Automated Selection of Interaction Effects in Sparse Kernel Methods to Predict Pregnancy Viability

Vanya Van Belle
ESAT-SCD / iMinds-KU Leuven Future Health Department,
KU Leuven, Leuven, Belgium
Department of Mathematics and Statistics,
Liverpool John Moores University, Liverpool, UK
Email: vanya.vanbelle@esat.kuleuven.be

Paulo Lisboa
Department of Mathematics and Statistics,
Liverpool John Moores University, Liverpool, UK
Email: P.J.Lisboa@ljmu.ac.uk

***Abstract*—Support vector machines are highly flexible and generalizing mathematical models that can be used to build prediction models. Their success on a mathematical field is not followed by their application in practice due to their black-box nature. The RBF kernel is often used but the good performance cannot be accompanied by an interpretation of the results. We present a method to visualize the different components of an RBF kernel and propose a method to select the relevant ones. The proposed method is able to automatically detect important main and two-way interaction effects while still obtaining interpretable prediction models. The method is illustrated on a large dataset to predict the viability of pregnancies at the end of the first trimester based on initial scan findings.**

I. Introduction

Support vector machines (SVM) [17] and Least-squares SVMs (LS-SVM) [13], [12] for classification are very flexible and generalizable methods to predict the occurrence of an event of interest (e.g. disease, viability, malignancy). Although these methods are very popular in mathematical and engineering communities, their application in real life problems where interpretation of the results is important is still rare. The main threshold for their application is the black-box nature of the resulting prediction models when using flexible kernels. An alternative is the use of additive or componentwise kernels [9] or anova kernels [11]. The disadvantage of these kernels is that they either assume that interactions are not present, prior knowledge is necessary to decide which interactions should be included or too many interactions are included.

In order for machine learning methods to be used in domains where interpretability of the results is an issue (e.g. in medicine), several issues are important. First it should be possible to perform feature selection in the input space. When building prediction models, it is important to identify which inputs are important and should be collected to make predictions for future observations. Correct selection of input variables will result in a reduction in cost, over and under treatment of patients etc. Secondly, interactions should be allowed for and preferably detected in an automated way. Thirdly, it should be clear what the contribution of each input is and how the final estimate is obtained.

This work is intended to tackle the issues raised above. We start from an SVM for classification with a standard RBF kernel, but restrict the contributions of the RBF kernel to main and two-way interaction effects such that all components can be visualized for interpretation by the end user. Secondly, a sparsity mechanism [14], [4] is used to restrict the number of these components, such that feature selection is automatically performed.

This paper is organized as follows. Section II shortly introduces the standard SVM approach. Section III starts by proposing a truncated version of the RBF kernel such that only main and two-way interaction effects are taken into account. An iteratively reweighted l_1 regularized SVM is then used with the components extracted from the previous defined kernel, to perform component selection. Seciont IV illustrates the proposed method on the prediction of the viability of pregnancies at the end of the first trimester, based on initial scan findings.

II. Support vector machines

Support vector machines for classification were proposed by Vapnik [17] as follows:

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

where $x_i \in \mathbb{R}^d$ represents the input variables of observation i, $y_i \in \{-1, 1\}$ represents the corresponding class label, $\varphi(\cdot)$ is a feature map mapping the inputs to a higher dimensional space and γ is a positive regularization constant. The idea behind the SVM is that by mapping the input space to a higher dimensional input space, non-linear separable problems can be separated using a linear hyperplane in the higher dimensional feature space. The model is most often used in its dual form:

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

where $\alpha_i,\ i = 1, \ldots, N$ are the Lagrange multipliers. Thanks to the kernel trick, the mapping can be done implicitly when choosing a positive semidefinite kernel

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

The estimated class label is then found as

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

where $\sum_{i=1}^{N} \alpha_i y_i K(x_i, x_\star) + b$ is often referred to as the latent variable.

III. Component selection in the RBF kernel

This Section starts with the introduction of an alternative to the RBF kernel, restricting the contributions to components that are visualizable, to ensure that the resulting models can be interpreted by experts in the application domain. We then propose to use a sparsity mechanism to reduce the number of components.

