Elsevier Editorial System(tm) for Artificial Intelligence in Medicine
Manuscript Draft

Manuscript Number: AIIM-D-12-00314R3

Title: White box radial basis function classifiers with component selection for clinical prediction models

Article Type: Original Research Paper

Keywords: Interpretable support vector machines; radial basis functions; white box methods; feature selection; clinical decision support

Corresponding Author: Dr. Vanya Van Belle, Dr. Ir.

Corresponding Author's Institution: KU Leuven

First Author: Vanya Van Belle, Ir.

Order of Authors: Vanya Van Belle, Ir.; Paulo Lisboa

ESAT/SISTA
KU LEUVEN
KASTEELPARK ARENBERG 10/2446
B-3001 HEVERLEE (LEUVEN)
BELGIUM

KATHOLIEKE
UNIVERSITEIT
LEUVEN

HEVERLEE, 2013–09–30

Dear Editor,

Please find in attachment the again revised version of the manuscript *White box radial basis function classifiers with component selection for clinical prediction models*, which we would like to be considered for publication in *Artificial Intelligence in Medicine*. The manuscript was adapted to incorporate the comments of the reviewer. A detailed description of the changes is provided with the submission. Additional changes were included as requested by the editor.

We have no personal financial interest in this subject, and none of the authors has any conflict of interest. All authors have read and approved the final version of this manuscript.

On behalf of all authors,

Vanya Van Belle
ESAT-SCD/SISTA
Kasteelpark Arenberg 10/2446
B-3001 Leuven
BELGIUM

DR. IR. VANYA VAN BELLE
TEL. +32-16-321065 FAX +32-16-321970
E-mail: vanya.vanbelle@esat.kuleuven.be

Answers to the review of *White box RBF classifiers with component selection for clinical prediction models*

We thank the reviewers for their constructive comments. We changed the manuscript according to the advice of the reviewers. Concrete changes are indicated below. The references were completed such that they contain the editors, publisher and location of the publisher for workshop and conference proceedings. For two papers, the authors were not able to find this information.

Reviewer 1

1) You added "However, the proposed methodology can also be extended to more complex components when necessary". What do you mean by that? Should it be discussed in the Conclusion instead? Do you mean that you can extend your work to k-way interactions effects? If yes this is not obvious nor trivial, as you limited your equations to two-way interaction to be able to interpret the results.

The text "However, the proposed methodology can also be extended to more complex components when necessary" was moved to the conclusion and updated as follows: "While the proposed methodology is focused on linear and bivariate interactions, it can in principle be extended to more complex effects. This would be non-trivial to interpret, although the methodology for sparse modeling does extend to the consideration of third or higher-order terms, this would add considerable complexity to the method. Nevertheless it is possible to do this in principle and it may be that noise in the data removes much of the complexity. This matter is of interest for further work."

2) "restricting the components to the components" => restricting the components to the ones ?

We thank the reviewer for this comment and changed manuscript accordingly.

3) In an added paragraph: "It is therefor proposed" => therefore

We thank the reviewer for this comment and changed manuscript accordingly.

White box radial basis function classifiers with component selection for clinical prediction models

Vanya Van Belle[a,b,1], Paulo Lisboa[b]

[a] *Department of Electrical Engineering / iMinds Future Health Department, KU Leuven, Kasteelpark Arenberg 10 / 2446, 3001 Leuven, Belgium*

[b] *Department of Mathematics and Statistics, Liverpool John Moores University, Byrom Street, Liverpool L3 5UX, UK*

Abstract

Objective: To propose a new flexible and sparse classifier that results in interpretable decision support systems.

Methods: Support vector machines (SVM) for classification are very powerful methods to obtain classifiers for complex problems. Although the performance of these methods is consistently high and non-linearities and interactions between variables can be handled efficiently when using non-linear kernels such as the radial basis function (RBF) kernel, their use in domains where interpretability is an issue is hampered by their lack of transparency. Many feature selection algorithms have been developed to allow for some interpretation but the impact of the different input variables on the prediction still remains unclear. Alternative models using additive kernels are restricted to main effects, reducing their usefulness in many applications. This paper proposes a new approach to expand the RBF kernel into interpretable and visualizable components, including main and two-way interaction effects. In order to obtain a sparse model representation, an iterative l_1-regularized parametric model using the interpretable components as inputs is proposed.

Results: Results on toy problems illustrate the ability of the method to select the correct contributions and an improved performance over standard RBF classifiers in the presence of irrelevant input variables. For a 10-dimensional x-or problem, an SVM using the standard RBF kernel obtains an area under the receiver operating characteristic curve (AUC) of 0.947, whereas the proposed method achieves an AUC of 0.997. The latter additionally identifies the relevant components. In a second 10-dimensional artificial problem, the underlying class probability follows a logistic regression model. An SVM with the RBF kernel results in an AUC of 0.975, as apposed to 0.994 for the presented method. The proposed method is applied to two benchmark datasets: the Pima Indian diabetes and the Wisconsin breast cancer dataset. The AUC is in both cases comparable to those of the

[1]Corresponding author:
Tel.: +32 16 32 10 65
Fax: +32 16 32 19 86
E-mail address: vanya.vanbelle@esat.kuleuven.be

standard method (0.826 versus 0.826 and 0.990 versus 0.996) and those reported in the literature. The selected components are consistent with different approaches reported in other work. However, this method is able to visualize the effect of each of the components, allowing for interpretation of the learned logic by experts in the application domain.
Conclusions: This work proposes a new method to obtain flexible and sparse risk prediction models. The proposed method performs as well as a support vector machine using the standard RBF kernel, but has the additional advantage that the resulting model can be interpreted by experts in the application domain.

Keywords:
Interpretable support vector machines, radial basis functions, white box methods, feature selection, clinical decision support

---

1. Introduction

Machine learning methods [1–3] are increasingly used to classify data. They are specifically powerful in higher dimensions and when the effects of the variables are assumed to be non-linear or interacting with each other. A disadvantage of these methods is their inherent black-box nature and as such the resulting models do not reveal any information on the contribution of each specific input variable on the predicted outcome. In many applications, such as medical and financial decision making, interpretability of the prediction model is considered more important than best performance. The use of standard machine learning methods in practice is therefore hampered in these domains.

Interpretability of prediction models can have different meanings. In this work we will concentrate on two parts of interpretable models. Firstly, unnecessary variables should be discarded in the final model. Secondly, the impact of the value of the different input variables on the prediction should be clear. Both of these requirements have been studied in the literature, but weaknesses in the proposed approaches still remain and methods simultaneously tackling both aspects are rare. Different feature selection methods for support vector machines (SVM) and in extension for least-squares support vector machines have been proposed. Three main approaches can be identified. A first approach filters irrelevant inputs out before building the classifier on the selected set. One possibility is to rank inputs according to some criterion, e.g. Fisher's criterion, Pearson correlation or mutual information criteria [4, 5]. More advanced approaches such as relief and focus have been proposed in [6–8]. Although filter approaches are very efficient w.r.t. computation, this approach might not be optimal [9, 10]. A second approach involves wrappers that use the performance of a specific classifier to rank subsets of variables. The least informative input (or set of inputs) is removed in an iterative procedure until convergence. One example is the recursive feature elimination SVM [11], that iteratively eliminates the input with the lowest difference in the margin when calculating the kernel matrix without this input. Similar approaches using different ranking functions were proposed in [12, 13]. More recent work has focused on the embedding of feature selection within the classifier. Many of these approaches solve the feature selection task by replacing the 2-norm in standard SVMs by

a 0-norm, a 1-norm or approximations and combinations of these [14–18]. A drawback of these approaches is that feature selection is performed in the primal model formulation, restricting its use to linear models. Several methods are reported to deal with feature selection in the dual formulation. However, these methods most often result in sparsity in the features determined in feature space and not in the input space. Since the resulting features can not be interpreted in function of the input variables, these methods are not suitable for applications where interpretability is an issue. Only some approaches study the combination of feature selection in input space while optimizing the dual problem formulation as (a relaxation of) mixed integer programming problems [19, 20]. Maldonado [21] proposed to learn an anisotropic kernel, where the bandwidth w.r.t. the different inputs was varied and inputs with a large bandwidth are subsequently eliminated.

Another approach that is often used to enable interpretation of SVMs are rule extraction methods [22, 23]. However, the approach of these methods is quite different from the one presented in this manuscript. The learned rules give an explanation of the model but they are not equal to the model. The rules only mimic the original model and are thus an approximation of the learned logic of the SVM. Decision rules are a binary approximation to the smooth response function. Our method makes the response function explicit in its variable specific components and for pairwise interactions. Additionally, there is no mechanism controlling the difference in performance between the original model and the learned rules. The intention of this work is to provide flexible methods that are interpretable by design, and contain an explicit control mechanism on the performance.

