Proceedings of the 42nd IEEE
Conference on Decision and Control
Maui, Hawaii USA, December 2003

Least Squares Support Vector Machines and Primal Space Estimation

Marcelo Espinoza, Johan A.K. Suykens, Bart De Moor
K.U. Leuven, ESAT-SCD-SISTA
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
Tel. +32/16/32.17.09, Fax. +32/16/32.19.70
{marcelo.espinoza,johan.suykens}@esat.kuleuven.ac.be

***Abstract*— In this paper a methodology for estimation in kernel-induced feature spaces is presented, making a link between the primal-dual formulation of Least Squares Support Vector Machines (LS-SVM) and classical statistical inference techniques in order to perform linear regression in primal space. This is done by computing a finite dimensional approximation of the kernel-induced feature space mapping by using the Nyström technique in primal space. Additionally, the methodology can be applied for a fixed-size formulation using active selection of the support vectors with entropy maximization in order to obtain a sparse approximation. Examples for different cases show good results.**

Keywords: Least Squares Support Vector Machines, Nyström Approximation, Fixed-Size LS-SVM, Kernel Based Methods, Sparseness, Primal Space Regression.

I. Introduction

Kernel based estimation techniques, such as Support Vector Machines (SVMs) and Least Squares Support Vector Machines (LS-SVMs) have shown to be powerful nonlinear classification and regression methods [10], [16], [19]. Both techniques build a linear model where the inputs have been transformed by means of a (possibly infinite dimensional) nonlinear mapping φ. This is done in dual space by means of the Mercer's theorem, without computing explicitly the mapping φ. The SVM model solves a quadratic programming problem in dual space, obtaining a sparse solution [1]. The LS-SVM formulation, on the other hand, solves a linear system under a least squares cost function [12], where the sparseness property can be obtained by sequentially pruning the support value spectrum [13]. The SVM and LS-SVM dual formulations are quite advantageous when working with large dimensional input spaces, or when the dimension of the input space is larger than the sample size. The LS-SVM training procedure involves a selection of the kernel parameter and the regularization parameter of the cost function, that usually can be done by cross-validation or by using Bayesian techniques [9]. In this way, the solutions of the LS-SVM can be computed using an eventually infinite-dimensional φ based on a non-parametric estimation in the dual space.

However, the primal-dual structure of this problem can be exploited further. Working in the primal (or feature) space directly allow us to make use of traditional parametric methodologies for linear estimation. Not only it is possible to compute statistical properties of the parameters, inputs and predictions, but also it is a practical alternative when working with large datasets, where the dual estimation becomes less attractive as it requires the resolution of a linear system with as many unknowns as datapoints.

In this paper we propose a methodology for model estimation in primal space by using an explicit approximation of the kernel-induced nonlinear mapping φ. Based on the eigendecomposition of the kernel matrix and the use of Nyström techniques (as proposed within the framework of the fixed-size LS-SVM in [14]), the problem of multicollinearity is avoided and it is possible to perform a linear regression using Ordinary Least Squares (OLS) instead of ridge-regression.

This paper is organized as follows. Section II describes the general setting of the LS-SVM and the method for computing explicitly the nonlinear mapping through an eigenvalue decomposition of the kernel matrix. Classical inference for OLS is reviewed in Section III. The proposed use of the nonlinear variables as regressors under an OLS setting is presented and described in Section IV. Some examples are given in Section V.

II. Function Estimation using LS-SVM

The main elements of the primal-dual LS-SVM formulation are described, as well as the Nyström techniques available for an explicit approximation of the nonlinear mapping in the primal space.

A. Primal-Dual LS-SVM Formulation

The standard framework for LS-SVM estimation is based on a primal-dual formulation. Given the dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ the goal is to estimate a model of the form

$$y = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}) + b \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$ and $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_h}$ is the mapping to a high dimensional (and possibly infinite dimensional) feature space. The following optimization problem is formulated:

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

$$\text{s.t.} \quad y_i = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i, \quad i = 1, \ldots, N.$$

With the application of the Mercer's theorem on the kernel matrix $\boldsymbol{\Omega}$ as $\boldsymbol{\Omega}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{\varphi}(\boldsymbol{x}_j), i, j = 1, \ldots, N$ it is not required to compute explicitly the nonlinear mapping $\boldsymbol{\varphi}(\cdot)$ as this is done implicitly through the use of positive definite kernel functions K. For $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ there are usually the following choices: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T\boldsymbol{x}_j$ (linear kernel); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T\boldsymbol{x}_j/c + 1)^d$ (polynomial of degree d, with c a tuning parameter); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2/\sigma^2)$ (radial basis function, RBF), where σ is a tuning parameter.