A. Truncated RBF kernel

The RBF kernel

$$K_{\mathrm{RBF}}(x, z) = \exp\left(-\frac{||x - z||_2^2}{\sigma^2}\right), \tag{3}$$

with σ a tuning parameter, is often used in combination with SVMs because it can handle non-linearities, interactions and is bounded. However, the choice for this kernel has some implications w.r.t. the application domain. The obtained classifier is completely black-box and the user is only provided with the estimated class. An explanation of how the method reaches this particular estimate is not available.

The literature describes several alternative kernels to allow for interpretable results, including additive kernels such as the polynomial kernel, componentwise kernels [9], and the clinical kernel [5]. The advantage of these kernels is found in their additive structure, such that the model can be represented by means of different terms, which add up to the latent variable. For many practical problems however, these approaches lack an important aspect of model building: the ability to include interaction effects. The RBF kernel automatically uses all variables and all possible interactions between them. Additive kernels are restricted to main effects and are not able to detect important interactions. An alternative is therefore the use of anova kernels [11], where interaction terms are manually added. However, the use of anova kernels implies the availability of prior knowledge such that relevant interactions can be included. Alternatively, a large number or all interaction effects are used, reducing the interpretability of the results.

In application domains where interpretation of the results is an issue, the users are mainly interested in main effects and interaction effects between two input variables. Higher order interaction effects do not occur often in these domains and/or become very difficult to interpret. To accomplish the requirements of the end user without restricting the mathematical methods to non-flexible methods, we propose the truncated RBF kernel as

$$K^{\mathrm{tr}}_{\mathrm{RBF}}(x, z) = \frac{2}{d(d-1)} \sum_{p=1}^{d} \sum_{q>p}^{d} K_{\mathrm{RBF}}(x^{p,q}, z^{p,q})\,, \tag{4}$$

a linear combination of RBF kernels evaluated on all pairs of input variables. Based on the kernel properties discussed in [7], it can be shown that this new kernel is still positive definite and thus a valid kernel to apply the kernel trick. The above kernel includes all possible main and two-way interaction effects that are also present in the standard RBF kernel. The difference between this and the original RBF kernel lies completely in the truncation of the Taylor expansion of the RBF kernel after the two-way interactions. When using the truncated RBF kernel in combination with the standard SVM classifier, the class estimate is found as

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

where the partial contribution $\hat{y}^{p,q}$ are a weighted sum of RBF kernels evaluated on two-dimensional vectors.

B. Component selection

When the goal is not only the visualize how the estimate is obtained, but features need to be selected, the partial components can be separated into main and interaction effects and used as input variables in an iteratively reweighted l_1 regularized SVM [2]. The partial contribution $\hat{y}_{p,q}$ contains the main effects of x^p ($\hat{y}^p$) and x^q ($\hat{y}^q$) and the interaction between both ($\hat{y}^{p,q} - \hat{y}^p - \hat{y}^q$). All these terms are then normalized to zero mean and unit standard deviation ($\tilde{y}^p$ and $\tilde{y}^{p,q}$) and used as inputs for the following method

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

subject to

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

with χ^p equal to 1 in the first iteration and defined as

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

in later interations. Here, ε is a small constant (e.g. 0.005) and c is a parameter making the trade-off between sparsity and performace. This parameter needs to be tuned for the problem at hand. We use the 1-norm as in [15] but further improve the sparsity by iteratively reweighting as in [4]. As in the non-negative garrote estimator [3], [19], the coefficients β are restricted to be positive.

IV. Prediction of pregnancy viability

The proposed method is applied for the prediction of the viability of pregnancies at the end of the first trimester. A dataset with infomation on 1435 single pregnancies is available. All patients attended a single early pregnancy unit after a positive pregnancy test. Demographic data and presence or absence of symptoms were recorded and a sonography was performed. The patients were followed and the nuchal scan (at 11-14 weeks of gestation) was used as ground truth. Out of 1435 pregnancies, 885 were viable at the nuchal scan. A selected subset of variables based on correlations and availability of the inputs in clinical practice, was selected for this analysis. Missing variables were imputed [10]. The input variables are: maternal age, pbac bleeding score (0=no bleeding, 1=light bleeding, ..., 4=bleeding with cloths), vas (visual analog score) pain score, previous delivery after 24 weeks (binary), previous early miscarriage (binary), mean gestational sac diameter, mean yolk sac diameter and presence of a fetal heart rate. The data was split into a training ($n_{\text{train}} = 955$) and test set ($n_{\text{test}} = 480$).