In order to allow for an explanation of the model's prediction, models are often restricted to be additive [24, 25]. Thanks to the additive structure, the contribution of each input variable to the prediction is clear. However, several classification problems can not be solved using a sum of main effects. The use of ANOVA models [26], extending the additive structure to incorporate a number of predefined interaction terms, offers a solution to this problem. In its general form, the ANOVA decomposition is composed as the sum of the main effects and all possible combinations of inputs. For most practical applications demanding an interpretable prediction model, reducing this decomposition to main and two-way interaction effects is sufficient [27, 28]. An additional advantage of this approach is the possibility to visualize the effects and thus enable validation of the resulting models by experts in the application domain. ANOVA models for component selection where proposed in [29–31]. The kernel approach taken by Gunn and Kandola [32] for regression problems is most strongly related to the work presented here for classification. They replace the kernel by means of a weighted sum of kernels. The problem is then solved by iteratively solving two convex optimization problems: (i) solve the problem in the Lagrange multipliers, fixing the weights in the sum of kernels; and (ii) solve the problem in the weights, fixing the Lagrange multipliers. Their approach is restricted to kernels without hyperparameters to reduce computational load.

The goal of this work is to combine component selection with SVMs using the radial basis function (RBF) kernel in order to obtain flexible but interpretable models. We propose to replace the RBF kernel by a truncated version, containing only main and two-way interaction effects. Using this kernel, a standard SVM is solved. In a second

step, the different contributions to the prediction of the SVM classifier are calculated and used as input variables for a linear and iteratively reweighted l_1-regularized SVM. The result is a white box RBF classifier with component selection. In this work, we explicitly choose to restrict the components to main and two-way interaction effects to facilitate the visualization of the effect of the different components on the prediction. In most clinical research, main effects are considered and when assumed necessary, interactions are added [27, 28].

The remainder of the paper is organized as follows. Section 2 starts with introducing the notations used throughout the paper and summarizes support vector machines for classification. In Section 2.2 we illustrate how the RBF kernel can be represented as a sum of kernels evaluated on subsets of the input variables. Section 2.3 proposes a method to obtain sparse results. Section 2.4 indicates how the results can be interpreted in clinical practice. Section 3 discusses the model selection aspects of this work. Our approach is illustrated on toy problems and real life classification problems in Section 4. Section 5 summarizes some final conclusions.

2. A white box RBF classifier

In this Section, we propose a novel approach to obtain sparse and interpretable classifiers that are able to select relevant (non-)linear and interaction effects. The standard RBF kernel is truncated to only include main and two-way interaction effects. These effects are then combined in a sparse way by solving an iteratively reweighted l_1-regularized SVM in primal space.

2.1. Support vector classifier

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be a set of observations, with $x_i \in \mathbb{R}^d$ the input variables of observation i and $y_i \in \{-1, 1\}$ the corresponding class label. The standard SVM for classification [1] is then formulated as

$$\begin{aligned} \min_{w,b,\epsilon} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{N} \epsilon_i \\ \text{subject to} \quad & \begin{cases} y_i \left(w^T \varphi(x_i) + b \right) \geq 1 - \epsilon_i, & \forall\ i = 1, \ldots, N \\ \epsilon_i \geq 0, & \forall\ i = 1, \ldots, N\,. \end{cases} \end{aligned} \tag{1}$$

In this notation, $\varphi(\cdot)$ represents a feature map, mapping the input variables into a (possibly infinite) feature space; $w \in \mathbb{R}^{d_\varphi}$ is a coefficients vector and γ is a strict positive regularization parameter making the trade-off between smoothness and correct classification of the training data. When solving this problem in primal space, the feature map needs to be specified explicitly and a prediction for a new point $x_\star$ is obtained from

$$\hat{y} = \text{sign}(w^T \varphi(x_\star) + b)\,.$$

Defining the Lagrangian of problem (1), and deriving the Karush-Kuhn-Tucker conditions yields the dual problem formulation

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \varphi(x_i)^T \varphi(x_j)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i \\ \text{subject to} \quad & \begin{cases} \displaystyle\sum_{i=1}^{N}\alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq \gamma, \quad \forall\, i = 1,\ldots,N\,. \end{cases} \end{aligned} \tag{2}$$

An advantage of this approach is that the feature map $\varphi(x)$ does not need to be constructed explicitly. Any continuous function $K(x, x_\star)$ for any points x and $x_\star$ satisfying Mercer's condition [33] can be expressed as an inner product

$$K(x, x_\star) = \varphi(x)^T\varphi(x_\star)\,.$$

The classifier then becomes

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{N}\alpha_i y_i K(x_i, x_\star) + b\right)\,.$$

In many applications, the RBF is chosen as the kernel since it is able to model non-linearities and interactions between variables automatically and is bounded. A drawback of using a non-additive kernel like the RBF is that the resulting classifier is a black-box model, not revealing any information on the way the predictions are obtained. In the next Section, it is shown how the RBF kernel can be approximated to obtain a white box classifier.

2.2. Truncated radial basis functions

Several additive kernels, such as the polynomial and clinical kernel [34], can be used to enable interpretability. However, in practice not all problems can be solved by main effects. ANOVA kernels offer a solution to this problem [26], but prior knowledge is needed in order to define which terms should be included in the ANOVA decomposition.

In this work, we propose to expand the RBF kernel and to truncate its contributions to main and two-way interaction effects as follows. The RBF kernel is defined as $K_{\text{RBF}}(x, z) = \exp\left(-\frac{||x-z||_2^2}{\sigma^2}\right)$, with x and $z \in \mathbb{R}^d$. Using the Taylor expansion of the exponential function $\exp(x) = \sum_{n=0}^{\infty}\frac{x^n}{n!}$, the RBF kernel can be written as

$$K_{\text{RBF}}(x, z) = \sum_{n=0}^{\infty}\frac{(-1)^n(||x-z||_2^2)^n}{n!\sigma^{2n}}\,.$$

Using the multinomial theorem

$$(x^1 + x^2 + \cdots + x^d)^n = \sum_{k_1+\cdots+k_d=n}\binom{n}{k_1,\ldots,k_d}\prod_{1\leq p\leq d}(x^p)^{k_p}\,,$$

with x^p the p^{th} variable of x, this becomes

$$
\begin{aligned}
K_{\text{RBF}}(x,z) &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \left[\sum_{p=1}^{d} (x^p - z^p)^{2n} + \sum_{\substack{\sum_{l=1}^{d} k_l = n \\ k_l \neq n}} \binom{n}{k_1, \ldots, k_d} \prod_{1 \leq p \leq d} (x^p - z^p)^{2k_p} \right] \\
&= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \sum_{p=1}^{d} (x^p - z^p)^{2n} \\
&\quad + \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \sum_{\substack{k_p + k_q = n \\ k_p, k_q \neq n}} \binom{n}{k_p, k_q} (x^p - z^p)^{2k_p} (x^q - z^q)^{2k_q} \qquad (3) \\
&\quad + \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \sum_{\substack{\sum_{l=1}^{d} k_l = n \\ k_l \neq n \\ k_l + k_m \neq n}} \binom{n}{k_1, \ldots, k_d} \prod_{1 \leq p \leq d} (x^p - z^p)^{2k_p} .
\end{aligned}
$$

The first term in (3) represents the contributions of single input variables (main effects), the second term represents all two-way interaction effects and the last term represents all interaction effects with more than two variables involved. In order for the results to be interpretable and explainable, we will focus on the first two terms since these can be visualized. For most applications where interpretability is an issue it suffices to take two-way interactions into account. Using equation (3), the RBF kernel evaluated on a 2-dimensional test point $x^{p,q} = [x^p, x^q]^T$ and an rbf center $z^{p,q} = [z^p, z^q]^T$ can be expressed as

$$
\begin{aligned}
K_{\text{RBF}}(x^{p,q}, z^{p,q}) &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \left[(x^p - z^p)^{2n} + (x^q - z^q)^{2n} \right] \qquad (4) \\
&\quad + \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \sum_{\substack{k_p + k_q = n \\ k_p, k_q \neq n}} \binom{n}{k_p, k_q} (x^p - z^p)^{2k_p} (x^q - z^q)^{2k_q} \\
&= \exp\left(-(x^p - z^p)^2 / \sigma^2 \right) + \exp\left(-(x^q - z^q)^2 / \sigma^2 \right) \\
&\quad + \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\sigma^{2n}} \sum_{\substack{k_p + k_q = n \\ k_p, k_q \neq n}} \binom{n}{k_p, k_q} (x^p - z^p)^{2k_p} (x^q - z^q)^{2k_q} ,
\end{aligned}
$$

and contains the main effects of both input variables and their interaction effect. The truncated RBF kernel is then defined as the summation of RBF kernels evaluated for

every pair of coordinates p and q:

$$K^{\text{tr}}_{\text{RBF}}(x,z) = \frac{2}{d(d-1)} \sum_{p=1}^{d} \sum_{q>p}^{d} K_{\text{RBF}}(x^{p,q}, z^{p,q}) \,.$$

