q + I_{n_\alpha}/\gamma \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline d^y \end{array}\right], \tag{41}$$

where $\Omega_{\mathcal{V}}^Q \in \mathbb{R}^{n_\alpha \times n_\alpha}$, *the vector* $\alpha = \left(\alpha^1,\ldots,\alpha^Q\right)^T \in \mathbb{R}^{n_\alpha}$. *The estimate can be evaluated at a new point* $x_* = \left(x_*^{(1)},\ldots,x_*^{(Q)}\right)$ *as follows*

$$\hat{f}(x_*) = \sum_{q=1}^Q \sum_{i\in\mathcal{A}_q} \hat{\alpha}_i^q K\left(x_i^{(q)}, x_*^{(q)}\right) + \hat{b}, \tag{42}$$

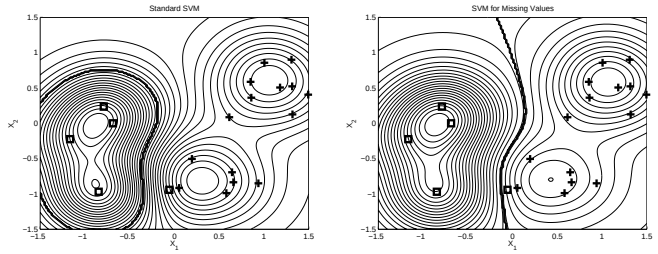
where $\hat{\alpha}_i^q$ *and* $\hat{b}$ *are the solution to (41).*

Proof: The Lagrangian of the primal problem (39) becomes

$$\mathcal{L}_\gamma\left(w_q, z_i^q, b; \alpha\right) = \mathcal{J}_\gamma(w_q, z_i^q, b)$$
$$-\sum_{q=1}^Q \sum_{i\in\mathcal{A}_q} \alpha_i^q\left(w_q^T\varphi_q\left(x_i^{(q)}\right) - z_i^q\right), \tag{43}$$

where $\alpha \in \mathbb{R}^{n_\alpha}$ is a vector with all Lagrange multipliers α_i^q for all $q = 1,\ldots,Q$ and $i \in \mathcal{A}_q$. The minimization of the Lagrangian with respect to the primal variables w_q, b and z_i^q is characterized by

$$\begin{cases} w_q = \sum_{i\in\mathcal{A}_q} \alpha_i^q \varphi_q\left(x_i^{(q)}\right) & \forall q \\ \sum_{v_k\in\mathcal{V}_{iq}} \frac{1}{|\mathcal{U}_k|}\left(\sum_{p=1}^Q z_{v_{k,p}}^p + b - y_k\right) = -\frac{1}{\gamma}\alpha_i^q & \forall q, \forall i \in \mathcal{A}_q \\ \sum_{i=1}^N \frac{1}{|\mathcal{U}_i|}\sum_{u_i\in\mathcal{U}_i}\left(\sum_{q=1}^Q z_{u_{i,q}}^q + b - y_i\right) = 0 & \\ z_i^q = w_q^T\varphi_q\left(x_i^{(q)}\right), & \forall q,\ \forall i \in \mathcal{A}_q. \end{cases} \tag{44}$$



(a) (b)

Fig. 2. *An artificial example ("X" denote positive labels, "□" are negative labels) showing the difference between* **(a)** *the standard SVM using only the complete samples, and* **(b)** *the modified SVM using the all samples using the modified risk $\mathcal{R}^*_{emp}$ as described in Section II.A. While the former results in an unbalanced solution, the latter approximates better the underlying rule $f(X) = \mathcal{I}(X_1 > 0)$ with an improved generalization performance.*

One can eliminate the primal variables w_q and z_i^q from this set using the first and the last expression, resulting in the set

$$\begin{cases} \sum_{p=1}^{Q} \sum_{j \in \mathcal{A}_p} \left[\sum_{v_k \in \mathcal{V}_{iq}} \frac{1}{|\mathcal{U}_k|} K_p \left(x_{v_{k,p}}^{(p)}, x_j^{(p)} \right) \right] \alpha_j^p \\ \qquad + n_{iq} b + \frac{1}{\gamma} \alpha_i^q = d_{iq}^y \qquad \forall q,\ \forall i \in \mathcal{A}_q \\ \sum_{q=1}^{Q} \sum_{j \in \mathcal{A}_q} \alpha_i^q = 0. \end{cases} \tag{45}$$