A. Model building

The training set was 10 times split into a training (two thirds of the data) and validation set (one third). In each randomization, the model was build on the training set and evaluated on the validation set. Model hyperparameters γ (equation (1)) and σ (equations (3) and (4)) were tuned using 10-fold cross validation and coupled simulated annealing [18]. The parameters are 10 times randomly initialized, where σ is scaled according to $\sqrt{d}$, with d the dimension of the input space. The area under the receiver operating characteristic (AUC) [8] is used as model selection criterion. The parameter γ^* in equation (5) was tuned by means of a linesearch in the range $[0.01 - 1000]$, where c was fixed at $c = 1$. The value leading to the highest 10-fold cross-validation AUC was selected. To find an optimal value of c in equation (6), the model with fixed γ^* was evaluated for c in $[0.01, 1]$. When more than 1 component was selected for $c = 0.01$, the value of γ^* was decreased. When less than $3d$ components were selected, the maximal value of c to be evaluated was increased. Within each training-validation split, two thirds (random with equal prevalence) of the training data was used to train the model, the remainder of the data was used to calculate the performance and tune c. The AUC of the model with most components (corresponding to the highest value of c) was used as baseline model and the performance of increasingly sparse models (when lowering the value of c) is compared with this model using the method described in [6]. The lowest value of c for which the p-value obtained by this model is larger than 0.05 is selected as the optimal value. This approach is able to select a sparser model without statistical evidence of a reduced AUC. The baseline model was updated whenever the sparser model obtained a higher AUC.

B. Results

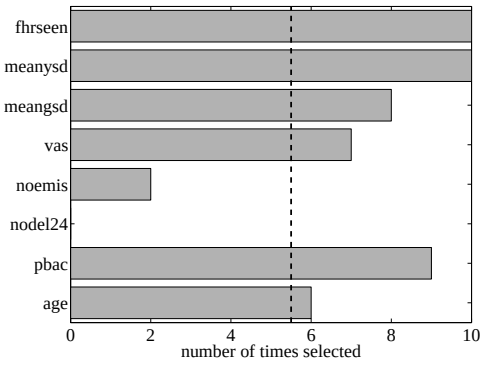
The results for the 10 randomizations are shown in Table I. The mean area under the ROC curve (AUC), accuracy (ACC) and balanced error rate (BER) are reported together with their standard deviation. The proposed method performs as well as the standard RBF kernel, but results in sparser models.

TABLE I
MEAN AND STANDARD DEVIATION OF THE PERFORMANCE ON THE VALIDATION SETS OVER 10 RANDOMIZATIONS. THE PROPOSED METHOD IS COMPARED WITH THE STANDARD RBF KERNEL AND THE RBF KERNEL USING THE SUBSET OF VARIABLES SELECTED BY THE PROPOSED METHOD. AUC: AREA UNDER THE RECEIVER OPERATING CURVE; ACC: ACCURACY; BER: BALANCED ERROR RATE.

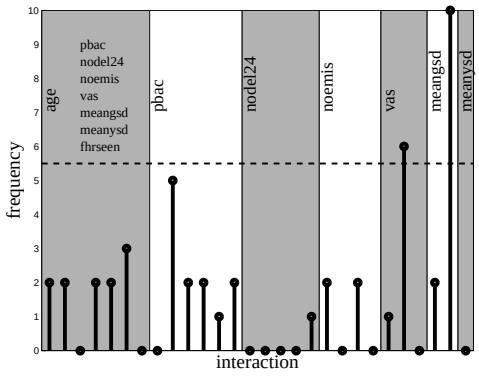
method	AUC	ACC	BER
RBF	0.895	0.813	0.200
	(0.012)	(0.010)	(0.012)
RBF (subset)	0.898	0.816	0.196
	(0.011)	(0.015)	(0.020)
proposed method	0.899	0.813	0.197
	(0.012)	(0.016)	(0.019)

Figure 1 illustrates the frequency at which components are selected. Part (a) of the Figure represents the selection of main effects, part (b) the selection of interaction effects. The first gray area indicates interactions with age. The order of the variables is indicated within this gray area. The next white area represents all interactions with pbac bleeding score (but not the one with age), etc. The main effects of age, pbac bleeding score, vas pain score, mean gestational sac, main yolk sac and the presence of a fetal heart rate are selected in most of the cases. Figure 1 (b) clearly indicates the necessity of the inclusion of an interaction effect between gestational sac size and the presence of the fetal heart rate. Additionally, the interaction between the pain score and the diameter of the yolk sac seams important.