Replacing the RBF kernel with its truncated version, the prediction of the classifier for a new point $x_\star$ is obtained from

$$\begin{aligned}
\hat{y} &= \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K^{\text{tr}}_{\text{RBF}}(x_i, x_\star) + b \right) \\
&= \text{sign}\left(\frac{2}{d(d-1)} \sum_{i=1}^{N} \alpha_i y_i \left(\sum_{p=1}^{d} \sum_{q>p}^{d} K_{\text{RBF}}(x_i^{p,q}, x_\star^{p,q}) \right) + b \right) \\
&= \text{sign}\left(\frac{2}{d(d-1)} \sum_{p=1}^{d} \sum_{q>p}^{d} \sum_{i=1}^{N} \alpha_i y_i K_{\text{RBF}}(x_i^{p,q}, x_\star^{p,q}) + b \right) \\
&= \text{sign}\left(\sum_{p=1}^{d} \sum_{q>p}^{d} \hat{y}^{p,q} + b \right) \,.
\end{aligned}$$

2.3. Parsimonious RBF classifiers

In order to achieve a sparse model representation, including only a subset of main and interaction effects, the following approach is proposed. First, the different terms $\hat{y}^{p,q}$ are split into main and two-way interaction effects. Second, they are used as inputs in a parametric model with sparsity constraints on the coefficients, such that only some of them will be different from zero. The exact methodology is explained below.

The partial contributions $\hat{y}^{p,q}$ are weighted sums of RBF kernels that are evaluated in 2-dimensional vectors, i.e. each RBF kernel is evaluated for a specific pair of covariate dimensions. As such, the partial contribution $\hat{y}^{p,q}$ contains the main effects of both variables and an interaction effect. In order to be able to select all of these effects separately, $\hat{y}^{p,q}$ is extracted in three components: (i) $\hat{y}^{p}$ which is built upon the first term in equation (4), (ii) $\hat{y}^{q}$ which is built upon the second term in equation (4), and (iii) a contribution of the interaction expressed as $\hat{y}^{p,q} - \hat{y}^{p} - \hat{y}^{q}$. Note that this extraction can not be made on the level of the kernel due to the necessity of the kernel in the SVM classifier to be positive semidefinite.

Before these components are used as inputs for a parametric model, they are standardized as usual. Let $\tilde{y}^{p}$ be the normalized version of $\hat{y}^{p}$ with zero mean and a standard deviation of 1 and $\tilde{y}^{p,q}$ the normalized version of $\hat{y}^{p,q} - \hat{y}^{p} - \hat{y}^{q}$ and denote these as the partial contributions or components of the predictor. These partial contributions are then used as inputs for a linear and iteratively reweighted l_1-regularized SVM classifier [17] with non-negative coefficients, such that only some of the components will be selected. The model is then formulated as:

$$\min_{\beta,b^*,\epsilon^*} \sum_{p=1}^{d} \chi^p \beta^p + \sum_{p=1}^{d} \sum_{q>p}^{d} \chi^{p,q} \beta^{p,q} + \gamma^* \sum_{i=1}^{N} \epsilon_i^*$$

subject to

$$\begin{cases} y_i \left(\sum_{p=1}^{d} \beta^p \tilde{y}^p + \sum_{p=1}^{d} \sum_{q>p} \beta^{p,q} \tilde{y}^{p,q} + b^* \right) \geq 1 - \epsilon_i^*, & \forall\ i = 1, \ldots, N \\ \epsilon_i^* \geq 0, & \forall\ i = 1, \ldots, N \\ \beta^p \geq 0, & \forall\ p = 1, \ldots, d \\ \beta^{p,q} \geq 0, & \forall\ p = 1, \ldots, d; q = p+1, \ldots, d\,, \end{cases} \tag{5}$$

where χ^p equals 1 in the first iteration and is defined as

$$\chi^p = \frac{1}{\varepsilon + c\beta^p}\,, \tag{6}$$

in the next iterations. Here, ε is a small, predefined constant (e.g. 0.005) and c a parameter to control the sparsity of the solution [35]. The value of χ^p is chosen such that components with a small (large) coefficient receive a large (small) weight. As such, small coefficients will be further penalized and will shrink to zero in subsequent iterations. Larger coefficients will receive less impact on the cost function and the corresponding components will be used in the final model. Method (5) is iterated until the average of the absolute value of the difference between the β-vectors in two iterations is less then 10^{-8}. The 1-norm penalty was first introduced by [36] as the Least Absolute Shrinkage and Selection Operator in the context of linear regression. In equation (5) the coefficients are restricted to be non-negative since all the components are assumed to positively correlate with the outcome. This is similar to the non-negative garrote estimator [29, 37], which was originally proposed to shrink the estimates from least-squares regression.

The procedure to obtain an interpretable and sparse classifier is summarized in Algorithm 1. A description on how the different parameters are tuned in our experiments follows in Section 3. The results can be further improved by iterating until the selected set of components remains unchanged. In practical applications, this is achieved after two to four iterations.

2.4. Clinical interpretation of the results

In this Section, we indicate how the proposed method will enable clinical interpretation of the resulting models. In a clinical setting, prognostic indices are often used to indicate the severity of an illness. An example is found in the body mass index, where a high body mass index indicates higher risks for diabetes and cardiovascular diseases. Additionally, cut-offs are used to make formal diagnosis. In the case of the body mass index, the cut-off is set to 30 in order to diagnose obesity. For a standard SVM approach, the prognostic

Algorithm 1 Procedure to obtain a sparse white box RBF classifier.

1: Determine the optimal tuning parameters γ and σ for the truncated RBF kernel in (2).
2: Given the optimal value of γ and σ, solve equation (2) to obtain α.
3: Given α, estimate the partial contributions $\tilde{y}^p$ and $\tilde{y}^{p,q}$.
4: Given the partial contributions, determine the optimal value of γ^* in equation (5) with a fixed value of $c = 1$.
5: Given γ^*, determine the optimal value of c in equation (5).
6: Given c and γ^*, solve equation (5) to obtain the sparse model representation.
7: Obtain the prediction as

$$\hat{y} = \text{sign}\left(\sum_{p=1}^{d} \beta^p \tilde{y}^p + \sum_{p=1}^{d}\sum_{q\geq p}^{d} \beta^{p,q}\tilde{y}^{p,q} + b^*\right).$$

index is defined as $\sum_{i=1}^{N} \alpha_i y_i K(x_i, x_\star)$. Since it is unknown how each input contributes to this weighted kernel sum, interpretation of the decision process is not possible. For the proposed method, the prognostic index becomes $\sum_{p=1}^{d} \beta^p \tilde{y}^p + \sum_{p=1}^{d}\sum_{q\geq p}^{d} \beta^{p,q}\tilde{y}^{p,q}$, which is a (sparse) linear combination of contributions that are allocated to specific inputs. In order to visualize these contributions, figures can be made representing $\beta^p\tilde{y}^p$ for main effects and $\beta^{p,q}\tilde{y}^{p,q}$ for interaction effects, where the components with $\beta = 0$ are dropped. In each of these figures, a higher contribution will indicate that the corresponding input value will target the prediction towards the positive class and a lower contribution to the negative class. The bias term b^* serves as cut-off. When presenting the weight (kg) and height (m) of a person together with the outcome of obesity, the presented method would select one component, being the interaction between weight and height. $\tilde{y}^{\text{weight,height}}$ will be related to the function $f(\text{weight}, \text{height}) = \frac{\text{weight}}{\text{height}^2}$. The b^* will be the value best dividing obese from non-obese persons using $\beta^{\text{weight,height}}\tilde{y}^{\text{weight,height}}$ as prognostic index.

As will be further discussed in Section 4.3, it is possible that the presented method selects different components with slight modifications on the dataset. In those circumstances, the method indicates that different models are able to achieve the same performance using different inputs. When this occurs in a clinical application, a discussion between model developer and end user should be initiated and the clinically most relevant model should be selected. Two models that perform equally good from a statistical point of view might be very different from a clinical perspective, and one of them might even be impractical or not logical. The goal of this work is exactly to enable the detection of clinically irrelevant models with a good performance. This approach is also able to identify problems with the data, which would not become apparent when using black-box models.

3. Tuning of the parameters

The performance of the proposed approach depends on the value of several parameters. In addition to the tuning of the parameters involved in standard SVMs, other parameters

need to be set to an appropriate value and the optimal value of some of them are related. In the experiments, the parameters were tuned according to the following scheme (see Figure 1).

