Computational Statistics and Data Analysis 56 (2012) 611–628

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A mixed effects least squares support vector machine model for classification of longitudinal data

Jan Luts [a,b,*], Geert Molenberghs [c,d], Geert Verbeke [d], Sabine Van Huffel [a,b], Johan A.K. Suykens [a,b]

[a] Department of Electrical Engineering (ESAT), Research Division SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

[b] IBBT-K.U. Leuven Future Health Department, Leuven, Belgium

[c] I-BioStat, Universiteit Hasselt, Agoralaan 1, B-3590 Diepenbeek, Belgium

[d] I-BioStat, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B-3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 12 November 2010
Received in revised form 9 September 2011
Accepted 10 September 2011
Available online 18 September 2011

Keywords:
Classification
Longitudinal data
Least squares
Support vector machine
Kernel method
Mixed model

ABSTRACT

A mixed effects least squares support vector machine (LS-SVM) classifier is introduced to extend the standard LS-SVM classifier for handling longitudinal data. The mixed effects LS-SVM model contains a random intercept and allows to classify highly unbalanced data, in the sense that there is an unequal number of observations for each case at non-fixed time points. The methodology consists of a regression modeling and a classification step based on the obtained regression estimates. Regression and classification of new cases are performed in a straightforward manner by solving a linear system. It is demonstrated that the methodology can be generalized to deal with multi-class problems and can be extended to incorporate multiple random effects. The technique is illustrated on simulated data sets and real-life problems concerning human growth.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

During the past decade, kernel-based approaches for estimation and learning problems have gained significant interest (Schölkopf and Smola, 2001). Kernel methods are typically used in the machine learning community to deal with all kinds of classification problems involving input vectors with a fixed dimension. For instance, the popular kernel-based technique support vector machines (SVMs) has successfully been used for many applications (Vapnik, 1998). On the other hand, for many real-life problems the data are characterized by their non-fixed dimensional nature. For example, the analysis of longitudinal data, which consist of repeated measurements on the same units (or often called cases) over a period of time, generally requires the use of methods which can deal with unbalanced data. Within the context of longitudinal data, unbalanced means that there is an unequal number of observations for each case at non-fixed time points. While SVMs are not straightforwardly applicable to such kind of data, the classical approach in the statistics community is to analyze longitudinal data by means of mixed effects models (Verbeke and Molenberghs, 2000). Importantly, these models are able to take the relationship between observations from the same case into account. Estimation of the fixed and random effects is commonly based on the theory of best linear unbiased prediction (BLUP).

* Corresponding author at: Department of Electrical Engineering (ESAT), Research Division SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. Tel.: +32 16 321065; fax: +32 16 321970.
E-mail address: jan.luts@esat.kuleuven.be (J. Luts).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.09.008

To classify longitudinal profiles in different groups, Verbeke and Lesaffre (1996) proposed a linear mixed effects model with heterogeneity in the random effects. The normality assumption for the random effects was replaced by a mixture of normal distributions with a common covariance matrix. Maximum likelihood estimates were obtained based on the expectation-maximization algorithm. A multi-class classification rule was established by assigning a case to the component to which it had the highest posterior probability to belong. Marshall and Barón (2000) designed a linear/quadratic discriminant analysis method for unbalanced longitudinal data using mixed effects models. Linear models as well as models that were nonlinear with respect to the mean and the random effects were included. An iterative algorithm was proposed to estimate the class means and covariance matrices, which were subsequently used in various discriminant functions. Variance components were allowed to vary between classes. The final classification was performed based on an extended Mahalanobis distance measure. James and Hastie (2001) extended linear discriminant analysis to a functional linear discriminant analysis (FLDA) classifier for irregularly sampled curves. An expectation-maximization algorithm was used to fit a mixed model with splines as random and fixed effects. Final class assignment was based on posterior probabilities. Several extensions such as functional quadratic discriminant analysis (i.e. class-specific covariance matrices) and functional regularized discriminant analysis were proposed. Wouters et al. (2007) presented a two-step approach called doubly hierarchical discriminant analysis to classify high-dimensional longitudinal data. This approach first performs modeling of the data via a class of flexible linear mixed models. The resulting set of estimated parameters, which represented a summary of the data, was then fed to a hierarchical linear discriminant classifier. A decision tree composed of binary classifiers was constructed and posterior probabilities were computed for class assignment. Fieuws et al. (2008) focused on the classification of multivariate longitudinal profiles. Univariate linear, nonlinear and generalized mixed models were combined into a multivariate mixed model in order to take the correlation between the different profiles into account. Association between the different profiles was assumed to be fully captured by the association between their random effects. Parameter estimates were obtained based on combining all possible bivariate mixed models. Posterior probabilities were calculated for class assignment. Other studies, e.g. Brown et al. (2001) and De la Cruz-Mesía et al. (2007), addressed the problem of classification of longitudinal profiles from a Bayesian point of view. The former described a multivariate model to classify longitudinal data, that were only measured at fixed time points, and used an expectation-conditional maximization algorithm. The latter developed a semi-parametric approach for unbalanced longitudinal data and used Markov chain Monte Carlo methods. In addition, the topic of longitudinal data classification is also related to the field of functional data analysis (González-Manteiga and Vieu, 2007; Ramsay and Silverman, 2005). Functional data can be interpreted as an observation over a certain continuum, e.g. time. Even though these observations are discrete, they can be perceived as values that vary smoothly over time, and therefore they are called functional observations. Within this context, James (2002) developed a technique to incorporate (multiple) functional predictors in generalized linear models. An expectation-maximization algorithm was described to fit the spline-based functional logistic regression model. Finally, Müller (2005) showed an example of the use of a functional binomial regression model with a logistic link for classification of sparse longitudinal data.

Various approaches have been used to deal with nonlinear profiles in mixed effects models. For instance, the FLDA approach by James and Hastie (2001) included natural cubic spline functions to model the observations of each case. Wouters et al. (2007) employed a fractional polynomial mixed model to extract features as input for discriminant analysis. Liu et al. (2007) studied the use of kernels for regression analysis by means of linear mixed models, but longitudinal data were not considered. More recently, explicit connections between the traditional approaches for longitudinal data fitting and kernel machines were described (Pearce and Wand, 2009). The correspondence between reproducing kernel Hilbert space optimization and estimated BLUPs was illustrated. Whereas the study by Pearce and Wand (2009) showed that kernel methods can be used for regression modeling of longitudinal data, the present study focuses on classification of this type of data using a kernel-based approach. To this end, the least squares support vector machine (LS-SVM) classifier (Suykens et al., 2002; Suykens and Vandewalle, 1999) is extended to a mixed effects LS-SVM model for classification of longitudinal profiles.

The proposed method consists of a regression modeling and a consecutive prediction step (Section 2). This kernel method combines the regularization property of support vector machines with mixed effects models. The methodology is able to deal with nonlinearity in the data and it addresses binary, as well as multi-class classification (Section 3.1). A model formulation with multiple random effects and corresponding classification strategy are developed (Section 3.2). Model selection via parameter tuning is discussed in Section 4. The techniques are applied to various simulated and real-life problems (Section 5).

2. Random intercept LS-SVM

This section starts with a brief review of the standard LS-SVM classifier (Section 2.1). In the next sections, an LS-SVM model is developed for longitudinal data classification. The random intercept LS-SVM approach includes two steps. Firstly, a regression is done using training data for which the true class labels are known in order to estimate the model parameters (Section 2.2). Secondly, new longitudinal profiles with unknown class labels can be classified based on the obtained model parameters from the previous step (Section 2.3). Finally, an alternative approach to this classification step is established (Section 2.4).

2.1. LS-SVM classifier

The standard LS-SVM classifier is a binary classifier and requires a data set of N input vectors and corresponding class labels: $\{x_i, z_i\}_{i=1}^N$, with $x_i \in \mathbb{R}^d$ being the input pattern and $z_i \in \{-1, +1\}$ the class indicator. The proposed optimization problem is defined as

$$\min_{w,e,b} \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2, \tag{1}$$

subject to

$$z_i(w^T \varphi(x_i) + b) = 1 - e_i, \quad i = 1, \ldots, N, \tag{2}$$

with $e = [e_1 \ \ldots \ e_N]^T, \varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d_h}$ a mapping from the input space to a high-dimensional feature space with dimension d_h, w a vector of the same dimension as $\varphi(\cdot)$, γ a positive regularization constant and b a bias term. In contrast with SVMs, LS-SVMs employ a least squares loss function with error variables e_i and equality constraints are used in place of inequality constraints. This formulation results in solving a set of linear equations, while SVMs require solving a quadratic programming problem. Whereas the primal problem in (1) is expressed in terms of the feature map, the resulting linear system in the dual space is expressed in terms of the kernel function. The method requires tuning of the regularization constant γ and kernel parameters. The resulting LS-SVM classifier can be used to classify a test data set $\{x_j^*, z_j^*\}_{j=1}^G$. The classifier is formulated as

$$\hat{z_j^*} = \operatorname{sign}\left(\sum_{i=1}^{N} \hat{\beta}_i z_i K(x_j^*, x_i) + \hat{b} \right), \tag{3}$$

in the dual space with x_j^* the jth test case to be classified, Lagrange multipliers β_i and the positive definite kernel function $K(\cdot, \cdot)$ with $K(x_j^*, x_i) = \varphi(x_j^*)^T \varphi(x_i)$. In practice, a linear, polynomial or radial basis function (RBF) kernel is often used. In the remainder of this study the RBF kernel (i.e. $K(x_j^*, x_i) = \exp(- \parallel x_j^* - x_i \parallel_2^2 / \sigma^2)$) is employed.

The LS-SVM has been shown to be a core model, leading to a generic framework that allows to define new methods in a systematic and constructive way (De Brabanter et al., 2010, 2011; Luts et al., 2010; Suykens et al., 2010). Examples include weighted versions to incorporate robustness, sparse representations and a fixed-size approach for large-scale data. The models provide connections to parametric statistics, as well as non-parametric statistics, via primal and dual formulations. The LS-SVM, having minimal underlying assumptions, is well-suited to deal with high-dimensional input data. In addition, reliable numerical implementations are available (e.g. a linear system for classification and regression).

2.2. Training: random intercept LS-SVM regression

The longitudinal training data can be represented as y_{ij} which is the observed value at time x_{ij} for case i, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, with N the number of cases, n_i the number of observations for case i and the total number of observations equals $M = \sum_{i=1}^N n_i$. The class labels are known and are denoted by $z_i \in \{-1, +1\}$ for case i. The training step consists of a regression modeling. Based on the theory of BLUP (Robinson, 1991), the proposed optimization problem for random intercept LS-SVM regression is formulated as

$$(\hat{w}, \hat{e}, \hat{b}, \hat{B_0}, \hat{B_1}) = \operatorname*{argmin}_{w,e,b,B_0,B_1} \frac{1}{2} w^T w + \frac{\lambda_1}{2} \sum_{i=1}^{N} b_i^2 + \frac{\lambda_2}{2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} e_{ij}^2, \tag{4}$$

subject to the model

$$y_{ij} = B_0 + B_1 z_i + b_i + \left(\frac{z_i - 1}{2} \right) w_1^T \varphi_1(x_{ij}) + \left(\frac{z_i + 1}{2} \right) w_2^T \varphi_2(x_{ij}) + e_{ij}, \quad i = 1, \ldots, N, \ j = 1, \ldots, n_i, \tag{5}$$

with weighting vector $w = [w_1^T w_2^T]^T$, a vector of random intercepts $b = [b_1 \ldots b_N]^T$, error variables $e = [e_{11} \ \ldots \ e_{1n_1} e_{21} \ldots e_{Nn_N}]^T \in \mathbb{R}^M$, positive regularization constants λ_1 and λ_2, model parameters B_0 and B_1, and mappings $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$. The Lagrangian for the problem in (4)–(5) is