To stabilize the results, a final model is trained using the complete training set. Only those components that are selected more than 5 times in the 10 randomization are included. Figure 2 illustrates the estimated effects of the selected components. Given the other inputs, a low age (< 20) and a high age (> 35) decrease the chance on a viable pregnancy. The absence of bleeding is a good sign, whereas heavier bleeding increases the risk on an unviable pregnancy, up to the level of 2. Bleeding more heavily no longer influences the results. Interpretation of the other main effects can only be done when taking the interaction effects into account as well. Figure 3 visualizes the combined main and interaction effects. The main and interaction effects of the pain score seem to counteract as the combined effects of pain score and mean yolk sac diameter seem to indicate a main effect of yolk sac diameter instead of an interaction effect. The interaction is only relevant for a small number of points (dots in Figure 3(a)). The effect of



(a)



(b)

Fig. 1. Frequency at which components are selected: (a) main effect; (b) interaction effects. The main effects of age, bleeding and pain score, gestational and yolk sac diameter and the presence of a fetal heart rate are selected in most randomizations. The interaction effect between the gestational sac size and the presence of a fetal heart rate is always selected, indicating the importance of this interaction effect. The interaction between the vas pain score and the yolk sac diameter is also selected in most randomizations.

mean gestational sac size depends on whether or not a fetal heart rate was visualized. In the presence of a heart rate, a higher gestational sac is better. In the absence of a heart rate, the prognosis can still be positive whenever the gestational sac is small. This indicates that the pregnancy is still in a very early stage, where a fetal heart rate can not be detected yet. Absence of a heart rate in combination with a large gestational sac indicates problems with the fetus and results in a poor prognosis.

Table II summarizes the performance of this final model on the test set. Using only 8 components (and 6 variables) the proposed method is able to perform as well as the standard RBF kernel. Restricting the variables used in the RBF kernel to the ones selected by the proposed method, maintains the performance. This is a clear indication that the proposed method is able to select the relevant inputs.

The results are w.r.t. selected components and performance comparable to the results reported in [1], [16], where the same data where used. However, the inclusion of the interaction effect was non-automated in the first and based on prior knowledge in the second paper. The approach presented here, offers a way to combine feature selection in input space

TABLE II
PERFORMANCE AND 95% CONFIDENCE INTERVALS FOR THE INDEPENDENT TEST SET. THE PROPOSED METHOD IS COMPARED WITH THE STANDARD RBF KERNEL AND THE RBF KERNEL USING THE SUBSET OF VARIABLES SELECTED BY THE PROPOSED METHOD. AUC: AREA UNDER THE RECEIVER OPERATING CURVE; ACC: ACCURACY; BER: BALANCED ERROR RATE.

method	AUC	ACC	BER
RBF	0.927	0.846	0.162
	0.903-0.946	0.808-0.873	0.130-0.200
RBF (subset)	0.933	0.860	0.146
	0.907-0.952	0.827-0.885	0.118-0.184
proposed method	0.918	0.838	0.154
	0.893-0.942	0.800-0.867	0.140-0.204

with automated relevance determination of main and two-way interaction effects.