Tuning of the bandwidth of the (truncated) RBF and the regularization parameter γ in equation (2) is performed by means of coupled simulated annealing [38]. The parameter values were randomly initialized, where σ was scaled with $\sqrt{d}$. The procedure started from 10 different initializations. The parameter combination leading to the best 10-fold cross-validation area under the receiver operating characteristic curve (AUC) was selected. Once γ and σ are tuned, the Lagrange parameters α and b can be defined and the partial contributions can be calculated.

To select which components are relevant, c and γ^* need to be tuned. However, the optimal values of these parameters are related. A high value of γ^* inhibits a sparse solution, whatever the value of c. Tuning both parameters simultaneously would necessitate the use of a risk measure capturing the trade-off between sparsity and performance. Since it is not clear in advance which trade-off is realistic, this choice is left open for discussion. In the experiments, the value of γ^* was tuned by means of 5-fold cross-validation, with $c = 1$. The grid over which γ^* was varied was defined as an exponential grid on $[0.01, 1000]$. The AUC was used as model selection criterion. Using the tuned value of γ^*, the value of c was varied, and the 5-fold cross-validation AUC was reported. To reduce computational load, the range of values over which c is varied was restricted to values for which the resulting coefficients vector β yielded 1 to $3d$ non-zero elements. The optimal value of c was defined as the lowest value yielding an AUC on 5-fold cross validation that did not lead to a significant reduction in AUC ($p>0.05$) according to the test of DeLong [39]. In order to be able to compare the results over folds, a logistic regression model was trained in all training folds and applied to the test fold, in order to convert the uncalibrated latent variables to calibrated probabilities [40].

4. Results

This Section illustrates the use of the presented method on artificial and real-life data. In the artificial experiments, a training and test set were created. The tuning (as explained above) was performed on the training set and the performance is reported on the test set. In both real-life applications, the available data was 10 times randomly split into a training (two thirds of the total data) and validation set (the remainder of the total data). The mean performance on the test data is reported. The components that are selected more than 5 times are then used to train the final model on the complete data set. The results are then visualized for clinical interpretation.

Toy problems illustrate the ability of the model to detect the relevant components, whilst being as performant as standard SVMs. Two datasets from the UCI machine learning depository [41] are used to compare our results with results from other methods described in the literature. In all the experiments γ and γ^* were scaled with $\frac{N}{N_+}$ and $\frac{N}{N_-}$, where N_* indicates the number of observations in class $*$, for elements belonging to the positive and negative class respectively. The method was iterated until the set of selected

components did no longer change or the maximal number of iterations (here 10) was exceeded. In each iteration the components selected in the previous iteration were taken into account in addition to the main effects of the variables involved in a selected interaction effect. The reported confidence intervals where calculated by means of the bias corrected and accelerated percentile method using 1000 bootstrap samples. The AUC, accuracy (acc) and balanced error rate (ber) are reported, together with a 95% confidence interval (95% CI) or standard deviation (std).

4.1. Artificial example 1: the x-or problem

In this first experiment, the x-or problem is considered in three different settings: 2-dimensional, 4-dimensional and 10-dimensional. In all three settings, only the first two variables are relevant. All variables are independently drawn from a uniform distribution and are in the range $[0, 1]$. The observations belong to the first class if $x_1 \leq 0.5$ and $x_2 > 0.5$ or $x_1 > 0.5$ and $x_2 \leq 0.5$ and to the second class otherwise. The proposed method (l_1-svm-rbf$^{\text{tr}}$) selects a single component in all three cases: the interaction between the first and second input variable (see Figure 2). Table 1 compares the results of the presented method with two standard SVMs using an RBF kernel: one using all variables, and one using those variables that were selected by our method. Note that when our method selects a single interaction effect, the use of the standard RBF kernel involves using that effect together with both main effects. Since our method is able to build a classifier only using a selected set of variables, the performance does not drop when increasing the number of irrelevant features. The standard SVM classifier suffers from overfitting when more irrelevant features are included. The performance increases again when restricting the used feature set to the ones selected by our method.

4.2. Artificial example 2

In this second experiment, a classification problem with an underlying logistic regression function is used to illustrate the ability of the model to select the correct relevant variables. A dataset with 10 variables is created by means of a multivariate Gaussian distribution. The variables are uncorrelated except for the first three variables, whose correlation matrix is

$$\begin{bmatrix} 1 & 0.8 & 0.2 \\ 0.8 & 1 & 0.1 \\ 0.2 & 0.1 & 1 \end{bmatrix}.$$

The probability of an observation to belong to the positive class is modeled by

$$P(\text{class } 1|\mathrm{x}_1, \ldots, \mathrm{x}_{10}) = \frac{\exp(5\mathrm{x}_1 + 5\mathrm{x}_3 + 10\mathrm{x}_1\mathrm{x}_6)}{1 + \exp(5\mathrm{x}_1 + 5\mathrm{x}_3 + 10\mathrm{x}_1\mathrm{x}_6)}.$$

The method detects a main effect for x_3 and an interaction effect for x_1 and x_6. The main effect of x_1 is not selected. This is due to the fact that the split in main and interaction effects, without specification of the form of these effects, is not unique in an additive model. Additionally, the effect size of x_1 is smaller than the effect sizes for both other relevant

components, and might not be reflected in the AUC. Figure 3 illustrates the selected effects. Table 2 compares the results with the standard approach using all features and the selected subset (x_1, x_3, x_6). The performance of all methods are comparable but l_1-svm-rbf$^{\text{tr}}$ offers a way to interpret the results.

4.3. Stability analysis

In a last artificial example, the stability of the selected components and obtained performance is tested. The setting is the same as in the previous example, but the underlying model is now defined as

$$P(\text{class } 1|\mathrm{x}_1, \ldots, \mathrm{x}_{10}) = \frac{\exp(5\mathrm{x}_2 + 10\mathrm{x}_1\mathrm{x}_3)}{1 + \exp(5\mathrm{x}_2 + 10\mathrm{x}_1\mathrm{x}_3)}\,.$$

The dataset is split into a training and test set. Ten different initializations of the parameters γ and σ and different splits into folds are used to investigate the stability of the method. Variation in the parameters γ^* and c will be less since they are evaluated on a fixed grid. Their optimal value will vary by their dependence on the split in folds of the training set. The results are summarized in Table 3. In six out of ten initializations, the selected components are x_2 and x_1x_3. In the remaining four initializations, the correct components are selected but a subset of $\{x_1$,x_3,$x_2x_3\}$ is also selected. Due to correlations between variables and the non-unique split between main and interaction effects in an additive model, the method is not always able to select the components we expect. The method can therefore be stabilized by repeated subsampling of the training data. A final model can then be built on the complete training set, only including those components that are selected in the majority of the subsamples.

4.4. The Pima Indians Diabetes dataset

This dataset contains information on eight continuously measured variables for 768 females, aged 21 or more, of Pima Indian heritage. The goal is to predict whether these women have diabetes. Observations with a zero value for plasma glucose, body mass index or blood pressure (n=44) were assumed to be missing values and were removed from the dataset. The proposed method was performed on ten randomizations between training (two thirds of the data) and test set (one third of the data). The results are compared with a standard SVM using an RBF kernel with all inputs and the selected subset of inputs in Table 4. The proposed method is competitive with a standard SVM using an RBF kernel, but offers an interpretable model representation. Plasma glucose and body mass index are selected in all ten randomizations. Age is selected in 9 out of ten randomizations. In four cases, other variables are selected as well. Given these results, we trained a model on all the data, restricting the components to the ones that were selected in more than five randomization: the main effects of plasma glucose, body mass index and age. To make use of all available data, all data were used for this purpose. No performance measure of this final model is therefore reported. The model could be externally validated when new data is available. The estimated effects of the selected components are illustrated in Figure 4. The results show that it is possible to discriminate healthy persons from patients

with diabetes based on three inputs. It is seen that an increasing plasma glucose value corresponds to a higher contribution to the prognostic index and will influence the decision towards the positive class (diabetes). An increasing BMI has the same effect, up to 30 after which the effect plateaus. The increase after 40 will have a large variability due to the low number of patients it is based on. An increasing age also contributes to a higher prognostic index and thus targets to decision towards diabetes. After the age of 50, the effect drops. These results are consistent with the literature.

To validate the feature selection process, the results are compared with different results reported in the literature. Table 5 shows that the selected features were also identified as important features by different other types of feature selection and/or ranking methods. The obtained performance is comparable with the literature. Plasma glucose, body mass index and age are selected in respectively 8, 6 and 11 out of the 11 methods we compare with. Five of these methods use all three variables selected by the proposed method. We therefore conclude that the presented method is able to achieve a performance comparable to that of other methods reported in the literature, while restricting the number of necessary variables. The selected variables are among those that are used by most other methods found in the literature.

Figure 5 illustrates the sparsity performance (AUC) trade-off made by means of the value of c in equation (6) for the first randomization between training and test set. Using three components is as performing as using 19 components. The same pattern is seen in the other randomizations. Figure 6 visualizes the decision boundary. The conclusions correspond with those extracted from Figure 4.