$$\begin{aligned} \mathcal{L}_1(w, e, b, B_0, B_1, \alpha) = & \frac{1}{2} w^T w + \frac{\lambda_1}{2} \sum_{i=1}^{N} b_i^2 + \frac{\lambda_2}{2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} e_{ij}^2 \\ & - \sum_{i=1}^{N} \sum_{j=1}^{n_i} \alpha_{ij} \left(y_{ij} - B_0 - B_1 z_i - b_i - \left(\frac{z_i - 1}{2} \right) w_1^T \varphi_1(x_{ij}) - \left(\frac{z_i + 1}{2} \right) w_2^T \varphi_2(x_{ij}) - e_{ij} \right), \end{aligned} \tag{6}$$

with Lagrange multipliers $\alpha = [\alpha_{11} \ \ldots \ \alpha_{1n_1}\alpha_{21} \ \ldots \ \alpha_{Nn_N}]^T \in \mathbb{R}^M$. The conditions for optimality yield

$$\begin{cases} \dfrac{\partial \mathcal{L}_1}{\partial w_1} = 0 \rightarrow w_1 = -\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{n_i} \alpha_{ij}\left(\frac{z_i - 1}{2}\right)\varphi_1(x_{ij}), \\ \dfrac{\partial \mathcal{L}_1}{\partial w_2} = 0 \rightarrow w_2 = -\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{n_i} \alpha_{ij}\left(\frac{z_i + 1}{2}\right)\varphi_2(x_{ij}), \\ \dfrac{\partial \mathcal{L}_1}{\partial e_{ij}} = 0 \rightarrow e_{ij} = -\dfrac{\alpha_{ij}}{\lambda_2}, \quad i = 1, \ldots, N, j = 1, \ldots, n_i, \\ \dfrac{\partial \mathcal{L}_1}{\partial b_i} = 0 \rightarrow b_i = -\dfrac{1}{\lambda_1}\displaystyle\sum_{j=1}^{n_i}\alpha_{ij}, \quad i = 1, \ldots, N, \\ \dfrac{\partial \mathcal{L}_1}{\partial \alpha_{ij}} = 0 \rightarrow y_{ij} = B_0 + B_1 z_i + b_i + \left(\dfrac{z_i - 1}{2}\right) w_1^T\varphi_1(x_{ij}) + \left(\dfrac{z_i + 1}{2}\right) w_2^T\varphi_2(x_{ij}) + e_{ij}, \\ \quad i = 1, \ldots, N, j = 1, \ldots, n_i, \\ \dfrac{\partial \mathcal{L}_1}{\partial B_0} = 0 \rightarrow 0 = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{n_i}\alpha_{ij}, \\ \dfrac{\partial \mathcal{L}_1}{\partial B_1} = 0 \rightarrow 0 = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{n_i}\alpha_{ij} z_i. \end{cases} \tag{7}$$

After eliminating w_1, w_2, e_{ij} and b_i one obtains

$$\begin{cases} y_{ij} = B_0 + B_1 z_i - \dfrac{1}{\lambda_1}\displaystyle\sum_{k=1}^{n_i}\alpha_{ik} - \sum_{k=1}^{N}\sum_{l=1}^{n_k}\alpha_{kl}\left(\frac{z_k - 1}{2}\right)\left(\frac{z_i - 1}{2}\right) \\ \varphi_1(x_{kl})^T\varphi_1(x_{ij}) - \displaystyle\sum_{k=1}^{N}\sum_{l=1}^{n_k}\alpha_{kl}\left(\frac{z_k + 1}{2}\right)\left(\frac{z_i + 1}{2}\right)\varphi_2(x_{kl})^T\varphi_2(x_{ij}) - \dfrac{\alpha_{ij}}{\lambda_2}, \quad i = 1, \ldots, N, j = 1, \ldots, n_i, \\ 0 = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{n_i}\alpha_{ij}, \\ 0 = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{n_i}\alpha_{ij} z_i. \end{cases} \tag{8}$$

This can be written as the linear system

$$\left[\begin{array}{cc|c} 0_{2\times 2} & & 1_M{}^T \\ & & Z^T \\ \hline 1_M & Z & -K_1 - K_2 - \frac{\Gamma}{\lambda_1} - \frac{I_{M\times M}}{\lambda_2} \end{array}\right]\begin{bmatrix} B_0 \\ B_1 \\ \alpha \end{bmatrix} = \begin{bmatrix} 0_{2\times 1} \\ y \end{bmatrix}, \tag{9}$$

with $y = [y_{11} \ \ldots \ y_{1n_1}y_{21} \ \ldots \ y_{Nn_N}]^T \in \mathbb{R}^M$, $Z = [\underbrace{z_1 \ \ldots \ z_1}_{n_1}\underbrace{z_2 \ \ldots \ z_2}_{n_2}\ldots\underbrace{z_N \ \ldots \ z_N}_{n_N}]^T$, $x = [x_{11} \ \ldots \ x_{1n_1}x_{21} \ \ldots \ x_{Nn_N}]^T \in \mathbb{R}^M$, $\Gamma = \text{blockdiag}(1_{n_1\times n_1} \ \ldots \ 1_{n_N\times n_N})$, kernel matrix $K_1 \in \mathbb{R}^{M\times M}$ with elements $K_{1_{rs}} = (\frac{z_r-1}{2})(\frac{z_s-1}{2})\varphi_1(x_s)^T\varphi_1(x_r) = (\frac{z_r-1}{2})(\frac{z_s-1}{2})K_1(x_s, x_r)$ and kernel matrix $K_2 \in \mathbb{R}^{M\times M}$ with elements $K_{2_{rs}} = (\frac{z_r+1}{2})(\frac{z_s+1}{2})\varphi_2(x_s)^T\varphi_2(x_r) = (\frac{z_r+1}{2})(\frac{z_s+1}{2})K_2(x_s, x_r)$, where $K_1(\cdot,\cdot)$ and $K_2(\cdot,\cdot)$ are positive definite kernel functions. Values for the regularization constants (and kernel parameters) can be obtained via cross-validation procedures. This is further discussed in Section 4.

2.3. Prediction: random intercept LS-SVM classification

Now that the parameters w_1, w_2, B_0 and B_1 in (5) are estimated, these can be used to classify a longitudinal data set $\{\{x_{ij}^*, y_{ij}^*\}_{j=1}^{g_i}\}_{i=1}^{G}$ with unknown class labels z_i^*, where G represents the number of cases, g_i the number of observations for case i and $H = \sum_{i=1}^{G} g_i$ the total number of observations. The proposed optimization problem for prediction is formulated as

$$(\hat{e^*}, \hat{b^*}, \hat{z^*}) = \underset{e^*, b^*, z^*}{\text{argmin}} \frac{\lambda_1}{2}\sum_{i=1}^{G} b_i^{*2} + \frac{\lambda_2}{2}\sum_{i=1}^{G}\sum_{j=1}^{g_i} e_{ij}^{*2}, \tag{10}$$

subject to

$$y_{ij}^* = \hat{B_0} + \hat{B_1} z_i^* + b_i^* + \left(\frac{z_i^* - 1}{2}\right)\hat{w_1}{}^T\varphi_1(x_{ij}^*) + \left(\frac{z_i^* + 1}{2}\right)\hat{w_2}{}^T\varphi_2(x_{ij}^*) + e_{ij}^*, \quad i = 1, \ldots, G, \ j = 1, \ldots, g_i, \tag{11}$$

with random intercepts $b^* = [b_1^* \ \dots \ b_G^*]^T$, error variables $e^* = [e_{11}^* \ \dots \ e_{1g_1}^* e_{21}^* \dots \ e_{Gg_G}^*]^T \in \mathbb{R}^H$, class labels $z^* = [z_1^* \ \dots \ z_G^*]^T$, weights $\hat{w}_1 = -\sum_{i=1}^N \sum_{j=1}^{n_i} \hat{\alpha}_{ij} (\frac{z_i - 1}{2}) \varphi_1(x_{ij})$ and weights $\hat{w}_2 = -\sum_{i=1}^N \sum_{j=1}^{n_i} \hat{\alpha}_{ij} (\frac{z_i + 1}{2}) \varphi_2(x_{ij})$. The Lagrangian for the problem in (10)–(11) is

$$\begin{aligned} \mathcal{L}_2(e^*, b^*, z^*, \nu) = \frac{\lambda_1}{2} \sum_{i=1}^{G} b_i^{*2} + \frac{\lambda_2}{2} \sum_{i=1}^{G} \sum_{j=1}^{g_i} e_{ij}^{*2} - \sum_{i=1}^{G} \sum_{j=1}^{g_i} \nu_{ij} \Bigg(y_{ij}^* - \hat{B}_0 \\ - \hat{B}_1 z_i^* - b_i^* - \left(\frac{z_i^* - 1}{2} \right) \hat{w}_1{}^T \varphi_1(x_{ij}^*) - \left(\frac{z_i^* + 1}{2} \right) \hat{w}_2{}^T \varphi_2(x_{ij}^*) - e_{ij}^* \Bigg), \end{aligned} \tag{12}$$

with Lagrange multipliers $\nu = [\nu_{11} \ \dots \ \nu_{1g_1} \nu_{21} \ \dots \ \nu_{Gg_G}]^T \in \mathbb{R}^H$. The conditions for optimality result in

$$\begin{cases} \dfrac{\partial \mathcal{L}_2}{\partial e_{ij}^*} = 0 \to e_{ij}^* = -\dfrac{\nu_{ij}}{\lambda_2}, \quad i = 1, \dots, G, j = 1, \dots, g_i, \\ \dfrac{\partial \mathcal{L}_2}{\partial b_i^*} = 0 \to b_i^* = -\dfrac{1}{\lambda_1} \displaystyle\sum_{j=1}^{g_i} \nu_{ij}, \quad i = 1, \dots, G, \\ \dfrac{\partial \mathcal{L}_2}{\partial z_i^*} = 0 \to 0 = \displaystyle\sum_{j=1}^{g_i} \nu_{ij} \left(\hat{B}_1 + \dfrac{\hat{w}_1{}^T \varphi_1(x_{ij}^*)}{2} + \dfrac{\hat{w}_2{}^T \varphi_2(x_{ij}^*)}{2} \right), \quad i = 1, \dots, G, \\ \dfrac{\partial \mathcal{L}_2}{\partial \nu_{ij}} = 0 \to y_{ij}^* = \hat{B}_0 + \hat{B}_1 z_i^* + b_i^* + \left(\dfrac{z_i^* - 1}{2} \right) \hat{w}_1{}^T \varphi_1(x_{ij}^*) + \left(\dfrac{z_i^* + 1}{2} \right) \hat{w}_2{}^T \varphi_2(x_{ij}^*) + e_{ij}^*, \\ \quad i = 1, \dots, G, \ j = 1, \dots, g_i. \end{cases} \tag{13}$$

Elimination of e_{ij}^* and b_i^* gives

$$\begin{cases} 0 = \displaystyle\sum_{j=1}^{g_i} \nu_{ij} \left(\hat{B}_1 + \dfrac{\hat{w}_1{}^T \varphi_1(x_{ij}^*)}{2} + \dfrac{\hat{w}_2{}^T \varphi_2(x_{ij}^*)}{2} \right), \quad i = 1, \dots, G, \\ y_{ij}^* = \hat{B}_0 + \hat{B}_1 z_i^* - \dfrac{1}{\lambda_1} \displaystyle\sum_{k=1}^{g_i} \nu_{ik} + \left(\dfrac{z_i^* - 1}{2} \right) \hat{w}_1{}^T \varphi_1(x_{ij}^*) + \left(\dfrac{z_i^* + 1}{2} \right) \hat{w}_2{}^T \varphi_2(x_{ij}^*) - \dfrac{\nu_{ij}}{\lambda_2}, \\ \quad i = 1, \dots, G, j = 1, \dots, g_i. \end{cases} \tag{14}$$