Define the matrix $\Omega_{\mathcal{U}}^Q \in \mathbb{R}^{n_\alpha \times n_\alpha}$ such that

$$\Omega_{\mathcal{U}}^Q = \begin{bmatrix} \Omega_{s^1}^{(1)} & \dots & \Omega_{s^1}^{(Q)} \\ \Omega_{s^2}^{(1)} & \dots & \Omega_{s^2}^{(Q)} \\ \vdots & & \vdots \\ \Omega_{s^Q}^{(1)} & \dots & \Omega_{s^Q}^{(Q)} \end{bmatrix} \text{ where}$$

$$\Omega^q_{s^p, \pi_p(i)\pi_q(j)} = \sum_{v_k \in \mathcal{V}_{iq}} \frac{1}{|\mathcal{U}_k|} K_q \left(x_{v_{k,q}}^{(q)}, x_j^{(q)} \right), \tag{46}$$

for all $p, q = 1, \dots, Q$ and for all $i, j \in \mathcal{A}_q$ where $\pi_q : \mathbb{N} \to \mathbb{N}$ enumerates all elements of the set $\mathcal{A}_q$. Hence the result (41) follows. ■

IV. Experiments

A. Artificial dataset

A modified version of the Ripley dataset was analyzed using the proposed techniques in order to illustrate the differences between existing methods. While the original dataset consists of 250 samples to be used for training and model selection and 1000 samples for the purpose of testing, only 50 samples of the former where taken for the purpose of training in order to keep the computations tractable. The remaining 200 were used for the purpose of tuning the regularization constant and the kernel parameters. 15 observations out of the 50 are then considered as missing. Let the 50 training samples have a balanced class distribution. Numerical results are reported in Table I illustrating that the proposed method outperforms common practice of median imputation of the inputs and omitting the incomplete observations. Note that even without incorporating the multivariate structure and using the modification to the standard SVM, an increase in performance can be observed.

	PCC testset	STD
	Ripley Dataset (50;200;1000)	
Complete obs.	0.8671	0.0212
Median Imputation	0.8670	0.0213
SVM&mv (III.A)	0.8786	0.0207
cSVM&mv (III.B)	**0.8939**	**0.0089**
cSVM& $\mathcal{M}$ (II.D)	0.6534	0.1533
LS-SVM&mv (III.C)	0.8833	0.0184
cLS-SVM&mv (III.C)	0.8903	0.0208
	Hepatitis Dataset (85;20;50)	
Complete obs. cSVM	0.5800	0.1100
Median Imputation cSVM	0.7575	0.0880
SVM&mv (III.A)	0.7825	0.0321
cSVM&mv (III.B)	0.8375	0.0095
cSVM& $\mathcal{M}$ (II.D)	0.7550	0.0111
LS-SVM&mv (III.C)	0.7700	0.0390
cLS-SVM&mv (III.C)	**0.8550**	**0.0093**

TABLE I

Numerical results of the case studies described in Subsection IV.A and IV.B respectively based on a Monte Carlo simulation. Results are expressed in Percentage Correctly Classified (PCC) on the test-set. The roman capitals refer to the Subsection in which the method is described. In the case of the artificial dataset based on the Ripley dataset, the advantage of the proposed methods over median imputation of the inputs or the complete case analysis is outperformed, even without the use of the componentwise method. In the case of the Hepatitis dataset, the componentwise LS-SVM taking into account the missing values outperforms the other methods.

This setup was employed in a Monte-Carlo study of 500 randomizations were in each the assignment of data to training-, validation- and test-set is randomized and values of the training-set are indicated as missing at random. From the results, it may be concluded that the proposed approach outperforms median inputation even when one does not employ the componentwise strategy to recover the partially observed values per observation. Figure 2 displays the results of one single experiment with two components corresponding to X^1 and X^2 and their corresponding predicted output distributions.