4.5. The Wisconsin Breast Cancer dataset (original)

This dataset contains information on 699 women of whom 683 had complete information on all 9 variables. The variables in this dataset are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image and are all integers ranging from 1 to 10. Due to their ordinal nature, all variables were considered to be continuous. We used 10 randomizations between training (two thirds of the data) and the test set (one third of the data). In 7 cases, uniform shape and bare nuclei were selected; one case selected uniform shape and chromatin; 1 case selected bare nuclei and chromatin; one case selected uniform size and bare nuclei. Based on these results, a final model containing uniform shape and bare nuclei was trained on the complete data set. To make use of all available data, all data were used for this purpose. No performance measure of this final model is therefore reported. The model could be externally validated when new data is available.

The results are summarized in Table 6. The results are comparable but the presented method has the advantage of interpretability and sparsity in the number of components. The estimated effects are illustrated in Figure 7. It is seen that an increase in the uniformity of the cell shape or in the bare nuclei level correspond with a higher contribution to the prognostic index and will influence the decision towards the positive class (cancer). These results are consistent with the literature. The decision surface (see Figure 8) illustrates the

same and additionally shows that the estimated boundary separates both classes nearly perfectly.

To validate the feature selection process, the results are compared with different results reported in the literature. Table 7 shows that the selected features are among those selected by other feature selection methods. The performance is comparable to what is reported in the literature. Uniformity of cell shape and bare nuclei are selected in respectively 5 and 6 out of the 7 methods we compare with. Four of these methods use both variables selected by the proposed method. However, all methods from the literature use more variables than the proposed method. We therefore conclude that the presented method is able to achieve a performance comparable to that of other methods reported in the literature, while restricting the number of necessary variables. The selected variables are among those that are used by most other methods found in the literature.

5. Conclusions

This work proposed a novel approach to enable the use of support vector machines with RBF kernels in domains where interpretability of the resulting classifiers is an issue. An expansion of the RBF kernel in components that are visualizable allows validation of the estimated effects of the input variables by experts in the domain of the application. It was shown how the extracted components could be shrunk to obtain a sparse model representation. Results on toy and artificial problems illustrate the ability of the model to select relevant main and two-way interaction effects. Comparison of the results on two benchmark datasets illustrates that the proposed method is competitive with other classifiers, but has the advantage of being interpretable. While the proposed methodology is focused on linear and bivariate interactions, it can in principle be extended to more complex effects. This would be non-trivial to interpret, although the methodology for sparse modeling does extend to the consideration of third or higher-order terms, this would add considerable complexity to the method. Nevertheless it is possible to do this in principle and it may be that noise in the data removes much of the complexity. This matter is of interest for further work.

The proposed approach has the advantage that it yields an interpretable model without restricting the effect of the inputs to linear and/or main effects. In contrast to rule extraction methods, or other methods creating explanations of decision boundaries, this method is interpretable by design and as such does not result in an approximation of a black-box model. As for other parametric approaches, the results will not be satisfactory when the underlying structure is not realistic. However, this method has the advantage of being highly flexible. In addition, two-way interaction effects can be selected in an automated way. A disadvantage of the method w.r.t. black-box models is that the computational load and time largely increases with the number of inputs. It is therefore proposed to use the method with up to 20 inputs. Alternatively, the code could be parallelized to speed up computations.

The visualization of the resulting models enable to illustrate the model to experts in the application domains, such that the model can be validated in relation to expert knowledge

in addition to performance on unseen data. Stability analysis and the results of the sparsity mechanism can be used for interaction with these experts in order to select the model with the highest acceptation grade in the application domain. Thanks to the properties of this method, decision support developers and clinical practitioners will be able to interact. This type of interaction is indispensable for clinical decision support systems to be used in clinical practice.

Acknowledgments

Research supported by Research Council KUL: GOA MaNet, PFV/10/002 (OPTEC), several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, G.0108.11 (Compressed Sensing), G.0869.12N (Tumor imaging) , IWT: TBM070706-IOTA3, PhD Grants; iMinds; Belgian Federal Science Policy Office: IUAP P7/ (DYSCO, 'Dynamical systems, control and optimization', 2012-2017); EU: RECAP 209G within INTERREG IVB NWE programme, EU HIP Trial FP7-HEALTH/ 2007-2013 (n. 260777), ERC AdG A-DATADRIVE-B. VVB is a postdoctoral fellow of the Research Foundation - Flanders (FWO).

References

[1] Vapnik V. Statistical Learning Theory. Wiley and Sons, New York; 1998.

[2] Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters 1999;9(3):293–300.

[3] Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least Squares Support Vector Machines. World Scientific, Singapore; 2002.

[4] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7.

[5] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16(10):906–14.

[6] Kira K, Rendell LA. A practical approach to feature selection. In: Sleeman D, Edwards P, editors. Proceedings of the ninth international workshop on Machine learning. ML92; San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1-5586-247-X; 1992, p. 249–56.

[7] Sun Y. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 2007;29(6):1035 –51.

[8] Almuallim H, Dietterich TG. Learning with many irrelevant features. In: Dean T, McKeown K, editors. Proceedings of the Ninth National Conference on Artificial Intelligence. MIT Press, Cambridge, Massachusetts; 1991, p. 547–52.

[9] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997;97(1):273–324.

[10] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research 2003;3:1157–82.

[11] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002;46(1-3):389–422.

[12] Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. In: Leen TK, Dietterich TG, Tresp V, editors. Advances in Neural Information Processing Systems 13. Cambridge, MA, USA: MIT Press; 2000, p. 668–74.

[13] Rakotomamonjy A. Variable selection using svm based criteria. Journal of Machine Learning Research 2003;3:1357–70.

[14] Fung G, Mangasarian OL. A feature selection newton method for support vector machine classification. Tech. Rep. 02-03; Data Mining Institute, Computer Sciences Department, University of Wisconsin; Madison, Wisconsin; 2002. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/02-03.ps (Accessed: 02 April 2013).

[15] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence 1997;97:245–71.

[16] Weston J, Elisseeff A, Schölkopf B, Tipping M. Use of the zero norm with linear models and kernel methods. Journal of Machine Learning Research 2003;3:1439–61.

[17] Bradley PS, Mangasarian OL. Feature selection via concave minimization and support vector machines. In: Shavlik J, editor. Proceedings of the Fifteenth International Conference on Machine Learning. ICML; San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8; 1998, p. 82–90.

[18] Neumann J, Schnörr C, Steidl G. Combined svm-based feature selection and classification. Machine Learning 2005;61(1-3):129–50.

[19] Mangasarian OL, Kou G. Feature selection for nonlinear kernel support vector machines. In: Tung AKH, Zhu Q, Ramakrishnan N, Zaane OR, Shi Y, Clifton CW, et al., editors. Proceedings of the Seventh IEEE International Conference on Data Mining Workshops. ICDMW; IEEE, Piscataway. ISBN 0-7695-3033-8; 2007, p. 231–6.

[20] Tan M, Wang L, Tsang IWI. Learning sparse svm for feature selection on very high dimensional datasets. In: Fürnkranz J, Joachims T, editors. Proceedings of the 27th International Conference on Machine Learning. ICML; Madison, WI, USA: Omnipress; 2010, p. 1047–54.

[21] Maldonado S, Weber R, Basak J. Simultaneous feature selection and classification using kernel-penalized support vector machines. Information Sciences 2011;181(1):115–28.

[22] Etchells TA, Lisboa PJG. Orthogonal search-based rule extraction (osre) for trained neural networks: a practical and efficient approach. IEEE Transactions on Neural Networks 2006;17(2):374–84.

[23] Martens D, Huysmans J, Setiono R, Vanthienen J, Baesens B. Rule extraction from support vector machines: An overview of issues and application in credit scoring. In: Diederich J, editor. Rule Extraction from Support Vector Machines; vol. 80 of Studies in Computational Intelligence. Springer Berlin Heidelberg; 2008, p. 33–63.

[24] Hastie T, Tibshirani R. Generalized additive models. Florida, USA: Chapman and Hall; 1990.

[25] Pelckmans K, Goethals I, De Brabanter J, Suykens JAK, De Moor B. Componentwise Least Squares Support Vector Machines; chap. Support Vector Machines: Theory and Applications. Heidelberg, GE: Springer; 2005, p. 77–98.

[26] Stitson M, Gammerman A, Vapnik V, Vovk V, Watkins C, Weston J. Support vector regression with ANOVA decomposition kernels; chap. Advances in kernel methods: support vector learning. Cambridge, MA, USA: MIT Press; 1999, p. 285–91.

[27] Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Statistics for Biology and Health; New York, USA: Springer; 2009.