The optimization problem in the dual space is formulated as

$$\left[\begin{array}{c|c} 0_{G \times G} & A^{*T} \\ \hline A^* & -\frac{\Theta}{\lambda_1} - \frac{I_{H \times H}}{\lambda_2} \end{array} \right] \begin{bmatrix} z^* \\ \nu \end{bmatrix} = \begin{bmatrix} 0_G \\ y^* - \hat{B}_0 1_H + P^* \end{bmatrix}, \tag{15}$$

with $y^* = [y_{11}^* \ \dots \ y_{1g_1}^* y_{21}^* \ \dots \ y_{Gg_G}^*]^T \in \mathbb{R}^H$, $\Theta = \text{blockdiag}(1_{g_1 \times g_1} \ \dots \ 1_{g_G \times g_G})$, $A^* = \text{blockdiag}([\hat{B}_1 + \frac{\hat{w}_1{}^T \varphi_1(x_{11}^*)}{2} + \frac{\hat{w}_2{}^T \varphi_2(x_{11}^*)}{2} \ \dots \ \hat{B}_1 + \frac{\hat{w}_1{}^T \varphi_1(x_{1g_1}^*)}{2} + \frac{\hat{w}_2{}^T \varphi_2(x_{1g_1}^*)}{2}]^T \ \dots \ [\hat{B}_1 + \frac{\hat{w}_1{}^T \varphi_1(x_{G1}^*)}{2} + \frac{\hat{w}_2{}^T \varphi_2(x_{G1}^*)}{2} \ \dots \ \hat{B}_1 + \frac{\hat{w}_1{}^T \varphi_1(x_{Gg_G}^*)}{2} + \frac{\hat{w}_2{}^T \varphi_2(x_{Gg_G}^*)}{2}]^T)$ and $P^* = [\frac{\hat{w}_1{}^T \varphi_1(x_{11}^*)}{2} - \frac{\hat{w}_2{}^T \varphi_2(x_{11}^*)}{2} \ \dots \ \frac{\hat{w}_1{}^T \varphi_1(x_{1g_1}^*)}{2} - \frac{\hat{w}_2{}^T \varphi_2(x_{1g_1}^*)}{2} \ \frac{\hat{w}_1{}^T \varphi_1(x_{21}^*)}{2} - \frac{\hat{w}_2{}^T \varphi_2(x_{21}^*)}{2} \ \dots \ \frac{\hat{w}_1{}^T \varphi_1(x_{Gg_G}^*)}{2} - \frac{\hat{w}_2{}^T \varphi_2(x_{Gg_G}^*)}{2}]^T$. The resulting longitudinal LS-SVM classifier is defined as $\text{sign}(\hat{z}^*)$. Note that solving (15) with all G test cases or with one test case at the time will result in identical predictions. Section 4 provides a discussion on the selection of the values for the regularization constants λ_1 and λ_2 and the kernel parameters.

2.4. An alternative prediction formulation

In this section an alternative approach to the prediction step in Section 2.3 is formulated. Instead of obtaining the unknown class labels $\hat{z}^*$ from solving (15), the cost of assigning a test case, e.g. $\{[x_{i1}^* \dots x_{ig_i}^*]^T, [y_{i1}^* \dots y_{ig_i}^*]^T\}$, to a specific class can be calculated. Final prediction is obtained by assigning the case to the class which is associated with the lowest cost. The problem in (10)–(11) is redefined as

$$(\hat{e}^*, \hat{b}^*) = \underset{e^*, b^*}{\text{argmin}} \frac{\lambda_1}{2} b_i^{*2} + \frac{\lambda_2}{2} \sum_{j=1}^{g_i} e_{ij}^{*2}, \tag{16}$$

subject to

$$y_{ij}^* = \hat{B}_0 + \hat{B}_1 z_i^* + b_i^* + \left(\frac{z_i^* - 1}{2} \right) \hat{w}_1{}^T \varphi_1(x_{ij}^*) + \left(\frac{z_i^* + 1}{2} \right) \hat{w}_2{}^T \varphi_2(x_{ij}^*) + e_{ij}^*, \quad j = 1, \dots, g_i. \tag{17}$$

This formulation assumes that the true class label z_i^* is provided, while the estimates $\hat{B}_0$, $\hat{B}_1$, $\hat{w}_1$ and $\hat{w}_2$, obtained from the regression step in Section 2.2, are plugged in. The Lagrangian for this problem becomes

$$\begin{aligned}\mathcal{L}_3(e^*, b^*, \nu) = & \frac{\lambda_1}{2} b_i^{*2} + \frac{\lambda_2}{2} \sum_{j=1}^{g_i} e_{ij}^{*2} - \sum_{j=1}^{g_i} \nu_{ij} (y_{ij}^* - \hat{B}_0 - \hat{B}_1 z_i^* - b_i^* \\ & - \left(\frac{z_i^* - 1}{2}\right) \hat{w}_1{}^T \varphi_1(x_{ij}^*) - \left(\frac{z_i^* + 1}{2}\right) \hat{w}_2{}^T \varphi_2(x_{ij}^*) - e_{ij}^*),\end{aligned} \tag{18}$$

with Lagrange multipliers $\nu_i = [\nu_{i1} \ \dots \ \nu_{ig_i}]^T$ and the conditions for optimality yield

$$\begin{cases} \dfrac{\partial \mathcal{L}_3}{\partial e_{ij}^*} = 0 \to e_{ij}^* = -\dfrac{\nu_{ij}}{\lambda_2}, \quad j = 1, \dots, g_i, \\ \dfrac{\partial \mathcal{L}_3}{\partial b_i^*} = 0 \to b_i^* = -\dfrac{1}{\lambda_1} \displaystyle\sum_{j=1}^{g_i} \nu_{ij}, \\ \dfrac{\partial \mathcal{L}_3}{\partial \nu_{ij}} = 0 \to y_{ij}^* = \hat{B}_0 + \hat{B}_1 z_i^* + b_i^* + \left(\dfrac{z_i^* - 1}{2}\right) \hat{w}_1{}^T \varphi_1(x_{ij}^*) + \left(\dfrac{z_i^* + 1}{2}\right) \hat{w}_2{}^T \varphi_2(x_{ij}^*) + e_{ij}^*, \quad j = 1, \dots, g_i. \end{cases} \tag{19}$$

Elimination of e_{ij}^* and b_i^* yields

$$y_{ij}^* = \hat{B}_0 + \hat{B}_1 z_i^* - \frac{1}{\lambda_1} \sum_{k=1}^{g_i} \nu_{ik} + \left(\frac{z_i^* - 1}{2}\right) \hat{w}_1{}^T \varphi_1(x_{ij}^*) + \left(\frac{z_i^* + 1}{2}\right) \hat{w}_2{}^T \varphi_2(x_{ij}^*) - \frac{\nu_{ij}}{\lambda_2}, \quad j = 1, \dots, g_i. \tag{20}$$

The solution to this optimization problem can be found by solving the linear system

$$\left[-\frac{1_{g_i \times g_i}}{\lambda_1} - \frac{I_{g_i \times g_i}}{\lambda_2}\right] [\nu_i] = [y_i^* - (\hat{B}_0 + \hat{B}_1 z_i^*) 1_{g_i} - D_i^*], \tag{21}$$

with $y_i^* = [y_{i1}^* \dots y_{ig_i}^*]^T$, $D_i^* = [(\frac{z_i^*-1}{2}) \hat{w}_1{}^T \varphi_1(x_{i1}^*) + (\frac{z_i^*+1}{2}) \hat{w}_2{}^T \varphi_2(x_{i1}^*) \ \dots \ (\frac{z_i^*-1}{2}) \hat{w}_1{}^T \varphi_1(x_{ig_i}^*) + (\frac{z_i^*+1}{2}) \hat{w}_2{}^T \varphi_2(x_{ig_i}^*)]^T$. Following (19) the cost function in (16) equals $\frac{1}{2\lambda_1} (\sum_{j=1}^{g_i} \nu_{ij})^2 + \frac{1}{2\lambda_2} \sum_{j=1}^{g_i} \nu_{ij}^2$. This cost function can be calculated for the first, i.e. $z_i^* = -1$, and the second, i.e. $z_i^* = 1$, class by solving (21). The classification rule assigns the case $\{[x_{i1}^* \dots x_{ig_i}^*]^T, [y_{i1}^* \dots y_{ig_i}^*]^T\}$ to the class with the smallest cost.

3. Extensions

In the following sections possible extensions of the model in (5) are explored.

3.1. Multi-class classification

The alternative approach in Section 2.4 has been introduced because it allows direct generalization of the binary longitudinal LS-SVM classifier to a multi-class version. This section illustrates this extension for the development of a longitudinal LS-SVM classifier for C classes. Similar to Section 2.2, assume the longitudinal training data $\{\{x_{ij}, y_{ij}\}_{j=1}^{n_i}\}_{i=1}^N$, but with known class labels $z_{1i}, \dots, z_{Ci} \in \{-1, +1\}$ for case i. These class coding labels equal -1 except for $z_{ci} = 1$ if case i belongs to class c. As before, the training step consists of a regression modeling and it is formulated by the optimization problem

$$(\hat{w}, \hat{e}, \hat{b}, \hat{B}_0, \hat{B}_1, \dots, \hat{B}_C) = \underset{w,e,b,B_0,B_1,\dots,B_C}{\operatorname{argmin}} \frac{1}{2} w^T w + \frac{\lambda_1}{2} \sum_{i=1}^N b_i^2 + \frac{\lambda_2}{2} \sum_{i=1}^N \sum_{j=1}^{n_i} e_{ij}^2, \tag{22}$$

subject to the model

$$y_{ij} = B_0 + \sum_{c=1}^C \left(\frac{z_{ci} + 1}{2}\right) (B_c + w_c^T \varphi_c(x_{ij})) + b_i + e_{ij}, \quad i = 1, \dots, N, \ j = 1, \dots, n_i, \tag{23}$$

with weighting vector $w = [w_1^T \ \dots \ w_C^T]^T$, model parameters $B_0, B_1, \dots, B_C$, and mapping $\varphi_c(\cdot)$, $c = 1, \dots, C$. The dual solution to this optimization problem can be obtained by solving the linear system

$$\left[\begin{array}{c|c} 0_{(C+1)\times(C+1)} & \begin{matrix} 1_M{}^T \\ Z_1{}^T \\ \vdots \\ Z_C{}^T \end{matrix} \\ \hline \begin{matrix} 1_M & Z_1 & \dots & Z_C \end{matrix} & -\displaystyle\sum_{c=1}^C K_c - \frac{\Gamma}{\lambda_1} - \frac{I_{M\times M}}{\lambda_2} \end{array}\right] \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_C \\ \alpha \end{bmatrix} = \begin{bmatrix} 0_{(C+1)} \\ y \end{bmatrix}, \tag{24}$$

with $Z_c = \frac{1}{2}([\underbrace{z_{c1} \ldots z_{c1}}_{n_1} \underbrace{z_{c2} \ldots z_{c2}}_{n_2} \ldots \underbrace{z_{cN} \ldots z_{cN}}_{n_N}]^T + 1_M)$, kernel matrix $K_c \in \mathbb{R}^{M \times M}$ with elements $K_{c_{rs}} = (\frac{z_{cr}+1}{2})(\frac{z_{cs}+1}{2})K_c(x_s, x_r)$, where $K_c(\cdot,\cdot)$ is a positive definite kernel function. The obtained estimates for B_0, B_1, ...,B_C and $w = [w_1^T \ldots w_C^T]^T$ are now used to predict the class for a new longitudinal data set $\{\{x_{ij}^*, y_{ij}^*\}_{j=1}^{g_i}\}_{i=1}^{G}$ with unknown class labels. The approach from Section 2.4 can be used to classify the ith test case $\{[x_{i1}^* \ldots x_{ig_i}^*]^T, [y_{i1}^* \ldots y_{ig_i}^*]^T\}$. The multi-class extension for the problem in (16)–(17) is defined as

$$(\hat{e}^*, \hat{b}^*) = \underset{e^*, b^*}{\operatorname{argmin}} \frac{\lambda_1}{2} b_i^{*2} + \frac{\lambda_2}{2} \sum_{j=1}^{g_i} e_{ij}^{*2}, \tag{25}$$

subject to

$$y_{ij}^* = \hat{B}_0 + \sum_{c=1}^{C} \left(\frac{z_{ci}^* + 1}{2}\right) (\hat{B}_c + \hat{w}_c{}^T \varphi_c(x_{ij}^*)) + b_i^* + e_{ij}^*, \quad j = 1, \ldots, g_i. \tag{26}$$