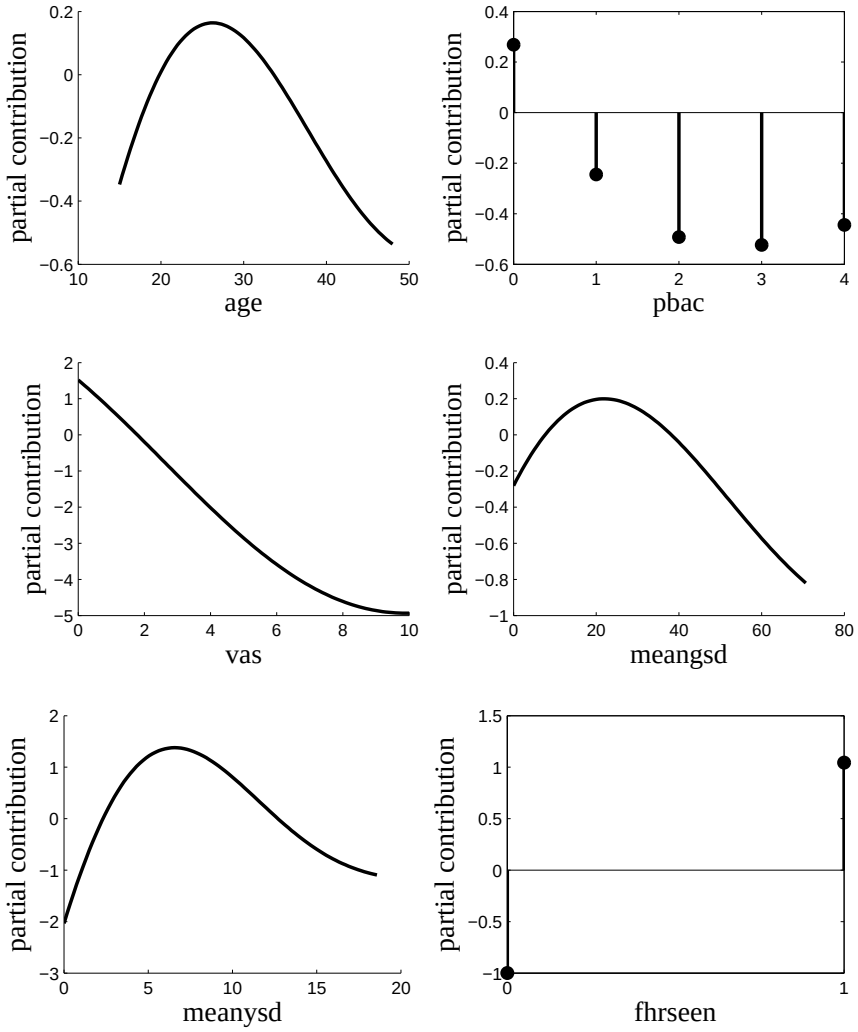
Although the results of the proposed method are not unique due to the additive structure of the model, the results are of major importance in domains with interpretability constraints. Thanks to the nice visualization of the selected components, made possible thanks to the use of the truncated kernel, the end user can interpret the results. This visualization will enable the discussion between model developer and end user w.r.t. the inclusion of relevant components and possibly result in better models. Additionally, since this process encourages the end user to be involved in the model building, the end users will become more engaged to use the developed model.

V. CONCLUSION

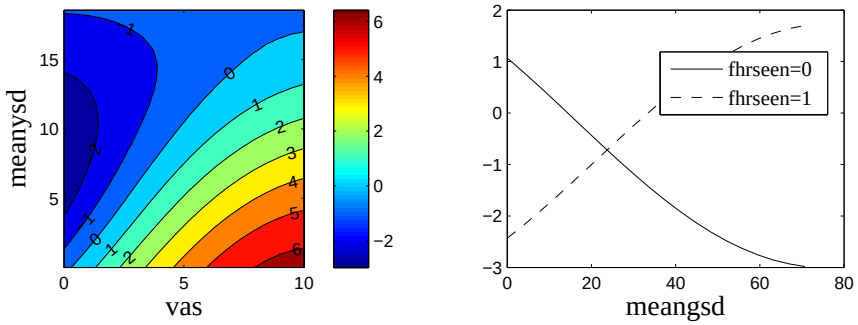
This paper presents a new kernel to ensure that the results of the SVM classifier can be visualized and interpreted by experts in the application domain. The different components, that add up to the latent variable, are extracted from this kernel and used as input variables for an iteratively reweighted l_1 regulatized SVM classifier in order to automatically select relevant components. As such, the proposed method is able to perform feature selection in the input space, visualize the relevant contributions and indicate how the model estimate is obtained. The method is illustrated on a large dataset to predict the viability of pregnancies at the end of the first trimester. The performance is comparable with that of the standard RBF kernel. The method offers to additional advantage that the end user can interpret the results. It additionally enables the discussion between model builders and end users, which is necessary to obtain models that will be used in practice.