From the Lagrangian $\mathcal{L}(\boldsymbol{w}, b, e; \alpha) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^N e_i^2 - \sum_{i=1}^N \alpha_i(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i - y_i)$, where $\alpha_i \in \mathbb{R}$ are the Lagrange multipliers, the conditions for optimality are given by:

$$\left\{\begin{array}{lcl} \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 & \rightarrow & \boldsymbol{w} = \sum_{i=1}^N \alpha_i \boldsymbol{\varphi}(x_i) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 & \rightarrow & \sum_{i=1}^N \alpha_i = 0 \\ \frac{\partial \mathcal{L}}{\partial e_i} = 0 & \rightarrow & \alpha_i = \gamma_i e_i, \qquad i = 1, \ldots, N \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 & \rightarrow & y_i = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i, \end{array}\right. \tag{3}$$

By elimination of $\boldsymbol{w}$ and e_i, the following linear system is obtained:

$$\left[\begin{array}{c|c} 0 & \mathbf{1}^T \\ \hline \mathbf{1} & \boldsymbol{\Omega} + \gamma^{-1}\mathbf{I} \end{array}\right]\left[\begin{array}{c} b \\ \hline \boldsymbol{\alpha} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \boldsymbol{y} \end{array}\right], \tag{4}$$

with $\boldsymbol{y} = [y_1, \ldots, y_N]^T, \boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$. The resulting LS-SVM model in dual space becomes

$$y(\boldsymbol{x}) = \sum_{i=1}^N \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b. \tag{5}$$

The ridge regression formulation is present in the cost function, and its regularization parameter γ avoids ill-conditioning due to possible multicollinearity among the n_h dimensions of $\boldsymbol{\varphi}$. Usually the training of the LS-SVM model involves an optimal selection of the tuning parameters σ (kernel parameter) and γ, which can be done using cross-validation techniques or Bayesian inference [9].

B. Nyström Approximation for Estimation in Primal Space

Explicit expressions for $\boldsymbol{\varphi}$ can be obtained by means of an eigenvalue decomposition of the kernel matrix $\boldsymbol{\Omega}$ with entries $K(\boldsymbol{x}, \boldsymbol{x}_j)$. Given the integral equation

$$\int K(\boldsymbol{x}, \boldsymbol{x}_j)\phi_i(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \lambda_i\phi_i(\boldsymbol{x}_j), \tag{6}$$

with solutions λ_i and ϕ_i, we can write

$$\boldsymbol{\varphi} = [\sqrt{\lambda_1}\phi_1, \sqrt{\lambda_2}\phi_2, \ldots, \sqrt{\lambda}_{n_h}\phi_{n_h}]. \tag{7}$$

Given the dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, it is possible to approximate the integral by a sample average [20], [21]. This will lead to the eigenvalue problem (Nyström approximation)

$$\frac{1}{N}\sum_{k=1}^N K(\boldsymbol{x}_k, \boldsymbol{x}_j)u_i(\boldsymbol{x}_k) = \lambda_i^{(s)}u_i(\boldsymbol{x}_j), \tag{8}$$

where the eigenvalues λ_i and eigenfunctions ϕ_i from the continuous problem can be approximated by the sample eigenvalues $\lambda_i^{(s)}$ and eigenvectors u_i as

$$\hat{\lambda}_i = \frac{1}{N}\lambda_i^{(s)}, \hat{\phi}_i = \sqrt{N}u_i. \tag{9}$$