The Lagrangian for this problem is

$$\mathcal{L}_4(e^*, b^*, \nu) = \frac{\lambda_1}{2} b_i^{*2} + \frac{\lambda_2}{2} \sum_{j=1}^{g_i} e_{ij}^{*2} - \sum_{j=1}^{g_i} \nu_{ij} \left(y_{ij}^* - \hat{B}_0 - \sum_{c=1}^{C} \left(\frac{z_{ci}^* + 1}{2}\right) (\hat{B}_c + \hat{w}_c{}^T \varphi_c(x_{ij}^*)) - b_i^* - e_{ij}^*\right), \tag{27}$$

and the conditions for optimality yield

$$\begin{cases} \dfrac{\partial \mathcal{L}_4}{\partial e_{ij}^*} = 0 \rightarrow e_{ij}^* = -\dfrac{\nu_{ij}}{\lambda_2}, \quad j = 1, \ldots, g_i, \\ \dfrac{\partial \mathcal{L}_4}{\partial b_i^*} = 0 \rightarrow b_i^* = -\dfrac{1}{\lambda_1} \displaystyle\sum_{j=1}^{g_i} \nu_{ij}, \\ \dfrac{\partial \mathcal{L}_4}{\partial \nu_{ij}} = 0 \rightarrow y_{ij}^* = \hat{B}_0 + \displaystyle\sum_{c=1}^{C} \left(\frac{z_{ci}^* + 1}{2}\right) (\hat{B}_c + \hat{w}_c{}^T \varphi_c(x_{ij}^*)) + b_i^* + e_{ij}^*, \quad j = 1, \ldots, g_i. \end{cases} \tag{28}$$

Elimination of e_{ij}^* and b_i^* gives

$$y_{ij}^* = \hat{B}_0 + \sum_{c=1}^{C} \left(\frac{z_{ci}^* + 1}{2}\right) (\hat{B}_c + \hat{w}_c{}^T \varphi_c(x_{ij}^*)) - \frac{1}{\lambda_1} \sum_{j=1}^{g_i} \nu_{ij} - \frac{\nu_{ij}}{\lambda_2}, \quad j = 1, \ldots, g_i. \tag{29}$$

The dual solution for this problem can be found by solving

$$\left[-\frac{1_{g_i \times g_i}}{\lambda_1} - \frac{I_{g_i \times g_i}}{\lambda_2}\right][\nu_i] = \left[y_i^* - \left(\hat{B}_0 + \sum_{c=1}^{C} \left(\frac{z_{ci}^* + 1}{2}\right) \hat{B}_c\right) 1_{g_i} - L_i^*\right], \tag{30}$$

with $L_i^* = [\sum_{c=1}^{C} (\frac{z_{ci}^*+1}{2}) \hat{w}_c{}^T \varphi_c(x_{i1}^*) \ \ldots \ \sum_{c=1}^{C} (\frac{z_{ci}^*+1}{2}) \hat{w}_c{}^T \varphi_c(x_{ig_i}^*)]^T$. The cost, i.e. $\frac{1}{2\lambda_1}(\sum_{j=1}^{g_i} \nu_{ij})^2 + \frac{1}{2\lambda_2} \sum_{j=1}^{g_i} \nu_{ij}^2$, of assigning a test case to a specific class can be calculated by solving (30) with the corresponding class coding labels (e.g. $z_{1i}^* = 1$ and $z_{2i}^* = \ldots = z_{Ci}^* = -1$ for class 1). Finally, the classification rule assigns the ith test case $\{[x_{i1}^* \ldots x_{ig_i}^*]^T, [y_{i1}^* \ldots y_{ig_i}^*]^T\}$ to the class with the smallest cost.

3.2. Multiple random effects

The proposed random intercept LS-SVM classifier (Section 2.3) can be extended to allow multiple random effects. Assume the longitudinal test data set $\{\{x_{ij}^*, y_{ij}^*\}_{j=1}^{g_i}\}_{i=1}^{G}$ with unknown class label z_i^* for case i. The optimization problem for prediction with r random effects is formulated as

$$(\hat{e}^*, \hat{b}^*, \hat{z}^*) = \underset{e^*, b^*, z^*}{\operatorname{argmin}} \frac{\lambda_3}{2} \sum_{i=1}^{G} b_i^{*T} R\, b_i^* + \frac{\lambda_4}{2} \sum_{i=1}^{G} e_i^{*T} E_i e_i^*, \tag{31}$$

subject to

$$y_{ij}^* = \hat{B}_0 + \hat{B}_1 z_i^* + \left(\frac{z_i^* - 1}{2}\right) \hat{w}_1{}^T \varphi_1(x_{ij}^*) + \left(\frac{z_i^* + 1}{2}\right) \hat{w}_2{}^T \varphi_2(x_{ij}^*) + b_i^{*T} u_{ij}^* + e_{ij}^*, \quad i = 1, \ldots, G, j = 1, \ldots, g_i, \tag{32}$$

with positive regularization constants λ_3 and λ_4, random effect covariates $u_{ij}^* = [u_{ij1}^* \ \ldots \ u_{ijr}^*]^T$, and the covariance matrices R^{-1} and E_i^{-1} of the random effects $b_i^* = [b_{1i}^* \ \ldots \ b_{ri}^*]^T$ and residuals $e_i^* = [e_{i1}^* \ \ldots \ e_{ig_i}^*]^T$, respectively. The Lagrangian for this problem can be written as

$$\mathcal{L}_5(e^*, b^*, z^*, \nu) = \frac{\lambda_3}{2}\sum_{i=1}^{G} b_i^{*T} R\, b_i^* + \frac{\lambda_4}{2}\sum_{i=1}^{G} e_i^{*T} E_i e_i^* - \sum_{i=1}^{G}\sum_{j=1}^{g_i} \nu_{ij}\left(y_{ij}^* - \hat{B}_0 - \hat{B}_1 z_i^* - \left(\frac{z_i^*-1}{2}\right)\hat{w}_1{}^T\varphi_1(x_{ij}^*) - \left(\frac{z_i^*+1}{2}\right)\hat{w}_2{}^T\varphi_2(x_{ij}^*) - b_i^{*T}u_{ij}^* - e_{ij}^*\right). \tag{33}$$

The corresponding conditions for optimality yield

$$\begin{cases} \dfrac{\partial \mathcal{L}_5}{\partial e_{ij}^*} = 0 \to 0 = \lambda_4 \displaystyle\sum_{k=1}^{g_i} e_{ik}^* E_{ijk} + \nu_{ij}, \quad i = 1,\ldots,G, j = 1,\ldots,g_i, \\ \dfrac{\partial \mathcal{L}_5}{\partial b_{ki}^*} = 0 \to 0 = \lambda_3 \displaystyle\sum_{l=1}^{r} R_{lk} b_{li}^* + \sum_{j=1}^{g_i} \nu_{ij} u_{ijk}^*, \quad i = 1,\ldots,G, k = 1,\ldots,r, \\ \dfrac{\partial \mathcal{L}_5}{\partial z_i^*} = 0 \to 0 = \displaystyle\sum_{j=1}^{g_i} \nu_{ij}\left(\hat{B}_1 + \frac{\hat{w}_1{}^T\varphi_1(x_{ij}^*)}{2} + \frac{\hat{w}_2{}^T\varphi_2(x_{ij}^*)}{2}\right), \quad i = 1,\ldots,G, \\ \dfrac{\partial \mathcal{L}_5}{\partial \nu_{ij}} = 0 \to y_{ij}^* = \hat{B}_0 + \hat{B}_1 z_i^* + \left(\dfrac{z_i^*-1}{2}\right)\hat{w}_1{}^T\varphi_1(x_{ij}^*) + \left(\dfrac{z_i^*+1}{2}\right)\hat{w}_2{}^T\varphi_2(x_{ij}^*) \\ \quad + \displaystyle\sum_{k=1}^{r} b_{ki}^* u_{ijk}^* + e_{ij}^*, \quad i = 1,\ldots,G, j = 1,\ldots,g_i. \end{cases} \tag{34}$$

Elimination of e_{ij}^* and b_{ki}^* gives

$$\begin{cases} 0 = \displaystyle\sum_{j=1}^{g_i} \nu_{ij}\left(\hat{B}_1 + \frac{\hat{w}_1{}^T\varphi_1(x_{ij}^*)}{2} + \frac{\hat{w}_2{}^T\varphi_2(x_{ij}^*)}{2}\right), \quad i = 1,\ldots,G, \\ y_i^* = \hat{B}_0 + \hat{B}_1 z_i^* + \left(\dfrac{z_i^*-1}{2}\right)[\hat{w}_1{}^T\Phi_{i1}^*]^T + \left(\dfrac{z_i^*+1}{2}\right)[\hat{w}_2{}^T\Phi_{i2}^*]^T - \dfrac{1}{\lambda_3}U_i^{*T}R^{-1}U_i^*\nu_i - \dfrac{1}{\lambda_4}E_i^{-1}\nu_i, \\ \quad i = 1,\ldots,G, \end{cases} \tag{35}$$

with $y_i^* = [y_{i1}^* \ \ldots \ y_{ig_i}^*]^T$, $\nu_i = [\nu_{i1} \ \ldots \ \nu_{ig_i}]^T$, $U_i^* = [u_{i1}^* \ \ldots \ u_{ig_i}^*]$, $\Phi_{i1}^* = [\varphi_1(x_{i1}^*) \ \ldots \ \varphi_1(x_{ig_i}^*)]$, $\Phi_{i2}^* = [\varphi_2(x_{i1}^*) \ \ldots \ \varphi_2(x_{ig_i}^*)]$, whereas E_{ijk} and R_{jk} denote the element on the jth row and kth column of the covariance matrices E_i and R, respectively. The dual solution to this problem can be found by solving

$$\left[\begin{array}{c|c} 0_{G\times G} & A^{*T} \\ \hline A^* & F^* \end{array}\right]\begin{bmatrix} z^* \\ \nu \end{bmatrix} = \begin{bmatrix} 0_G \\ y^* - \hat{B}_0 1_H + P^* \end{bmatrix}, \tag{36}$$

where $F^* = \text{blockdiag}(-\frac{1}{\lambda_3}U_1^{*T}R^{-1}U_1^* - \frac{1}{\lambda_4}E_1^{-1} \ \ldots \ - \frac{1}{\lambda_3}U_G^{*T}R^{-1}U_G^* - \frac{1}{\lambda_4}E_G^{-1})$, while A^* and P^* are identical to their formulations in (15). In case one assumes only a random intercept and a simple covariance structure for E_i, (36) simplifies to (15). The final class assignment is obtained by applying $\text{sign}(\hat{z}^*)$.

4. Model selection

Selection of appropriate values for the regularization parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and the kernel parameters is an important issue. Moreover, the use of multiple random effects requires the estimation of the covariance matrices R^{-1} and E_i^{-1}. These topics are discussed in detail in the following sections.

4.1. Random intercept LS-SVM

The typical procedure to tune regularization parameters in the context of conventional LS-SVMs is by means of cross-validation or repeated random sampling, resulting in training and validation data. On the other hand, some complications arise when using these methods to tune the parameters of the mixed effects LS-SVM models. The training and the prediction step each involve two regularization parameters (i.e. λ_1, λ_2) and additional kernel parameters (e.g. σ_1, σ_2). Although one could tune the parameters of the training step and the prediction step in a separate way, we decided to tune them together. This means that one gridsearch is performed and for a fixed combination of parameters (9) (or (24)) and (15) (or (21), (30)) are solved. Since the main goal of this paper is classification, the preferred performance measure is classification accuracy. Therefore, the combination of tuning parameters that maximizes the classification performance (i.e. at the level of the prediction step) on the validation data (cf. cross-validation or repeated random sampling) is chosen. The advantage of this approach is the lower computational cost of tuning only two parameters (kernel parameters excluded). Performing a separate model selection for the training and the prediction step would require tuning three parameters (kernel parameters excluded) since one of the lambdas in (10) (or (16), (25)) can be omitted then. Furthermore, cross-validation and random

sampling involve splitting data in a training set and validation set. Remark that data splitting was performed in such a way that cases instead of individual observations were repeatedly assigned to these sets. The final resulting model selection procedure for random intercept LS-SVM classification is summarized in Algorithm 1.