B. Benchmark dataset

A benchmark dataset of the UCI repository was taken to illustrate the effectiveness of the employed method on a real dataset. The hepatitis dataset consists of a binary classification task with 19 attribute values and a total of 155 samples and containing 167 missing values. A test-set of 50 complete samples and a validation-set of 20 complete samples were withdrawn for the purpose of model comparison and tuning the regularization constants.

These results suggest the appropriateness of the assumption of additive models in this case study even with regard to generalization performance. By omitting the components which have only a minor contribution to the obtained model, one additionaly gains insight in the model as illustrated in Figure 3.

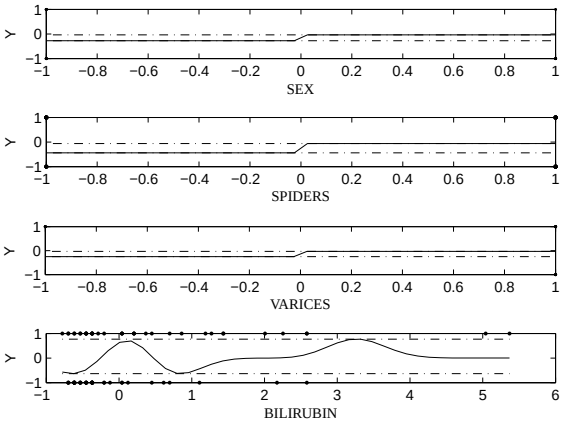


Fig. 3. *The four most relevant contributions for the additive classifier trained on the Hepatitis dataset using the componentwise LS-SVM as explained in Subsection III.C are function of the SEX of the patient, the attributes SPIDERS, VARICES and the amount of BILIRUBIN respectively.*

V. Conclusions

This paper studied a convex optimization approach towards the task of learning a classification rule from observational data when missing values occur amongst the input variables. The main idea is to incorporate the uncertainty due to the missingness into an appropriate risk function. An extension of the method is made towards multivariate input data by adopting additive models leading to componentwise SVMs and LS-SVMs respectively.

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

Pelckmans, K., De Brabanter J., Suykens, J.A.K., & De Moor, B. (2005a). Maximal variation and missing values for componentwise support vector machines. In *Proceedings of the international joint conference on neural networks (IJCNN 2005)*. Montreal, Canada: IEEE.

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Bousquet, O., Boucheron, S., & Lugosi, G. (2004). Introduction to statistical learning theory. *in* Advanced Lectures on Machine Learning Lecture Notes in Artificial Intelligence, *eds. O. Bousquet and U. von Luxburg and G. Rätsch, 3176*. (Springer)

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge University Press.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Chapman and Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Heidelberg: Springer-Verlag.

Hoeffding, W. (1961). *The strong law of large numbers for u-statistics.* Univ. North Carolina Inst. Statistics Mimeo Series, No. 302.

Lee, A. (1990). *U-statistics, theory and practice*. New York: Marcel Dekker.

Pelckmans, K., Goethals, I., De Brabanter, J., Suykens, J.A.K., & De Moor, B. (2005b). Componentwise least squares support vector machines. Chapter in *Support Vector Machines: Theory and Applications*, L. Wang (Ed.), Springer.

Pelckmans, K., Suykens, J.A.K., & De Moor, B. (2005c). Building sparse representations and structure determination on LS-SVM substrates. *Neurocomputing*, *64*, 137-159.

Pestman, W. (1998). *Mathematical statistics*. New York: De Gruyter Textbook.

Rubin, D. (1976). Inference and missing data (with discussion). *Biometrika*, *63*, 581-592.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Saunders, C., Gammerman, A., and Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th int. conf. on machine learning(ICML'98)* (p. 515-521). Morgan Kaufmann.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

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

Suykens, J.A.K., De Brabanter, J., Lukas, L., & De Moor, B. (2002). Weighted least squares support vector machines : robustness and sparse approximation. *Neurocomputing*, *48*(1-4), 85-105.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. World Scientific, Singapore.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, *58*, 267-288.

Vapnik, V. (1998). *Statistical learning theory*. Wiley and Sons.