[28] Harrell F. Regression Modeling Strategies. With applications to linear models, logistic regression, and survival analysis. Springer Series in Statistics; New York, USA: Springer; 2001.

[29] Breiman L. Better subset regression using the nonnegative garrote. Technometrics 1995;37(4):373–84.

[30] Lin Y, Zhang HH. Component selection and smoothing in multivariate nonparametric regression. Annals of Statistics 2006;34(5):2272–97.

[31] Ravikumar P, Lafferty J, Liu H, Wasserman L. Sparse additive models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2009;71(5):1009–30.

[32] Gunn SR, Kandola JS. Structural modelling with sparse kernels. Machine Learning 2002;48:137–63.

[33] Mercer J. Functions of positive and negative type and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society A 1909;209:415–46.

[34] Daemen A, De Moor B. Development of a kernel function for clinical data. In: Proceedings of the 31th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS). IEEE, Piscataway; 2009, p. 5913–7.

[35] Candès EJ, Wakin MB, Boyd S. Enhancing sparsity by reweighted l_1 minimization. Journal of Fourier Analysis and Applications 2008;14(5-6):877–905.

[36] Tibshirani R. The lasso method for variable selection in the cox model. Statistics in Medicine 1997;16(4):267–88.

[37] Yuan M, Lin L. On the non-negative garrote estimator. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2007;69:143–61.

[38] Xavier de Souza S, Suykens JAK, Vandewalle J, Bolle D. Coupled simulated annealing. IEEE Transactions on Systems, Man, and Cybernetics - Part B 2010;40(2):320–35.

[39] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988;44(3):837–45.

[40] Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Smola AJ, Bartlett P, Schoelkopf B, Schuurmans D, editors. Advances in large margin classifiers. Cambridge, MA, USA: MIT Press; 1999, p. 61–74.

[41] Frank A, Asuncion A. UCI machine learning repository. `http://archive.ics.uci.edu/ml`; 2010. (Accessed: 27 March 2013).

[42] Wang X, Wang S. Feature ranking by weighting and ise criterion of nonparametric density estimation. Journal of Applied Sciences 2009;9(6):1014–24.

[43] Dash M, Liu H, Yao J. Dimensionality reduction for unsupervised data. In: In Ninth IEEE International Conference on Tools with Artificial Intelligence. IEEE Computer Society, Piscataway; 1997, p. 532–9.

[44] Kononenko I. Estimating attributes: analysis and extensions of RELIEF. In: Bergadano F, De Raedt L, editors. Proceedings of the European conference on Machine Learning. ECML-94; Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN 3-540-57868-4; 1994, p. 171–82.

[45] Girolami M, He C. Probability density estimation from optimally condensed data samples. IEEE Transactions on Pattern Analysis and Machine Intelligence 2003;25(10):1253–64.

[46] Yacob YM, Mat Sakim HA, Mat Isa NA. Decision tree-based feature ranking using manhattan hierarchical cluster criterion. International Journal of Engineering and Physical Sciences 2012;6(4):187–93.

[47] Zhou XJ, Dillon TS. A statistical-heuristic feature selection criterion for decision tree induction. IEEE Transactions on Pattern Analysis and Machine Intelligence 1991;13(8):834–41.

[48] Hwang YS, Rim HC. Decision tree decomposition-based complex feature selection for text chunking. In: Wang L, Rajapakse JC, Fukushima K, Lee SY, Yao X, editors. Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02.; vol. 5. Piscataway, NJ, USA: IEEE; 2002, p. 2217–22.

[49] Mohammadi M, Gharehpetian GB. Application of core vector machines for on-line voltage security assessment using a decision-tree-based feature selection algorithm. IET Generation Transmission Distribution 2009;3(8):701.

[50] Quinlan JR. C4.5: programs for machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1993.

[51] Karegowda AG, Manjunath AS, Jayaram MA. Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management 2010;2(2):271–7.

[52] Goldberg DE. Genetic Algorithms in Search, Optimization and Machine Learning. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.; 1989.

[53] Balakrishnan S, Narayanaswamy R. Feature selection using FCBF in type II diabetes databases. International Journal of the Computer, the Internet and the Management 2009;17(SP 1):501–8.

[54] Setiono R. Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence in Medicine 2000;18(3):205–19.

[55] Taha I, Ghosh J. Symbolic interpretation of artificial neural networks. IEEE Transactions on Knowledge and Data Engineering 1999;11:448–63.

[56] Eleuteri A, Tagliaferri R, Milano L. A novel information geometric approach to variable selection in mlp networks. Neural Networks 2005;18(10):1309–18.

[57] Setiono R, Liu H. Neurolinear: From neural networks to oblique decision rules. Neurocomputing 1997;17(1):1–24.

[58] van de Laar P. Input selection based on an ensemble. Neurocomputing 2000;34(1-4):227–38.

Table 1: Comparison of the proposed method with standard SVMs using an RBF kernel on the x-or problems in three different settings. For every dataset, only the first two input variables contribute to the class labels.

method	auc (95% CI)	acc (95% CI)	ber (95% CI)
2-dimensional problem			
svm-rbf (all)	1.000 (0.999-1.000)	0.996 (0.972-1.000)	0.004 (0.000-0.024)
svm-rbf (subset)	1.000 (0.999-1.000)	0.996 (0.980-1.000)	0.004 (0.000-0.028)
l_1-svm-rbf$^{\text{tr}}$	1.000 (0.999-1.000)	0.996 (0.980-1.000)	0.004 (0.000-0.018)
4-dimensional problem			
svm-rbf (all)	0.995 (0.989-0.998)	0.956 (0.919-0.972)	0.046 (0.022-0.073)
svm-rbf (subset)	1.000 (0.999-1.000)	0.996 (0.980-1.000)	0.004 (0.000-0.028)
l_1-svm-rbf$^{\text{tr}}$	1.000 (1.000-1.000)	0.996 (0.976-1.000)	0.004 (0.000-0.027)
10-dimensional problem			
svm-rbf (all)	0.947 (0.917-0.966)	0.852 (0.800-0.888)	0.147 (0.109-0.196)
svm-rbf (subset)	1.000 (0.998-1.000)	0.992 (0.972-0.996)	0.007 (0.000-0.027)
l_1-svm-rbf$^{\text{tr}}$	0.997 (0.990-0.999)	0.976 (0.944-0.988)	0.027 (0.007-0.040)

Table 2: Comparison of the test set performance (artificial example 2) of the presented method with a standard SVM using an RBF kernel on the whole set of variables and the selected subset.

method	auc (95% CI)	acc (95% CI)	ber (95% CI)
svm-rbf (all)	0.975 (0.954-0.987)	0.908 (0.860-0.936)	0.084 (0.062-0.135)
svm-rbf (subset)	0.995 (0.984-0.998)	0.956 (0.920-0.976)	0.040 (0.025-0.082)
l_1-svm-rbf$^{\text{tr}}$	0.994 (0.983-0.998)	0.940 (0.900-0.960)	0.054 (0.037-0.099)

Table 3: Comparison of the test set performance in the stability analysis of the presented method with a standard SVM using an RBF kernel on the whole set of variables and the selected subset.

method	auc (std)	acc (std)	ber (std)
svm-rbf (all)	0.954 (0.004)	0.862 (0.009)	0.143 (0.008)
svm-rbf (subset)	0.982 (0.000)	0.944 (0.000)	0.058 (0.000)
l_1-svm-rbf$^{\text{tr}}$	0.975 (0.004)	0.921 (0.014)	0.078 (0.013)

Table 4: Comparison of the test set performance (mean and std) for the Pima Indians Diabetes dataset of the presented method with a standard SVM using an RBF kernel on the whole set of variables and the selected subset. The results illustrate that the presented method (l_1-svm-rnf$^{\text{tr}}$) is competitive with the standard SVM with the additional advantage of being interpretable.

method	auc	acc	ber
svm-rbf (all)	0.826 (0.014)	0.759 (0.019)	0.280 (0.019)
svm-rbf (subset)	0.826 (0.020)	0.762 (0.021)	0.271 (0.026)
l_1-svm-rbf$^{\text{tr}}$	0.826 (0.022)	0.767 (0.020)	0.269 (0.022)

Table 5: Comparison of the feature selection results on the Pima Indian Diabetes problem with results from the literature. For ranking methods, we indicated the set of variables with an equal number of variables as detected by the proposed method (l_1-svm-rbf$^{\text{tr}}$).

method	origin of results	reported accuracy	nb pregnant	plasma glucose	blood pressure	skin fold	serum insulin	body mass index	pedigree function	age
Wang and Wang (2009) [42]	[42]	0.75[a]	X	X						X
sud [43]	[42]	0.68[a]	X				X			X
relief-f [44]	[42]	0.66[a]	X	X						X
k-means [45]	[42]	0.75[a]			X	X				X
mhcc [46]	[46]	NA[b]	X	X		X		X	X	X
Zhou and Dillon (1991) [47]	[46]	NA[b]		X	X	X	X	X	X	X
Hwang and Rim (2002) [48]	[46]	NA[b]	X				X	X		X
Mojammado and Gharehpetian (2009)[49]	[46]	NA[b]	X	X						X
decision tree (C4.5) [50]	[51]	0.86[c]		X	X			X	X	X
genetic algorithm [52]	[51]	0.88[c]		X			X	X		X
fast corr.-based filtering [53]	[53]	0.78[c]		X				X	X	X
l_1-svm-rbf$^{\text{tr}}$		0.77		X				X		X

[a] Numbers are approximately since they were reported in a picture in the original work.