Fig. 1 further illustrates the importance of a careful model selection. The data that are used are identical to the training and validation data of the toy problem in Section 5.1.1. While Fig. 2 visualizes the optimal results, in the sense that the selected parameters maximize classification performance, Fig. 1 shows the effect of changing one parameter at a time on the estimated class means (dashed lines) and the case-specific predicted curve (red line) for the observed curve (green line). Decreasing the bandwidth of the RBF kernel results in less smooth class means and predicted curve. Increasing λ_2 strengthens this effect

Algorithm 1 (*Model selection and classification strategy for random intercept LS-SVM*)**.**

1: Start a gridsearch using cross-validation or repeated random sampling on training data to obtain values for λ_1 and λ_2. This includes step 2 and step 3:
2: Perform regression on training data by solving (9) or (24) to obtain estimates for $w, B_0, B_1\,(, \ldots, B_C)$.
3: Use these estimates to perform prediction on validation data by solving (15) or (21) for binary problems and (30) for multi-class problems and select the λ_1 and λ_2 with the best classification performance.
4: Finally, use the tuned λ_1 and λ_2 and corresponding estimates $\hat{w}, \hat{B_0}, \hat{B_1}\,(, \ldots, \hat{B_C})$ to perform prediction on test data by solving (15) or (21) for binary problems and (30) for multi-class problems.

since more emphasis is put on the sum of squared residuals. An increasing λ_1 value makes the class means even less smooth and shrinks the predicted curve toward the class mean.

4.2. Multiple random effects LS-SVM

The proposed method to deal with multiple random effects assumes that the covariance matrices R^{-1} and E_i^{-1} are known, while this is in practice not the case. A solution to this problem can be found by first applying the random intercept LS-SVM approach from Sections 2.2 and 2.3, in combination with the model selection procedure from Section 4.1, in order to estimate the mean (i.e. $\hat{B_0}, \hat{B_1}, \hat{w_1}, \hat{w_2}$) for each class. Then, for all observations of each case the corresponding residuals are computed by removing the estimated mean of the true class to which this case belongs. The residuals of all cases can then be used to estimate R^{-1} and E_i^{-1} using (restricted) maximum likelihood theory (Patterson and Thompson, 1971). Finally, values for the regularization constants λ_3 and λ_4 can be obtained via cross-validation or repeated random sampling of the data. Remark that since the problem in (31) requires a separate gridsearch, one of the lambdas could have been removed by dividing one through the other. In Section 3.2 both λ_3 and λ_4 were included to have a more consistent presentation. In practice, one lambda can be fixed while a gridsearch is used for the other one. The full procedure is summarized in Algorithm 2.

Algorithm 2 (*Model selection and classification strategy for LS-SVM with multiple random effects*)**.**

1: Start with a random intercept LS-SVM and proceed as described in Algorithm 1 to obtain the estimates $\hat{w}, \hat{B_0}, \hat{B_1}\,(, \ldots, \hat{B_C})$.
2: Compute the residuals for all observations of each case by removing the estimated mean, resulting from the random intercept LS-SVM step, while the given class label is used to decide about the estimated mean.
3: Use (restricted) maximum likelihood to estimate the covariance matrices R^{-1} and E_i^{-1} from the residuals of all cases.
4: Use cross-validation or repeated random sampling on training data to obtain values for λ_3 and λ_4. This includes step 5:
5: Perform prediction on validation data using the multiple random effects LS-SVM by solving (36) with the obtained $\hat{w}$, $\hat{B_0}$, $\hat{B_1}\,(, \ldots, \hat{B_C})$. Select the λ_3 and λ_4 with the best performance.
6: Finally, perform prediction on test data by solving (36) with $\hat{w}, \hat{B_0}, \hat{B_1}\,(, \ldots, \hat{B_C})$ and tuned λ_3 and λ_4.

5. Examples

This section provides examples of the application of the longitudinal LS-SVM classifier for 4 simulation studies and 2 real-life problems.

5.1. Simulation studies

5.1.1. Toy problem

For each class, 50 curves were generated according to the model in (5), resulting in the unbalanced data set in the upper right panel of Fig. 2. A simple variance structure, with a standard deviation of 80, was chosen for the normally distributed residual error and the standard deviation for the normally distributed random intercept was 400. The number of observations per curve varied between 5 and 10. This data set was 25 times split in a random way into a set for training (2/3 of the cases) and a set for validation (1/3 of the cases). An RBF kernel was used for the longitudinal LS-SVM classifier and the kernel bandwidths for both classes were constrained to be equal to reduce the computational cost. A gridsearch was

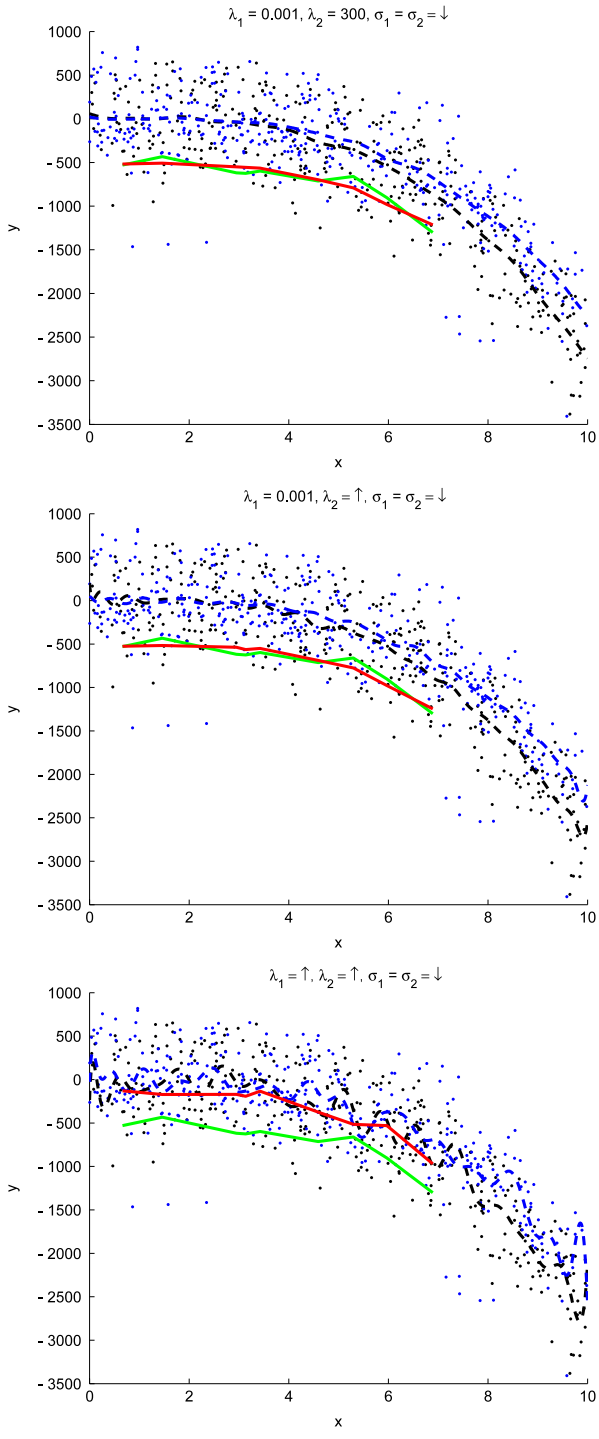

Fig. 1. Model selection. Top: decreasing the bandwidth σ of the RBF kernel results in wiggly estimates for the means (dashed lines) and the case-specific predicted curve (red line). The observed curve is visualized in green. Middle: putting more weight on the squared errors by increasing the regularization parameter λ_2 makes the estimated means and predicted curve more unstable. Bottom: increasing λ_1 shrinks the case-specific predicted curve to the estimated mean. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

performed in order to tune the parameters λ_1 ($\in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.5, 0.7, 1, 2, 3, 4, 5, 10, 20, 30, 40\}$), λ_2 ($\in \{10^{-2}, 0.1, 1, \ldots, 20, 30, 40, 80, 100, 150, 200, 300, 400, 600, 800\}$) and σ_1, σ_2 ($\in \{10^{-3}, 10^{-2}, 0.1, 1, \ldots, 25\}$). Use of the RBF kernel, training data splitting and the gridsearch settings were similar for all examples of the paper.

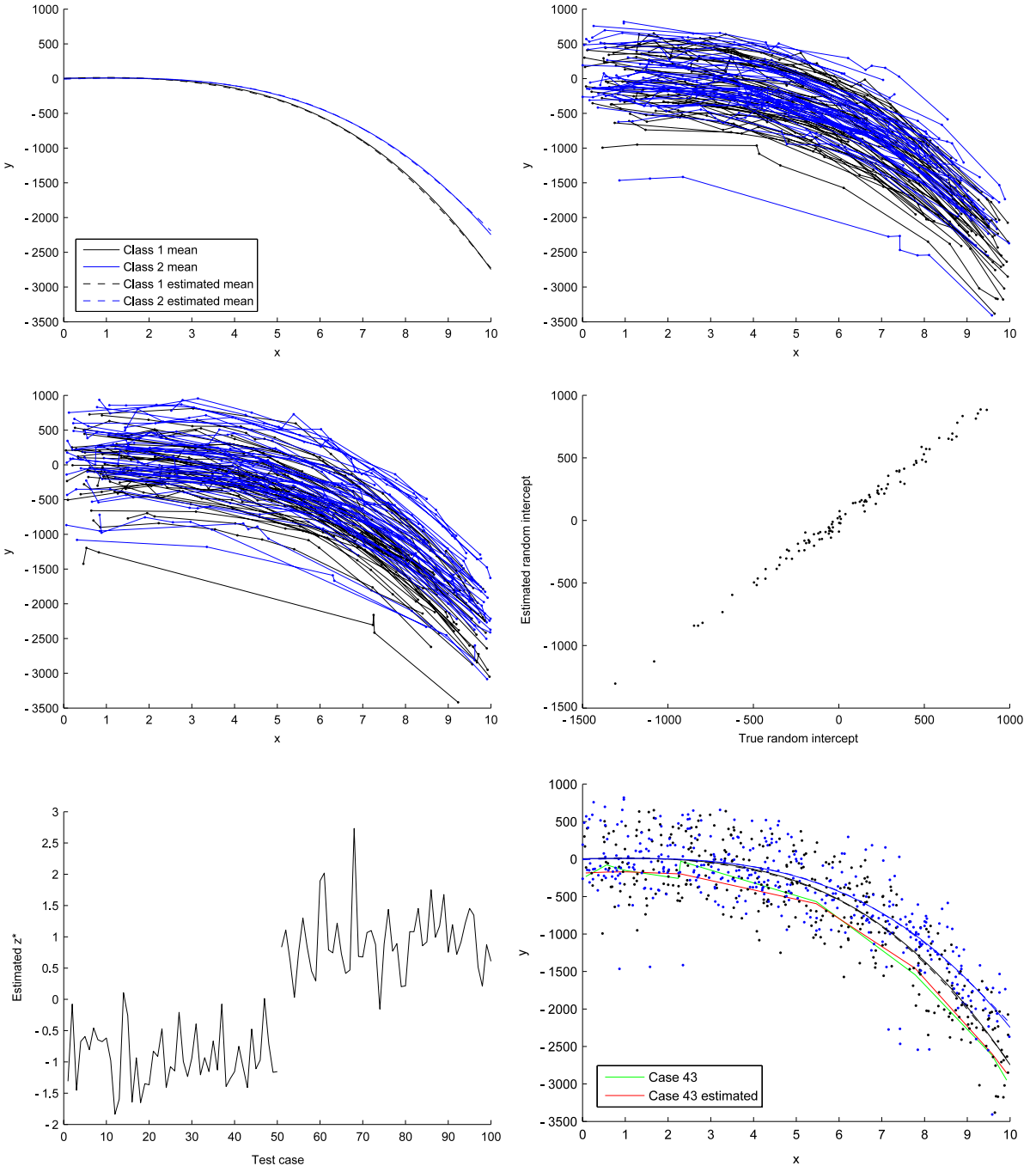

Fig. 2. Toy problem. Top left: simulated mean curves (solid) and estimated mean curves (dashed) for class 1 (black) and class 2 (blue). Top right: simulated training and validation data. Middle left: simulated test data. Middle right: simulated random effects versus estimated random effects. Bottom left: estimated z^*-values for test data. The first 50 cases belong to class 1 while the last 50 cases belong to class 2. Bottom right: case 43 (green) and after removing the residual error (red). Training and validation data, mean curves and estimated mean curves. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Empirical results showed that the classification approach in (15) produced identical predictions as compared to the alternative approach following (21). The combination of tuning parameters with the lowest (i.e. 2.19%) averaged balanced error rate (BER, i.e. the average of the class-specific error rates) on the validation data was: $\lambda_1 = 10^{-3}$, $\lambda_2 = 300$, $\sigma_1 = \sigma_2 = 8$. Fig. 2 illustrates that the estimated mean curves for class 1 and class 2 approximate the true mean curves, while the regularization parameters were in fact only optimized with respect to maximizing the BER. Next, the longitudinal