Based on this approximation, it is possible to compute the eigendecomposition of the kernel matrix $\boldsymbol{\Omega}$ and use its eigenvalues and eigenvectors to compute the required components of $\hat{\boldsymbol{\varphi}}(\boldsymbol{x})$ simply by applying (7) if $\boldsymbol{x} \in \{\boldsymbol{x}_i, y_i\}_{i=1}^N$ (is a training point), or for any point $\boldsymbol{x}^{(v)}$ by means of

$$\boldsymbol{\varphi}(\boldsymbol{x}^{(v)}) = \frac{1}{\sqrt{\lambda_i^{(s)}}}\sum_{k=1}^N u_{ki}K(\boldsymbol{x}_k, \boldsymbol{x}^{(v)}). \tag{10}$$

This finite dimensional approximation $\hat{\boldsymbol{\varphi}}(\boldsymbol{x})$ can be used in the primal problem (2) to estimate $\boldsymbol{w}$ and b.

C. Sparse Approximations and Large Scale Problems

It is important to emphasize that the use of the entire training sample of size N to compute the approximation of $\boldsymbol{\varphi}$ will yield at most N components, each one of which can be computed by (9) for all $\boldsymbol{x} \in \{\boldsymbol{x}_i, y_i\}_{i=1}^N$. However, if we have a large scale problem, it has been motivated in [14] to use of a subsample of $M \ll N$ datapoints to compute the $\hat{\boldsymbol{\varphi}}$. In this case, up to M components will be computed, and their properties can heavily depend on the selection of the subsample of size M. External criteria such as entropy maximization can be applied for an optimal selection of the subsample. In this case, given a fixed-size M, the aim is to select the support vectors that maximize the quadratic Renyi entropy [14], [5]

$$H_R = -\log\int p(x)^2 dx \tag{11}$$

that can be approximated by

$$\int \hat{p}(x)^2 dx = \frac{1}{N^2}\mathbf{1}^T\boldsymbol{\Omega}\mathbf{1}. \tag{12}$$

The use of this active selection procedure can be very important for large scale problems. We will show an example of its performance on Section V.

It is interesting to note that the equation (8) is related to applying kernel PCA in feature space. However, in our case the conceptual aim is to obtain a finite approximation of the mapping $\boldsymbol{\varphi}$ on feature space as good as possible. If we use the entire sample of size N, then only equations (9) are to be computed and therefore the components of $\hat{\boldsymbol{\varphi}}$ are directly the eigenvectors of the kernel matrix $\boldsymbol{\Omega}$.

III. OLS Estimation in Primal Space

A basic review of the main conceptual elements for practical estimation in linear regressions is presented in this Section.

A. Review of Linear Regression

In order to make an explicit link with the previous section, a change of notation is proposed. Let

$$\boldsymbol{z}_k = \hat{\boldsymbol{\varphi}}(\boldsymbol{x}_k)$$

and consider the new $\boldsymbol{z}_k \in \mathbb{R}^m$ as inputs to the linear regression

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + b\mathbf{1} + \boldsymbol{\varepsilon} \tag{13}$$

with $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N]^T \in \mathbb{R}^{N\times 1}, \boldsymbol{y} = [y_1, y_2, \ldots, y_N]^T \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{Z} = [\boldsymbol{z}_1; \boldsymbol{z}_2; \boldsymbol{z}_3; \ldots \boldsymbol{z}_N] \in \mathbb{R}^{N\times m}$. The vector of coefficients $\boldsymbol{\beta}$ corresponds to the selected m components of the vector $\boldsymbol{w}$ from the LS-SVM initial setting (1) associated to the m selected components of the $\hat{\boldsymbol{\varphi}}$. For ease of notation, consider the matrix of full regressors $\boldsymbol{Z_F} = [\mathbf{1}, \boldsymbol{Z}]$, and the vector of full coefficients $\boldsymbol{\beta_F} = [b, \boldsymbol{\beta}]^T$. The regression can be written as:

$$\boldsymbol{y} = \boldsymbol{Z_F}\boldsymbol{\beta_F} + \boldsymbol{\varepsilon}. \tag{14}$$

Under the assumptions of normality and independence of the residuals ε_k ($\varepsilon_k \rightsquigarrow N(0, \sigma_0^2)$), and exogeneity of $\boldsymbol{z}_k$ ($E(\varepsilon_k|\boldsymbol{z}_k) = 0$), the following results can be obtained[1]:

- Unbiased estimates of $\boldsymbol{\beta}, b$, as follows:

$$[\hat{b}, \hat{\boldsymbol{\beta}}] = \hat{\boldsymbol{\beta}}_{\boldsymbol{F}} = (\boldsymbol{Z_F}^T \boldsymbol{Z_F})^{-1} \boldsymbol{Z_F}^T \boldsymbol{y} \tag{15}$$

  which is the OLS estimator for the regression parameters.
- Variance of the estimated parameters, as:

$$\text{Var}([\hat{b}, \hat{\boldsymbol{\beta}}]) = \text{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{F}}) = \sigma_0^2 (\boldsymbol{Z_F}^T \boldsymbol{Z_F})^{-1} \tag{16}$$

- Individual hypothesis testing for each coefficient in $\boldsymbol{\beta_F}$. Given the stochastic assumptions on the residuals, it can be written that

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{F}} \rightsquigarrow N(\boldsymbol{\beta_F}, \sigma_0^2 (\boldsymbol{Z_F}^T \boldsymbol{Z_F})^{-1}) \tag{17}$$

  which allows to build individual statistics for hypothesis testing. Given a hypothesis of the form:

$$\begin{aligned} H_0 &: \beta_i = C \qquad \text{Null hypothesis,} \\ H_1 &: \beta_i \neq C \qquad \text{Alternative hypothesis,} \end{aligned}$$

  we compute the t-statistics as

$$t_{\beta_i} = \frac{\hat{\beta}_i - C}{\sqrt{\text{Var}(\hat{\beta}_i)}} \rightsquigarrow T_{(n-m)}, \tag{18}$$

  where $T_{(n-m)}$ is the so-called Student's-t distribution with $n - m$ degrees of freedom. In particular, it is possible to test the hypothesis of significance, with a null hypothesis of $H_0 : \beta_j = 0$ against the alternative $H_0 : \beta_j \neq 0$. By replacing $C = 0$ in the expression (18), we obtain the t-statistics that will decide if the null hypothesis can be rejected.

B. Validity of the assumptions

The assumptions of normality and independence of the residuals, implying a constant variance, may not be valid in practice. In this case, the unbiasedness of the estimates still hold, but the inference may not be valid based on the equation (16) and the use of heteroskedastic-correction mechanisms or Generalized Least Squares estimation may be required. Particularly, the application of the t-tests is based on the distributional assumptions of the $\boldsymbol{\beta_F}$, that still hold, but its computation would require the application of alternative equations that go beyond the scope of this paper.

The orthogonality of the initial regressors in (14) makes it possible to select the *relevant* regressors in a single step, based on the computed t-statistics of each of the coefficients in $\hat{\boldsymbol{\beta}}_{\boldsymbol{F}}$. This is, we compute the OLS estimates for $\hat{\boldsymbol{\beta}}_{\boldsymbol{F}}$, compute its t-statistics using $C = 0$ in (18), and select as relevant all those variables for which its associated t-statistics satisfy

$$|t| > c_t \tag{19}$$

[1]These are the so-called Gauss Markov conditions [3].

where c_t is a threshold that can be related to different model selection criteria. For instance, under AIC, $c_t = \sqrt{2}$; under BIC, $c_t = \sqrt{\log N}$ [4]. The traditional significance-based tests yield $c_t = 1.64$ for a 90% significance, $c_t = 1.96$ for a 95%, or $c_t = 2.57$ for a 99% significance, based on the OLS assumptions. This can be seen as another way to impose regularization, as we can remove some regressors and thus reduce the number of parameters in the final model based on distributional assumptions.

Denoting by s the final number of selected regressors ($s \leq m$), the final model becomes

$$\boldsymbol{y} = \boldsymbol{Z_F}^* \boldsymbol{\beta_F}^* + \boldsymbol{\varepsilon}. \tag{20}$$

where $\boldsymbol{Z_F}^* \in \mathbb{R}^{N \times s}, \boldsymbol{\beta_F}^* \in \mathbb{R}^{s \times 1}$. The asterisk denotes that this is the final model after (eventually) removing non-significant regressors.

Although appealing from a parsimony point of view, we should remind that the computational form of the tests described so far is based on the validity of the OLS assumptions, particularly the one related with the independence of the residuals. As the focus of this paper is to move from the dual space formulation to linear regression in primal space, the validity of OLS assumptions are not questioned. However, working in the primal space makes it possible to use traditional techniques like Generalized Least Squares if the assumptions for OLS are no longer valid [7].