[b] NA: not available. The original work reports an error measure.

[c] The original work does not mention whether this performance is reached on the training or the test set.

Table 6: Comparison of the test set performance (mean and std) on the Wisconsin Breast Cancer dataset of the presented method with a standard SVM using an RBF kernel on the whole set of variables and the selected subset for 10 randomizations of training and test set. The results illustrate that the presented method is competitive with the standard SVM with the additional advantage of being interpretable.

method	auc	acc	ber
svm-rbf (all)	0.996 (0.001)	0.968 (0.008)	0.037 (0.011)
svm-rbf (subset)	0.990 (0.004)	0.954 (0.014)	0.054 (0.018)
l_1-svm-rbf$^{\text{tr}}$	0.990 (0.004)	0.956 (0.010)	0.051 (0.013)

Table 7: Comparison of the feature selection results on the Wisconsin breast cancer dataset with results from the literature.

method	origin of results	reported accuracy	clump thickness	uniformity of cell size	uniformity of cell shape	marginal adhesion	epithelial cell size	bare nuclei	bland chromatin	normal nucleoli	mitoses
osre [22]	[22]	0.95[a]			X	X		X		X	
neurorule (set 1) [54]	[22]	>0.95	X	X		X		X		X	
neurorule (set 2) [54]	[22]	>0.95	X	X				X		X	
bio-re [55]	[22]	0.96	X		X				X	X	X
information geometric [56]	[56]	NA[b]	X		X	X		X			
neurolinear [57]	[56]	0.95			X	X		X		X	X
ensemble based [58]	[56]	NA[b]	X		X	X		X		X	
l_1-svm-rbf$^{\mathrm{tr}}$		0.96			X			X			

[a] Performance of the underlying network.

[b] NA: not available. The original work reports another measure.

Figure captions

Figure 1 Overview of the tuning of the parameters. In all experiments, the data were split into a training and test set. In a first phase the different components were obtained as follows (third column). An SVM with the truncated rbf kernel was used in combination with 10-fold cross validation (10F-CV) to tune the parameters γ and σ, as usual. These parameter values were then used to train the model and obtain the Lagrange parameters α and b. From these, the partial contributions or components $\tilde{y}$ are calculated. In a second phase (fourth column) the relevant components are selected. First, the parameter γ^* from model (5) is tuned by means of 5-fold cross-validation, keeping $c = 1$. Second, the value of c is tuned, keeping γ^* fixed. Third, the model is trained with the tuned parameter values to obtain the coefficients β and bias term b^*. In the final step, the performance of the method is calculated on the test set.

Figure 2 Illustration of the selected effects for the x-or problem. For each setting only one component was selected ($\beta \neq 0$). The figures represent $\beta^{p,q}\tilde{y}^{p,q}$ for the selected interaction effects. No main effects were selected. In each case, it is clearly seen that the selected component has opposite signs in opposite corners of the 2D grid. This corresponds to the x-or setting.

Figure 3 Artificial example 2. Selected main and interaction effects. The figures represent $\beta^p\tilde{y}^p$ and $\beta^{p,q}\tilde{y}^{p,q}$ for the selected components. An increase in the third variable corresponds with an increase in the prognostic index and will target the decision towards the positive class. When x_1 and x_6 have the same sign, an increase in the absolute value one of the values targets the decision towards a positive class. When x_1 and x_6 have opposite signs, an increase in the absolute value of one these values will target the decision towards the negative class. This corresponds with the hyperbolic interaction effect with a positive coefficient of both inputs in the underlying model.

Figure 4 Illustration of the selected features and their effects on the prediction of diabetes in the Pima Indian Diabetes dataset. The gray bars indicate the number of observations in the corresponding variable range. It is seen that an increasing plasma glucose value corresponds to a higher contribution to the prognostic index and will influence the decision towards the positive class (diabetes). An increasing BMI has the same effect, up to 30 after which the effect plateaus. The increase after 40 will have a large variability due to the low number of patients it is based on. An increasing age also contributes to a higher prognostic index and thus targets to decision towards diabetes. After the age of 50, the effect drops.

Figure 5 Illustration of the sparsity performance trade-off by means of the value of c in equation (6). The sparsity is represented by the median number of selected components (gray line). The cross-validation performance is represented by means of the black line. The upper bar indicates the p-value calculated by means of the method

of DeLong. The model with the highest $c-$value (corresponding to the model with the most components) is the reference model. Every model with a smaller $c-$value is compared with this reference. A sparser model obtaining a higher AUC than the current reference model becomes the new reference model (indicated by means of the triangles at the top). A light gray color indicates no significant differences between the AUCs of this and the reference model. A medium gray bar indicates a p-value between 0.01 and 0.05. A dark gray color indicates a p-value less than 0.01. The automated procedure, selecting the sparsest model where the DeLong p-value is larger than 0.05, selects 3 input variables. Inspection of this Figure shows that selecting these variables results in a cross validation performance equal to the one obtained using 19 components.

Figure 6 Decision boundary for the Pima Indian Diabetes problem. Both classes are well separated by the hyperplane. The dots and pluses indicate the observations for healthy patients and patients with diabetes, respectively.

Figure 7 Illustration of the selected features and their effects on the prediction of malignancy in the Wisconsin breast cancer dataset. The gray bars indicate the number of data points with the corresponding value of the input variable. It is seen that an increase in the uniformity of the cell shape or in the bare nuclei level correspond with a higher contribution to the prognostic index and will influence the decision towards the positive class (cancer).

Figure 8 Decision boundary for the Wisconsin breast cancer dataset. Both classes are well separated by the hyperplane. In order to improve the visualization, a random disturbance term is added to the variables, being integers in the dataset. The circles and pluses indicate the observations for malignant and benign tumors, respectively.

fig1.eps
Click here to download Figure: fig1.eps

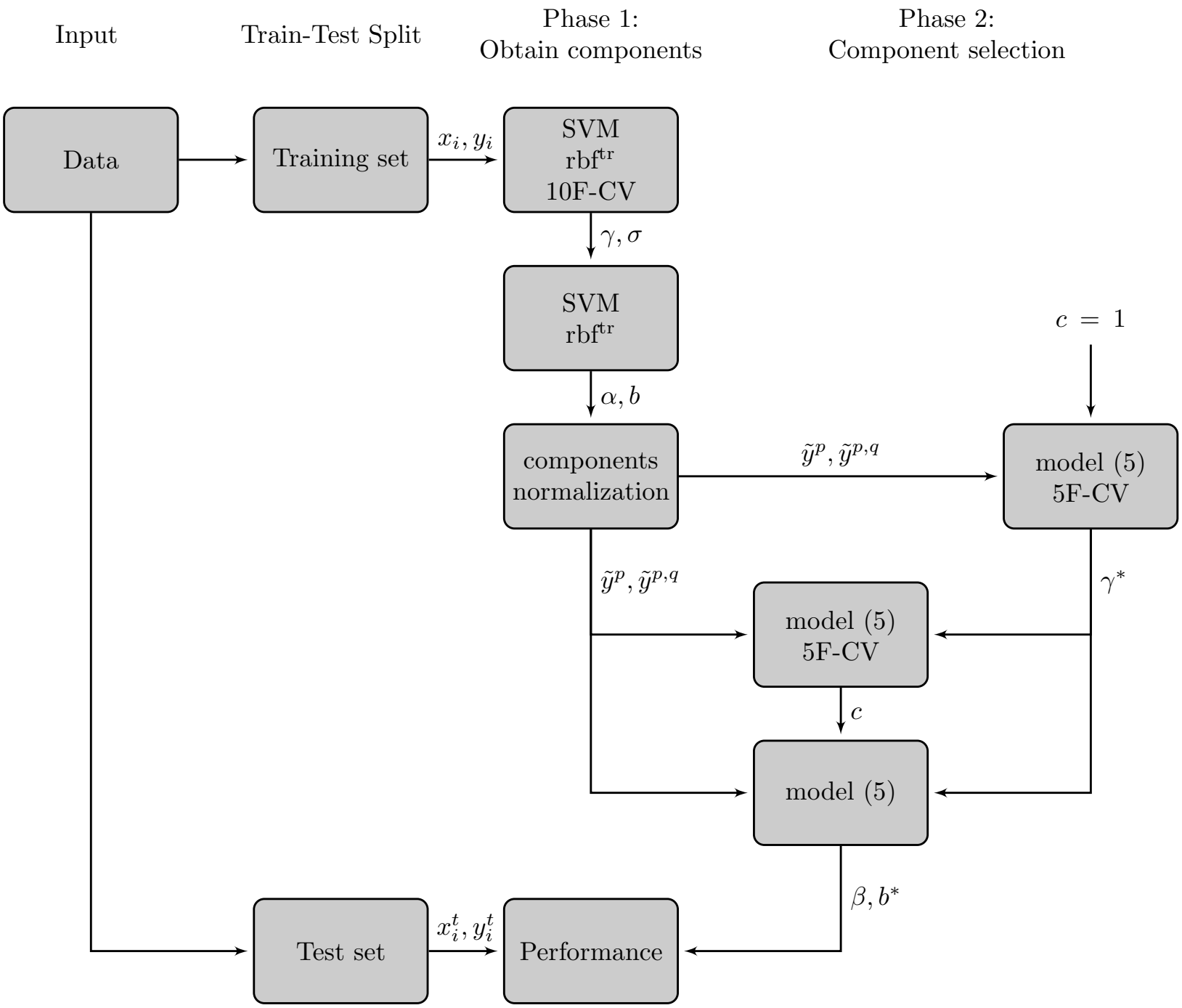
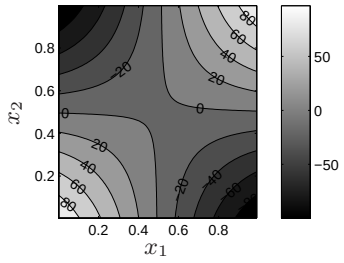
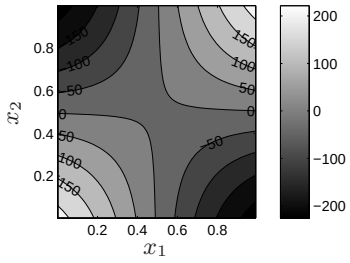