LS-SVM classifier was used to classify 50 new curves from class 1 and 50 new curves from class 2. The estimated random effects are close to the true random effects and the BER on the independent test set was 3%. Finally, Fig. 2 presents the estimated curve for case 43 after removing the residual error. The average over the squared residuals related to the prediction of y_{ij}^* (i.e. $\frac{1}{G}\sum_{i=1}^{G}\frac{1}{g_i}\sum_{j=1}^{g_i}(y_{ij}^* - \hat{B}_0 - \hat{B}_1\hat{z_i^*} - \hat{b_i^*} - (\frac{\hat{z_i^*}-1}{2})\hat{w_1}^T\varphi_1(x_{ij}^*) - (\frac{\hat{z_i^*}+1}{2})\hat{w_2}^T\varphi_2(x_{ij}^*))^2$) equaled 4477.5.

5.1.2. Sensitivity on tuning parameters

A training and validation data set with 50 curves for each class was generated according to the model in (5) to illustrate the influence of the tuning parameters of the longitudinal LS-SVM on the classification performance (Fig. 3). A simple variance structure, with a standard deviation of 60, was used for the normally distributed residual error and the standard deviation for the normally distributed random intercept was 600. The number of observations per curve varied between 5 and 7.

Parameter tuning based on 50 times random splitting resulted in the following settings: $\lambda_1 = 10^{-4}$, $\lambda_2 = 300$, $\sigma_1 = \sigma_2 = 9$. Fig. 3 visualizes the true mean curves and the estimated mean curves, using these settings, for class 1 and class 2. In order to study the sensitivity of the classifier's performance on the tuning parameters a grid was created around the optimum on the validation data: λ_1 ($\in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$), λ_2 ($\in \{80, 150, 200, 300, 400, 600\}$) and σ_1, σ_2 ($\in \{5, 7, 8, 9, 11, 13\}$). In a next step all these 180 combinations of tuning parameters were evaluated on 100 newly generated test data sets (e.g. Fig. 3, middle left panel) with 50 curves for each class, while training was done on the original training and validation data. The histogram of all 180 averaged (i.e. over the 100 repetitions) BERs is visualized in Fig. 3. The median BER was 19.50% while all 180 models fell in a range of 1.76%. This illustrates that the classifier's performance is rather insensitive with respect to (small) changes in tuning parameters. The bottom row of Fig. 3 presents the BER for different combinations of tuning parameters. Clearly, setting $\sigma_1 = \sigma_2 = 13$ or $\lambda_2 = 80$ results in larger BERs. Omitting these models from the histogram makes clear that all remaining models obtain BERs within a range of less than 1%.

5.1.3. Multi-class classification

The use of the multi-class approach from Section 3.1 is illustrated for a classification problem involving three classes (Fig. 4). Fifty curves were generated for each of the classes following the model in (23) with normally distributed residual error (with standard deviation 150) and random intercept (with standard deviation 600). The number of observations per curve varied between 5 and 7.

The set with the smallest (i.e. 8.96%) averaged error rate on the validation data based on 100 times random sampling was: $\lambda_1 = 2, \lambda_2 = 16, \sigma_1 = \sigma_2 = \sigma_3 = 2$. The upper left panel of Fig. 4 visualizes the estimated mean curves which approximate the true mean curves. The multi-class classifier was then used to classify 50 new curves from each class and resulted in an overall error rate of 5.33%. The average over the squared residuals related to the prediction of y_{ij}^* equaled 18308.2. The lower right panel of Fig. 4 shows $\frac{1}{\text{cost}_c}/(\frac{1}{\text{cost}_1} + \frac{1}{\text{cost}_2} + \frac{1}{\text{cost}_3})$ for each test case with $c = 1, \ldots, 3$. The first 50 test cases belong to class 1, the following 50 to class 2, etc.

5.1.4. Multiple random effects

To illustrate classification using models with multiple random effects 100 curves were generated for each of the classes. To simulate data normally distributed residual error with standard deviation 80 (i.e. uncorrelated errors) and three random effects (i.e. random intercept, random slope and random quadratic effect) with normal distribution and covariance matrix $[[400\ 10\ 5]^T[10\ 40\ 20]^T[5\ 20\ 20]^T]$ were assumed for the model in (32). The number of observations per curve varied between 5 and 10. Test data, consisting of 100 curves for each class, were simulated with similar settings. The top left panel of Fig. 5 visualizes the mean for each of the two classes, while the upper right panel visualizes the test data. The strategy in Algorithm 2 was followed for classifier building. First, a random intercept model was constructed using the training and validation data.

A gridsearch and repeated random sampling resulted in the following parameter settings: $\lambda_1 = 0.01$, $\lambda_2 = 80$, $\sigma_1 = \sigma_2 = 2$. The upper left panel of Fig. 5 demonstrates that the estimated mean curves approximated the true mean curves. After removing the estimated mean, restricted maximum likelihood was used on the training data to estimate the covariance matrix of the three random effects and residual error variance. Subsequently, a gridsearch, in combination with repeated random sampling, was used to tune the regularization constants: $\lambda_3 = 1600$ and $\lambda_4 = 950$. Finally, this classifier with three random effects was used on the independent test data and resulted in a BER of 9.5% (average over squared residuals for prediction of $y_{ij}^* = 5512.3$), while a LS-SVM with only a random intercept yielded a BER of 15.5% (average over squared residuals for prediction of $y_{ij}^* = 5643.0$). The cases that are misclassified by the LS-SVM with multiple random effects are visualized in the bottom panel of Fig. 5. It is clear that these remaining misclassifications are atypical profiles, which originate from the large overlap between both classes.

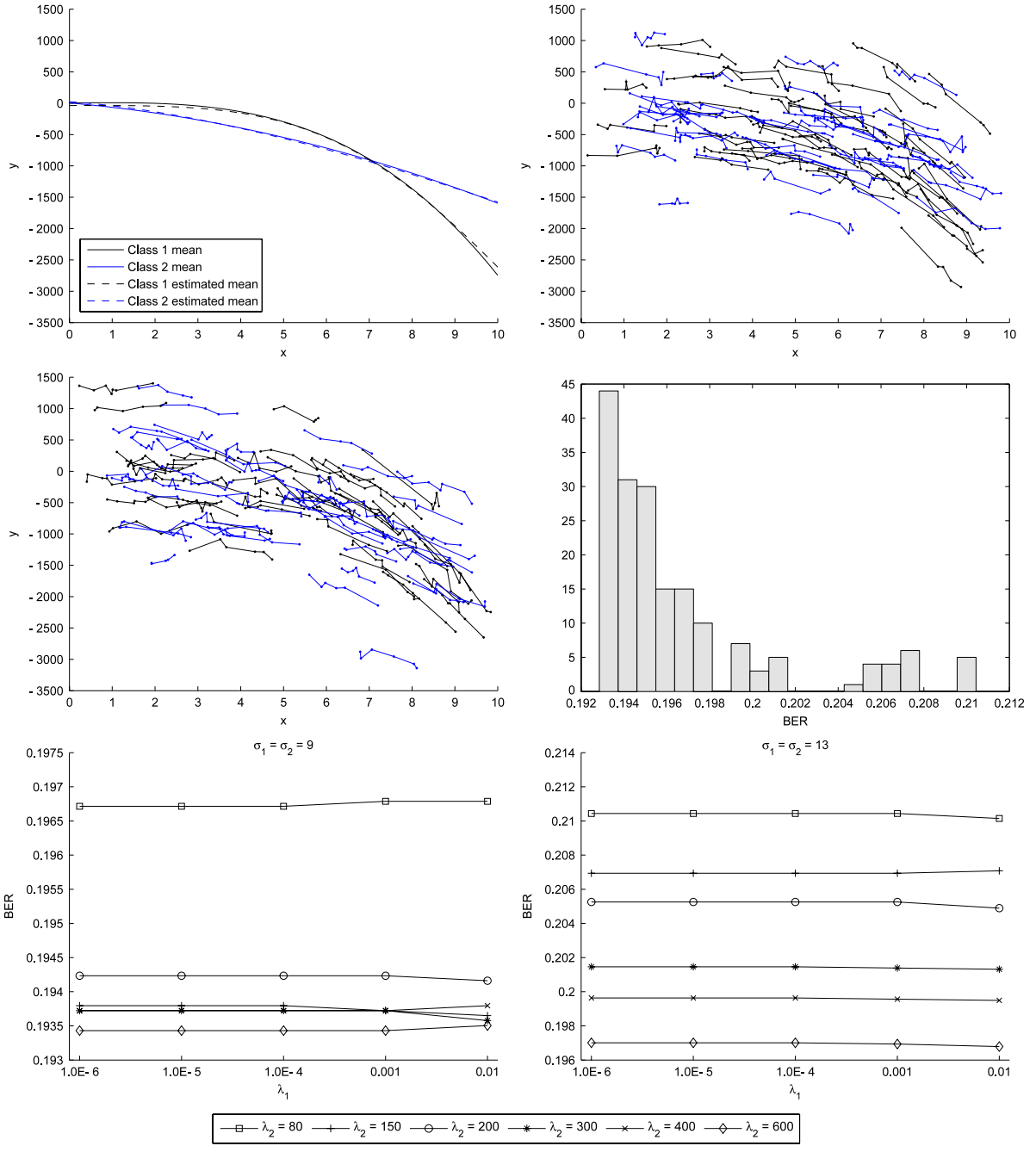

Fig. 3. Sensitivity analysis. Top left: simulated mean curves (solid) and estimated mean curves (dashed) for class 1 (black) and class 2 (blue). Top right: simulated training and validation data. Middle left: simulated test data. Middle right: histogram for averaged BER on test data. Bottom left: averaged BER on test data for $\sigma_1 = \sigma_2 = 9$. Bottom right: averaged BER on test data for $\sigma_1 = \sigma_2 = 13$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.2. Real-life data sets

5.2.1. Berkeley growth data

The Berkeley growth data set contains 54 growth curves of girls and 39 growth curves of males (Ramsay and Silverman, 2005). The classification problem is to predict the sex of a case based on the growth curve. Each original curve consists of 31 observations, taken at fixed time points. This experiment reports the performance of the random intercept LS-SVM classifier for various levels of unbalanced data (Fig. 6). Observations were deleted from each sample in a random manner in order to replicate unbalanced data. The percentage of unbalance was varied from 0 (i.e. 31 features) to 84 (i.e. 5 features).