ACKNOWLEDGMENT

We thank Dr. C. Bottomley for the collection of the data and for allowing us to use it. This research is 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



(a)



(b)

Fig. 2. Selected components: (a) main effect; (b) interaction effects. A low and high age and higher bleeding score correspond to a higher risk on a unviable pregnancy. To interpret the estimated effects for variables that are involved in main and interaction effects, both effects should be combined as in Figure 3.

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] C. Bottomley, V. Van Belle, E. Kirk, S. Van Huffel, D. Timmerman, and T. Bourne. Accurate prediction of pregnancy viability by means of a simple scoring system. *Human Reproduction*, 2012 (First published online: October 30, 2012).

[2] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proceedings of the*

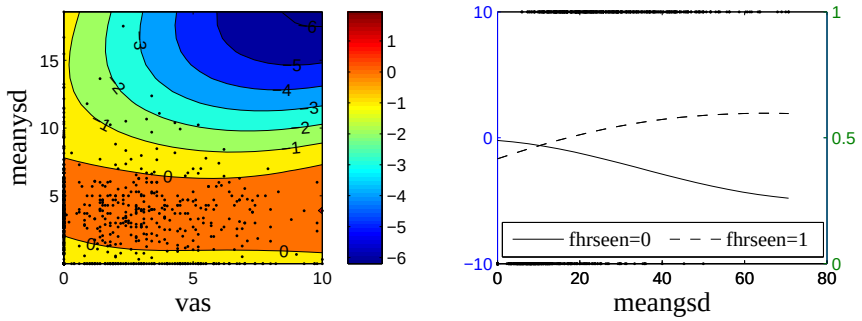


Fig. 3. Alternative representation of the effects of variables with main and interaction effects. The dots indicate the position of the training observations. (a) The combined effect of pain score and mean yolk sac diameter mainly illustrates the main effect of mean yolk sac diameter as presented in Figure 2. It should be discussed with the end user whether the effect of the pain score seems relevant in practice or could be eliminated. (b) The effect of mean gestational sac size depends on whether or not a fetal heart rate is visualized. When a heart rate is visualized, a larger gestational sac is better, since both observations then indicate that the pregnancy is of later gestation. When a fetal heart rate is absent, the outcome can still be positive for very early pregnancies, i.e. small gestational sacs. A larger gestational sac indicates problems with the fetus and corresponds to a poor prognosis.

Fifteenth International Conference on Machine Learning, ICML, pages 82–90, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[3] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[4] E. J. Candès, M. B. Wakin, and S. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

[5] A. Daemen and B. De Moor. 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)*, pages 5913–5917. IEEE, Piscataway, 2009.

[6] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[7] M. Genton, N. Cristianini, J. Shawe-taylor, and R. Williamson. Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

[8] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, April 1982.

[9] K. Pelckmans, J. A. K. Suykens, and B. De Moor. Componentwise support vector machines for structure detection. In *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005*. Springer Berlin / Heidelberg, 2005.

[10] J. Siddique and O. Harel. MIDAS: A sas macro for multiple imputation using distance-aided selection of donors. *Journal of Statistical Software*, 29(9):1–18, 2009.

[11] M. Stitson, A. Gammerman, V. Vapnik, V. Vovk, C. Watkins, and J. Weston. *Support vector regression with* ANOVA *decomposition kernels*, chapter Advances in kernel methods: support vector learning, pages 285–291. MIT Press, Cambridge, MA, USA, 1999.

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

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

[14] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, series B*, 58(1):267–288, 1996.

[15] R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):267–288, 1997.

[16] V. Van Belle, B. Van Calster, D. Timmerman, T. Bourne, C. Bottomley, L. Valentin, P. Neven, S. Van Huffel, J. A. K. Suykens, and S. Boyd. A mathematical model for interpretable clinical decision support with applications in gynecology. *PLoS One*, 7(3):e34312, 2012.

[17] V. Vapnik. *Statistical Learning Theory.* Wiley and Sons, New York, 1998.

[18] S. Xavier de Souza, J. A. K. Suykens, J. Vandewalle, and D. Bolle. Coupled simulated annealing. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 40(2):320–335, 2010.

[19] M. Yuan and L. Lin. On the non-negative garrote estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:143–161, 2007.