IV. Estimation in Primal Space: Implementation

In this section we describe the proposed methodology and its implementation, using the $\hat{\boldsymbol{\varphi}}$ as inputs to a linear regression in primal space. Specific distinctions can be applied for the case where we work with the entire training sample of size N, and for the case with a subsample of size M. A RBF kernel function of parameter σ will be used.

A. Description of the methodology

As discussed in Section II, it is possible to build an approximation of the nonlinear mapping $\boldsymbol{\varphi}$ using at most N components[2]. Here is where a major source of collinearity may arise, in the sense that the components of $\hat{\boldsymbol{\varphi}}$ computed from the eigenvectors associated with the smallest eigenvalues will have a numerical value close to zero, which can lead directly to ill-conditioning of the problem. Therefore, we will select only the *important* components of $\hat{\boldsymbol{\varphi}}$. This is, we will build up to $m < N$ components for $\hat{\boldsymbol{\varphi}}$, such that

$$\frac{\sum_{i=1}^{m} \lambda_i^{(s)}}{\sum_{i=1}^{N} \lambda_i^{(s)}} \geq c, \tag{21}$$

where $\lambda_k^{(s)}$ are the eigenvalues of the kernel matrix $\boldsymbol{\Omega}$ (in descending order, $\lambda_1^{(s)} > \lambda_2^{(s)} > \ldots$, etc) and c can be set arbitrarily close to one (c=0.90,0.95, 0.99). It is important to emphasize that the number of *relevant* components m will depend on the input data and it is not fixed beforehand. For

[2]This approximation has been studied recently in [15], with explicit bounds on the approximation accuracy of the eigenspectrum of the continuous problem.

instance, in the case of an RBF kernel, m will be influenced by the kernel parameter σ. Low values of σ will yield a large m, and vice versa.

Even with a subsample of size $M \ll N$, we can apply equation (21) to select m of them (and eventually it can be that we find $m = M$). Therefore, without loss of generality, for the remainder of the article we will refer to m as the number of selected eigenvalues, or equivalently, as the number of selected components of $\hat{\boldsymbol{\varphi}}$.

Therefore, by using the Nyström techniques it is possible to build a set of m orthogonal regressors $\hat{\boldsymbol{\varphi}}$ to be used in the linear regression (13) in order to perform the estimation in primal space. This model estimation can be done either using the entire sample to compute the initial $\hat{\boldsymbol{\varphi}}$, or using a subsample of size M in the spirit of the fixed-size LS-SVM. Probabilistic significance can be computed for each one of the m components, identifying those s components that are found to be statistically significant. In our case, we may use a $c_t = 1.6$ in (19) to ensure a reasonable probabilistic significance (larger than 90%).

The algorithm for each case can be described as follows.

1) Using the entire training sample (case M=N): We start with the estimation of the eigendecomposition of the kernel matrix $\boldsymbol{\Omega}$, and we approximate the $\boldsymbol{\varphi}$ using (7) and (9), selecting only m eigenvalues and eigenvectors according to (21) using c close to one. We use $c = 0.99$, and we see that no significant improvement of the model is obtained by increasing c closer to one. After having defined the set of m regressors, linear estimation is obtained by (14), obtaining an unbiased estimate of the $\boldsymbol{w}$ in (1). Significance tests can be computed from (18) using a predefined c_t, and eventually the model can be re-estimated after removing the non-significant components.

2) Using a subsample of size $M \ll N$: With an initial selection of a subsample of size M, active selection of this support vectors can be performed by entropy maximization. The initial eigendecomposition of $\boldsymbol{\Omega}$ is computed using only this M datapoints, selecting m eigenvectors according to (21). We are using the sparseness property which is explicit in SVMs and present in LS-SVMs, particularly suitable for large scale problems. We apply (7) and (9) for the datapoints in this subsample, and we approximate the components of $\boldsymbol{\varphi}$ for the remaining N datapoints using (10). In this way the set of regressors $\boldsymbol{Z_F} \in \mathbb{R}^{N \times m}$ is computed and the regression (14) is estimated. Again, it is possible to compute significance tests on the regressors, and eventually remove those non significantly different from zero.