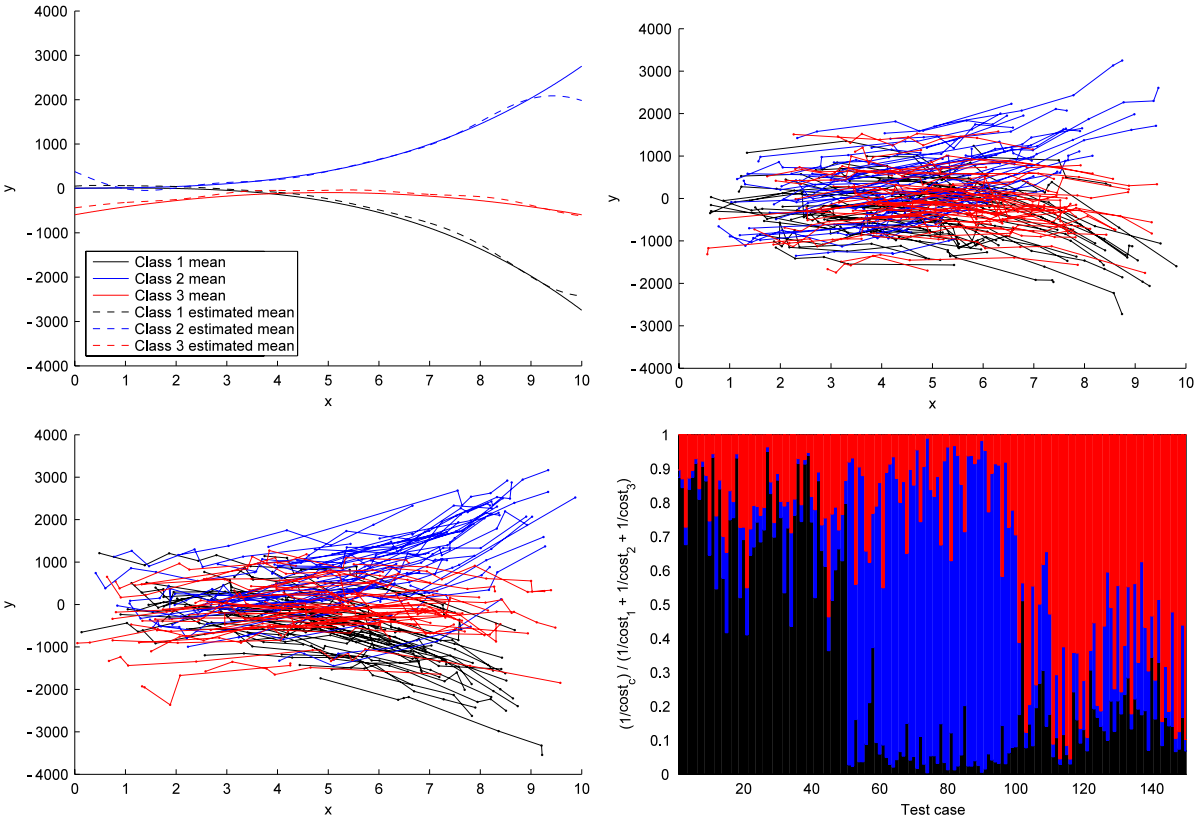

Fig. 4. Multi-class classification. Top left: simulated mean curves (solid) and estimated mean curves (dashed) for class 1 (black), class 2 (blue) and class 3 (red). Top right: simulated training and validation data. Bottom left: simulated test data. Bottom right: $\frac{1}{\text{cost}_c}/(\frac{1}{\text{cost}_1}+\frac{1}{\text{cost}_2}+\frac{1}{\text{cost}_3})$ for each test case with $c = 1, \ldots, 3$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The performance of FLDA is added for illustrative purposes. The original data were first randomly split into a set of 62 cases for training and validation and an independent test set of 31 cases. For nonzero percentages of unbalance the test data were 20 times randomly sampled and the averaged performance was computed.

Cubic splines were employed for FLDA and the number of knots q ($\in \{3, \ldots, 20\}$) and the rank constraint p ($\in \{1, \ldots, q+2\}$) were tuned. The resulting mean profiles, which are obtained by the longitudinal LS-SVM on a data set with 84% of unbalance, show the largest difference after the age of 13 (Fig. 6, upper right panel). The performances of FLDA and longitudinal LS-SVM are comparable: the BER for FLDA gradually decreases (with decreasing percentage of unbalanced data) from 39.78% to 6.62% while a decrease from 33.78% to 2.78% was obtained for the longitudinal LS-SVM.

5.2.2. Spinal bone mineral density data

The spinal bone mineral density data set consists of longitudinal profiles (i.e. 2, 3 or 4 measurements over a couple of years) from 153 females and 127 males (Bachrach et al., 1999). One of the aims of the study in Bachrach et al. (1999) was to examine gender differences in bone mineral acquisition, therefore, this variable was used as class variable in this experiment. The original data were randomly split into a set of 186 cases for model building and a separate, independent, set of 94 cases for model testing (Fig. 7). The performance of the random intercept LS-SVM classifier and FLDA were computed on the holdout data using the BER.

Cubic splines were used for the FLDA model and the number of knots q and the rank constraint p were tuned as in the previous example. The 186 cases were 60 times randomly split in a part for training (2/3 of the cases) and a part for validation (1/3 of the cases). The model selection procedure resulted in an FLDA model with 3 knots and the rank constraint was equal to 3, while the following parameters were obtained for the longitudinal LS-SVM classifier: $\lambda_1 = 0.5$, $\lambda_2 = 5$, $\sigma_1 = \sigma_2 = 9$. Fig. 7 presents the estimated mean profile for females as well as males. The longitudinal LS-SVM estimates show patterns that reach a plateau around age 16 for females and age 18 for males. The FLDA model and the LS-SVM resulted in an averaged BER of 31.69% and 29.54% on the validation data, respectively. Finally, the obtained classifiers were applied to the holdout data set. The FLDA model and the longitudinal LS-SVM classifier obtained a BER of 34.96% and 31.86% on the independent test set, respectively. The $\hat{z}_i^*$ values, which are obtained using the LS-SVM method, are visualized in Fig. 7. As expected, there is an increased classifier performance for cases with measurements at a lower age.

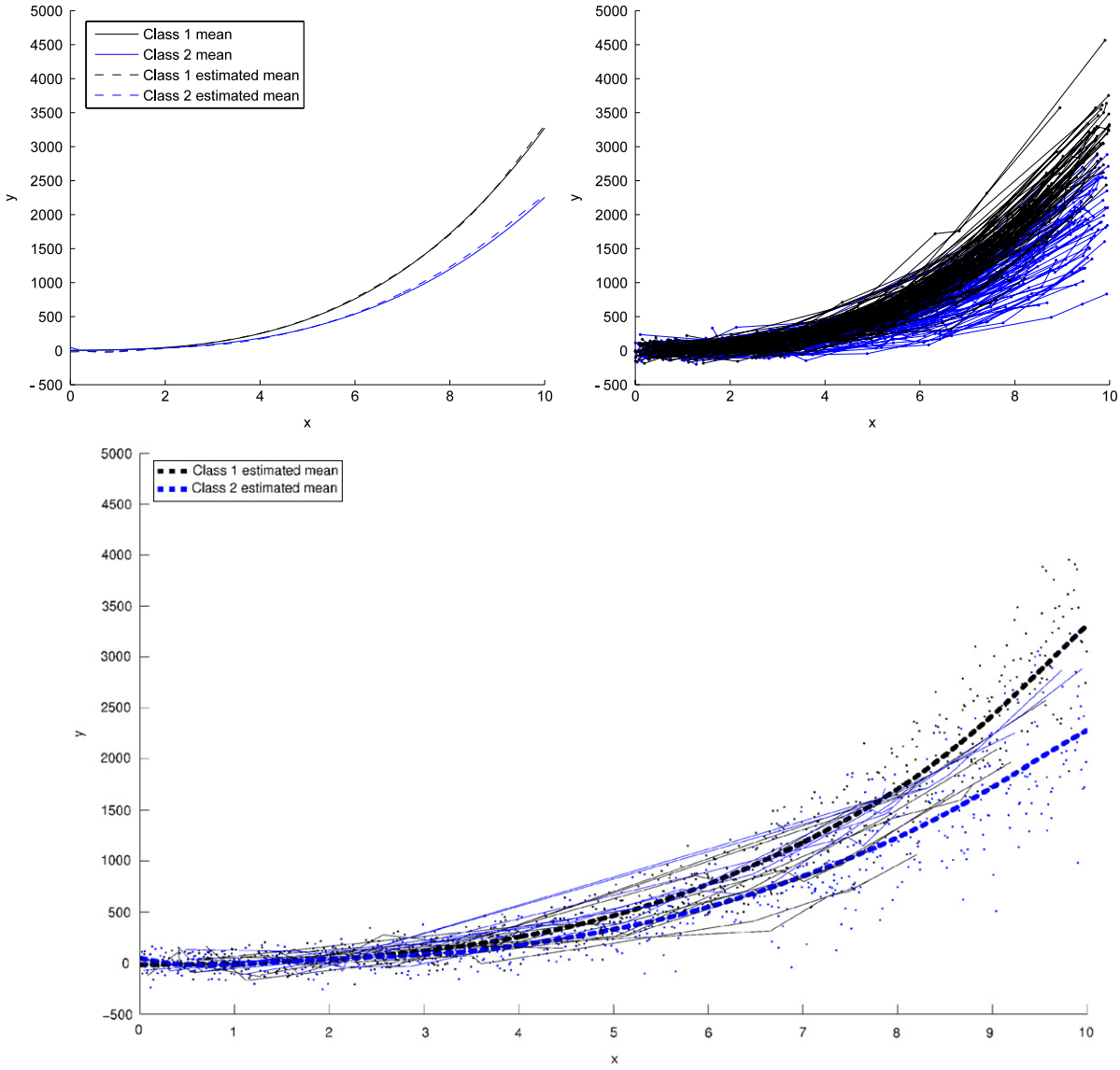

Fig. 5. Multiple random effects. Top left: simulated mean curves (solid) and estimated mean curves (dashed) for class 1 (black), class 2 (blue). Top right: simulated test data. Bottom: misclassified curves, estimated means (dashed) and training and validation data represented by dots. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6. Discussion

This study proposed a method to extend the least squares support vector machine (LS-SVM) classifier to a LS-SVM for classification of longitudinal profiles. Existing support vector machine-based classifiers are not able to directly deal with longitudinal profiles since these data are usually unbalanced and, therefore, a mixed effects approach was developed for the LS-SVM classifier. The proposed technique included a regression modeling and a prediction step. The method combined the regularization aspects of support vector machines with mixed effects models. Furthermore, the method incorporated nonlinear effects in longitudinal models via kernels while it enabled the kernel classifier to take within-subject correlation into account. It was demonstrated that the method addressed binary classification problems and an extension to multi-class classification was obtained in a straightforward way. Multiple random effects were included in the model formulation.