fig2.eps

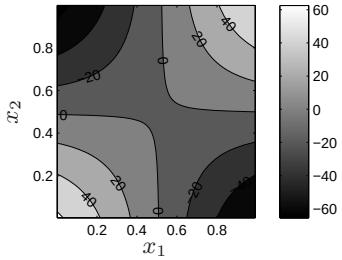
Click here to download Figure: fig2.eps



(a) 2D



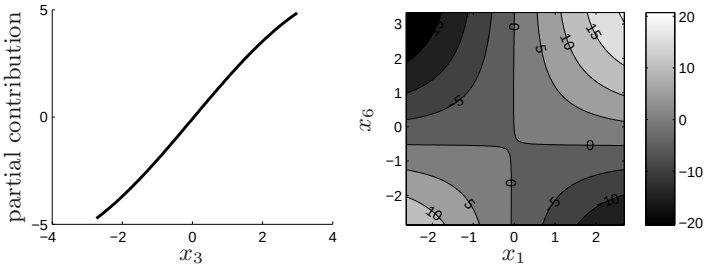
(b) 4D



(c) 10D

fig3.eps

Click here to download Figure: fig3.eps



(a) main effects (b) interaction effects

fig4.eps

Click here to download Figure: fig4.eps

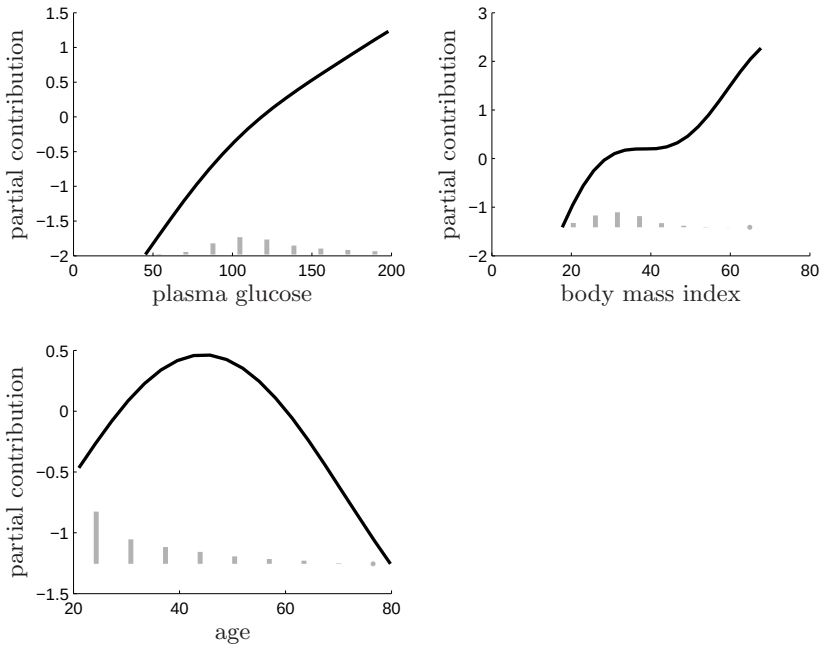


fig5.eps

Click here to download Figure: fig5.eps

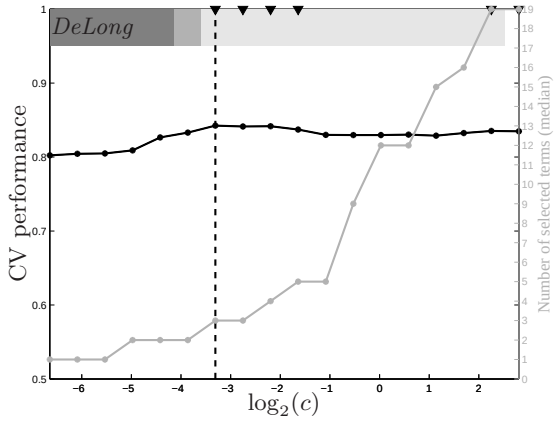


fig6.eps

Click here to download Figure: fig6.eps

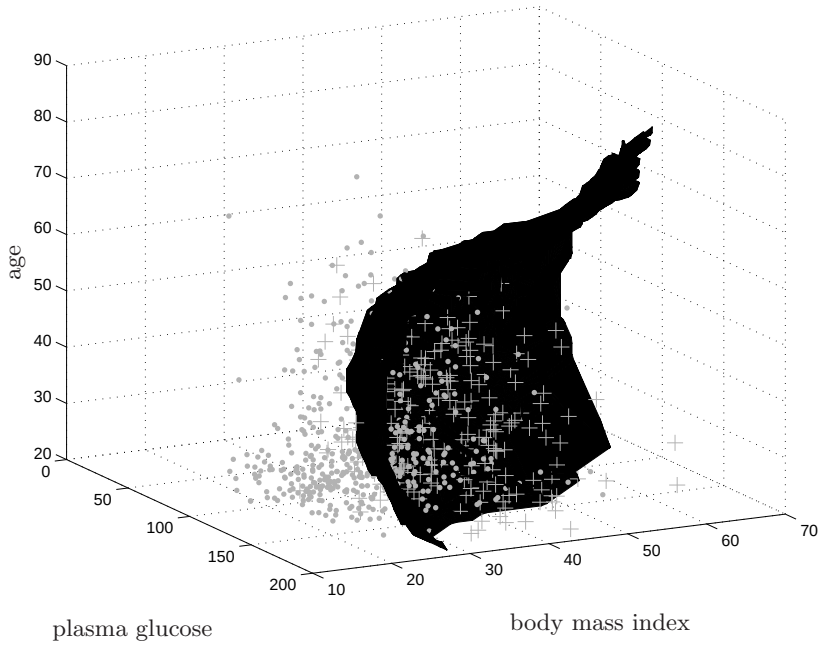


fig7.eps

Click here to download Figure: fig7.eps

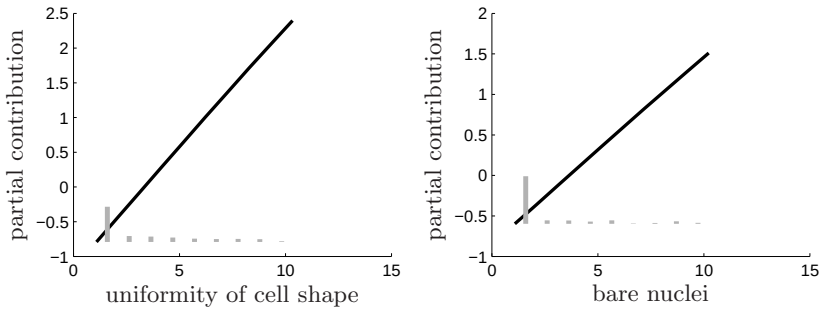


fig8.eps

Click here to download Figure: fig8.eps