B. Selection of the RBF tuning parameter

The value of the RBF kernel tuning parameter σ will influence directly the number of initial eigenvalues/eigenvectors m. Small values of σ will yield a large number of regressors m, and eventually it can lead to overfitting. By the contrary, a large value of σ can lead to a reduced number of regressors, making the model more parsimonious, but eventually not so accurate. In order to select an optimal kernel parameter, here we select the σ that minimizes the mean squared error (MSE) computed using 10-fold cross validation. This is the only tuning parameter the model requires, as there is no regularization term to be tuned as well.

V. Examples

Applications of the methodology presented so far are described in this Section, using 3 different examples. In order to make these examples as illustrative as possible, first we will apply the methodology for kernel parameter tuning and regression using the full sample, and also we will work with the case of a fixed-size subsample M. In this way, it is possible to verify the quality of the selection of support vectors and compare its performance with the case estimated using the full sample.

A. Description of the examples

Applications are implemented for the following cases:

1) *Function estimation*. This is a static nonlinear modeling problem, with unidimensional input $\boldsymbol{x}$ and noisy target values $y_k = \text{sinc}(2\pi x_k) + \epsilon_k$, with ϵ being a white noise of variance 0.1 and $x \in [-0.5, 0.5]$. The sample size is $N = 200$, and the subsample for the fixed-size application will be selected with size $M = 20$.
2) *Time Series forecasting*. This is the laser example of the Santa Fe competition [18] of time series prediction. Given 1000 historical datapoints, the goal is to predict the next 100 values using an iterative forecasting procedure. For our framework, we will fit a Nonlinear Auto-Regressive (NAR) [8], [11] model of the form $\hat{y}_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$, selecting $p = 50$. In our setting, the training sample is $N = 900$. The subsampling technique will be implemented with $M = 200$ datapoints. The out-of-sample iterative prediction will be computed for the next 100 values.
3) *Input-Output Model*. This is a example taken from the DaISy [2] datasets. The process is a liquid-satured steam heat exchanger, where water is heated by pressurized saturated steam through a copper tube. The output variable y_t is the outlet liquid temperature, and the input variable u_t is the liquid flow rate. We will fit a NARX model of the form $\hat{y}_t = f(y_{t-1}, \ldots, y_{t-p}, u_{t-1}, \ldots, u_{t-p})$ with $p = 5$, $N = 1800$, and $M = 200$ for subsampling. Out-of-sample predictions will be computed for the next 200 values. This is an example of a larger dataset, where working with the full sample N would require heavy computational cost. For comparison, the result of the same model under a linear estimation with the same p is reported.

It is important to emphasize that each one of these applications will be independently trained and estimated for the following cases:

- *Case I*. Using the full sample of size N to obtain the optimal hyperparameter, define the regressors and the final estimation.

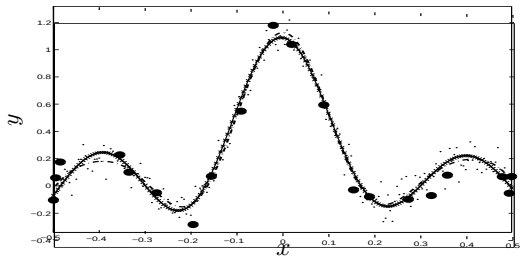

Fig. 1. Estimations for the noisy sinc function, for Case I ('-x' line) and Case II ('-.' line). The support vectors for Case II are depicted by the big dots.

- *Case II*. Using only a fixed-size subsample for finding the hyperparameter, the regressors and the final model.

The idea is to perform the training and estimation procedures independently, using only the available information for each Case. In other words, no information from Case I will be used in Case II, as for large scale problems the only feasible way to proceed would be using the subsampling method. The results reported on each case are:

1) The optimal σ found by minimizing the cross-validation MSE.
2) The value of m, the number of initial components selected for the regression in primal space. The number m comes as a result of using $c = 0.99$ in (21) for the initial selection of eigenvalues.
3) The value of s, the number of final selected regressors after estimation in primal space. The number s will come out as a result of using $c_t = 1.6$ in (19).
4) The MSE (mean squared error) both in-sample and out-of-sample.