An important strength of the proposed methodology was that it provided a very flexible tool to model (non)linear relationships between different variables. Moreover, the method was computationally efficient since only a linear system needed to be solved. The proposed multi-class model formulation also allowed to use a different kernel for each class. This offers the possibility to perform more optimal class-specific modeling. In addition, the proposed formulation enables

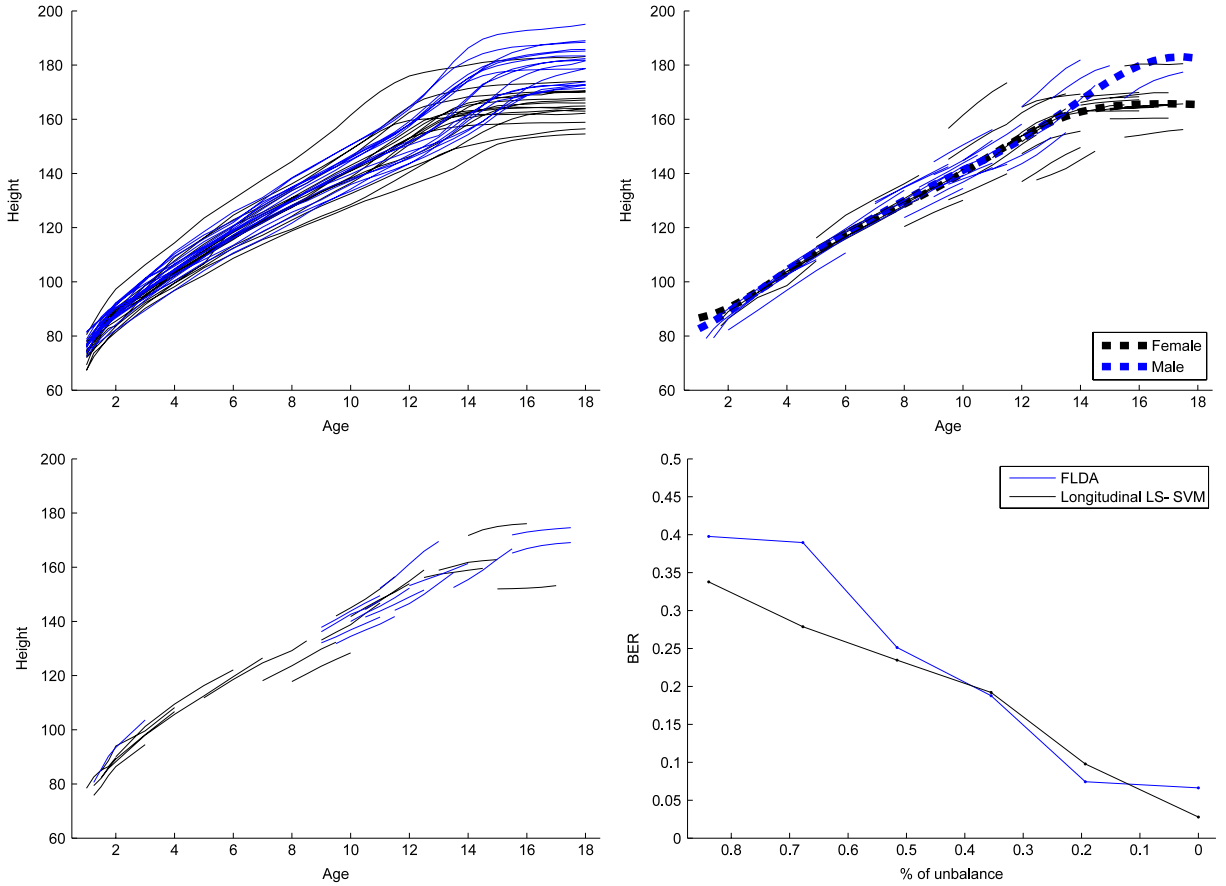

Fig. 6. Berkeley growth data. Top left: 15 growth curves of each class (0% of unbalance, female = black, male = blue). Top right: random training and validation data set with 84% of unbalance and estimated mean (dashed line) by longitudinal LS-SVM for females and males. Bottom left: random test data set with 84% of unbalance. Bottom right: BER for longitudinal LS-SVM and FLDA for various levels of unbalanced data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to specify heterogeneity in the covariance structure, i.e. class-dependent covariance structures. Another advantage of the proposed method was that, apart from the class prediction, a latent score for class membership could be obtained. Finally, the technique was applied to several simulated and real-life data sets concerning human growth, including a study to evaluate the sensitivity with respect to the choice of the regularization constants. The proposed methodology required tuning two regularization parameters, and possibly an additional kernel parameter. Usually cross-validation or repeated random sampling in combination with a gridsearch are used for this purpose, and these techniques require some computational time.

An interesting future perspective involves classification of large-scale data sets, i.e. data sets with large (e.g. $\geq$ 500,000) numbers of observations (n_i) or cases (N). Since the present approach operates in the dual space and thus solves a linear system of dimension proportional to the total number of observations M, it might become computationally intensive for such data sets. Therefore, future studies could focus on the use of approximation methods and incorporate them in the mixed effects LS-SVM classification methodology (De Brabanter et al., 2010; Espinoza et al., 2006). Furthermore, dealing with multivariate longitudinal data is a relevant topic as well. This type of data requires taking the additional correlation between various observed variables, measured at the same time point, into account. Finally, the incorporation of alternative loss functions represents an interesting research direction. For example, future studies will focus on the use of a binomial log-likelihood loss function (cf. logistic regression).

Acknowledgments

The authors would like to thank Laura K. Bachrach and Gareth M. James for making the spinal bone mineral density data available. Jan Luts is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO-Vlaanderen). The research is supported by the Research Council KUL: GOA-AMBioRICS, GOA-MaNet, several Ph.D./postdoc & fellow grants, Centers-of-excellence Optimization in Engineering (OPTEC), IDO 05/010 EEG-fMRI; Flemish Government: FWO: Ph.D./postdoc grants,

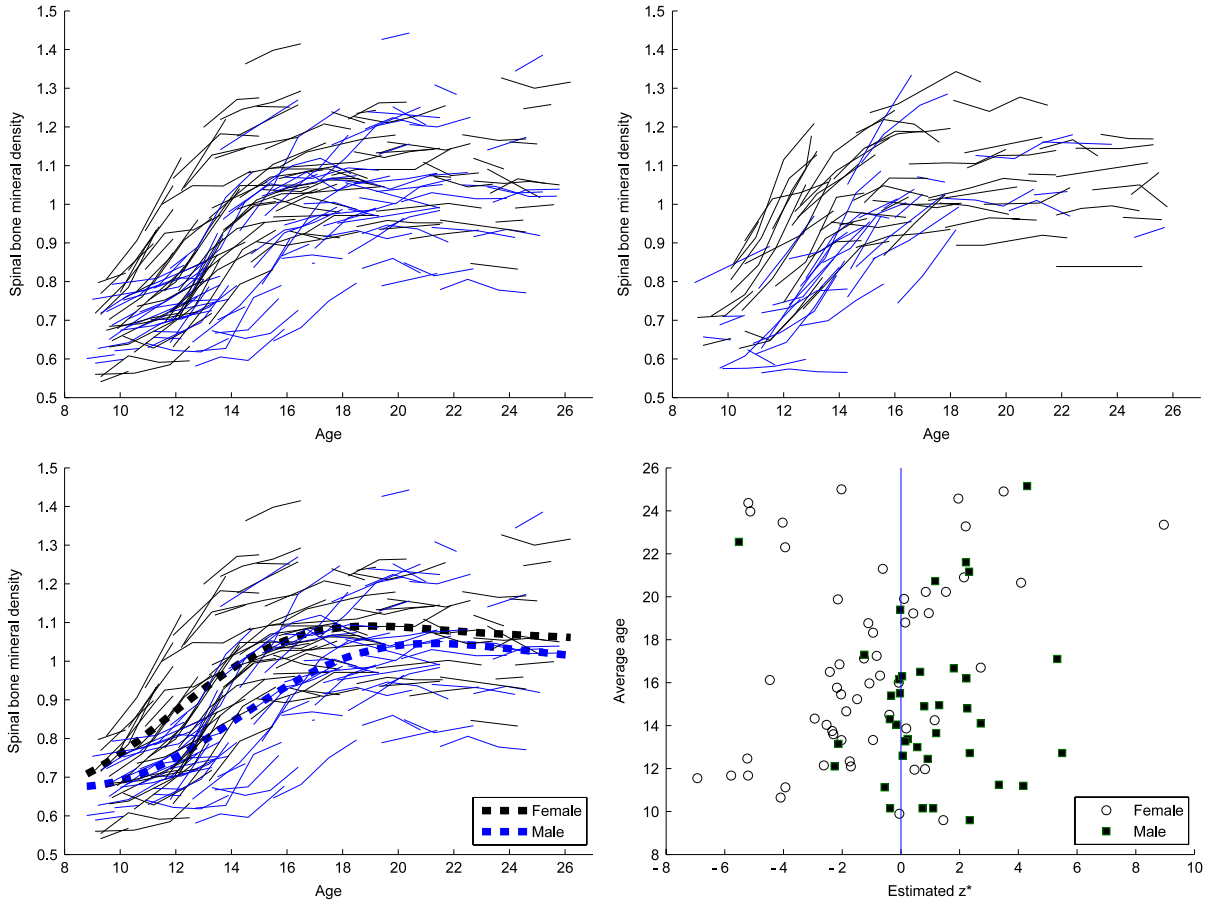

Fig. 7. Spinal bone mineral density data. Top left: training and validation data set (186 cases, female = black, male = blue). Top right: test data set (94 cases). Bottom left: estimated mean (dashed line) by longitudinal LS-SVM for females and males. Bottom right: $\hat{z}^*$ value versus average age for longitudinal LS-SVM on test data set. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

projects, G.0341.07 (Data fusion), G.0302.07 (Support vector machines and kernel methods), research communities (ICCoS, ANMMM); IWT: Ph.D. Grants; the Belgian Federal Government: IUAP P6/04 (Dynamical systems, control and optimization, 2007–2011); EU: FAST (contract no. FP6-019279-2).

References

Bachrach, L.K., Hastie, T., Wang, M.C., Narasimhan, B., Marcus, R., 1999. Bone mineral acquisition in healthy asian, hispanic, black, and caucasian youth: a longitudinal study. J. Clin. Endocrinol. Metab. 84 (12), 4702–4712.

Brown, P.J., Kenward, M.G., Bassett, E.E., 2001. Bayesian discrimination with longitudinal data. Biostatistics 2 (4), 417–432.

De Brabanter, K., De Brabanter, J., Suykens, J.A.K., De Moor, B., 2010. Optimized fixed-size kernel models for large data sets. Comput. Stat. Data Anal. 54 (6), 1484–1504.

De Brabanter, K., De Brabanter, J., Suykens, J.A.K., De Moor, B., 2011. Approximate confidence and prediction intervals for least squares support vector regression. IEEE Trans. Neural Netw. 22 (1), 110–120.

De la Cruz-Mesía, R., Quintana, F.A., Müller, P., 2007. Semiparametric Bayesian classification with longitudinal markers. J. R. Stat. Soc. Ser. C Appl. Stat. 56 (2), 119–137.

Espinoza, M., Suykens, J.A.K., De Moor, B., 2006. Fixed-size least squares support vector machines: a large scale application in electrical load forecasting. Comput. Manag. Sci. 3 (2), 113–129.

Fieuws, S., Verbeke, G., Maes, B., Vanrenterghem, Y., 2008. Predicting renal graft failure using multivariate longitudinal profiles. Biostatistics 9 (3), 419–431.

González-Manteiga, W., Vieu, P., 2007. Statistics for functional data. Comput. Stat. Data Anal. 51 (10), 4788–4792.

James, G.M., 2002. Generalized linear models with functional predictors. J. R. Stat. Soc. Series. B Stat. Methodol. 64 (3), 411–432.

James, G.M., Hastie, T.J., 2001. Functional linear discriminant analysis for irregularly sampled curves. J. R. Stat. Soc. Series. B Stat. Methodol. 63 (3), 533–550.

Liu, D., Lin, X., Ghosh, D., 2007. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 63 (4), 1079–1088.

Luts, J., Ojeda, F., Van de Plas, R., De Moor, B., Van Huffel, S., Suykens, J.A.K., 2010. A tutorial on support vector machine-based methods for classification problems in chemometrics. Anal. Chim. Acta 665 (2), 129–145.

Marshall, G., Barón, A.E., 2000. Linear discriminant models for unbalanced longitudinal data. Stat. Med. 19 (15), 1969–1981.

Müller, H.G., 2005. Functional modelling and classification of longitudinal data. Scand. J. Stat. 32 (2), 223–240.

Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58 (3), 545–554.

Pearce, N.D., Wand, M.P., 2009. Explicit connections between longitudinal data analysis and kernel machines. Electron. J. Stat. 3, 797–823.
Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer, New York.
Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects. Statist. Sci. 6 (1), 15–32.
Schölkopf, B., Smola, A.J., 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge.
Suykens, J.A.K., Alzate, C., Pelckmans, K., 2010. Primal and dual model representations in kernel-based learning. Stat. Surv. 4, 148–183.
Suykens, J.A.K., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9 (3), 293–300.
Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., 2002. Least Squares Support Vector Machines. World Scientific, Singapore.
Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.
Verbeke, G., Lesaffre, E., 1996. A linear mixed-effects model with heterogeneity in the random-effects population. J. Amer. Statist. Assoc. 91 (433), 217–221.
Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.
Wouters, K., Ahnaou, A., Cortinas Abrahantes, J., Molenberghs, G., Geys, H., Bijnens, L., Drinkenburg, W.H.I.M., 2007. Psychotropic drug classification based on sleep-wake behaviour of rats. J. R. Stat. Soc. Ser. C Appl. Stat. 56 (2), 223–234.