B. Simulation Results

The results are summarized on Table 1 and accompanying figures. In general, we observe a sequence of $M > m > s$ with satisfactory results on the performance of the models. It is important to note that the good performance of the cases when $M \ll N$ is due not only to the quality of the Nyström approximation, but also on the good selection of the support vectors by means of the Renyi quadratic entropy maximization. For the sinc function example, the support vectors are quite uniformly distributed (as seen in Figure 1, together with the performance of the predictions for both Cases). In the Laser problem, it is remarkable the way the support vectors spread around the zones where important changes on the levels of the series are taking place. With this selection of the support vectors, the results obtained for the iterative prediction are very close to those obtained using the entire sample, as seen in Figures 2 and 3 respectively. For the same Laser example, Figure 4 shows the evolution of the entropy during the support vector selection for the Laser and the Heat-Exchanger examples. The performance of the methodology on the 200 predicted values for the Heat-Exchanger example is shown in Figure 5.

Problem	σ	M	m	s	MSE_{IN}	MSE_{OUT}
Sinc Function						
Case I	1.0	200	10	5	0.005	0.006
Case II	0.8	20	11	6	0.005	0.006
Laser (Santa Fe)						
Case I	5.3	900	228	168	0.01	0.05
Case II	4.2	200	192	130	0.02	0.06
Example 3						
Case I	5.8	1800	28	20	0.04	0.06
Case II	4.7	200	55	34	0.04	0.07
Linear (same p)	-	1800	-	-	0.04	0.23

TABLE I

PERFORMANCE OF THE ESTIMATIONS IN PRIMAL SPACE. CASE I MAKES USE OF THE FULL SAMPLE ($M = N$), AND CASE II MAKES USE OF A FIXED-SIZE ($M \ll N$) VERSION.

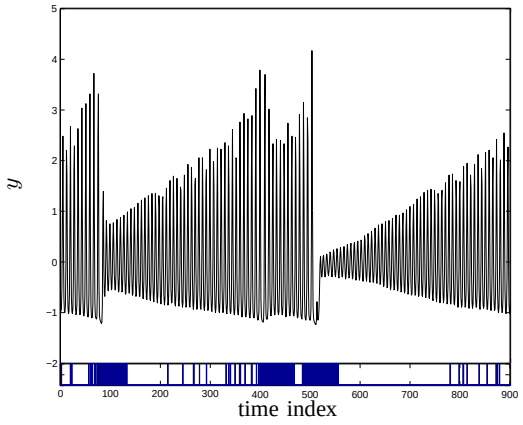

Fig. 2. Training sample for the Laser problem. Case I estimations make use of the full sample. The 200 selected support vectors can be visualized in terms of their time index position, indicated by the dark bars at the bottom.

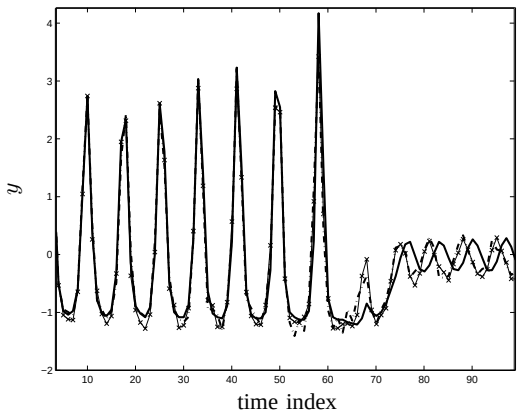

Fig. 3. Iterative prediction for the Laser example, for Case I ('-x' line), Case II ('-.' line), and 'true' values (full line).

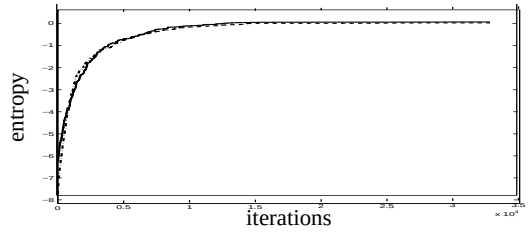

Fig. 4. Convergence of the Renyi entropy for the support vector selection in the Laser problem (full line) and the Heat Exchanger problem ('-.' line). Values have been normalized for comparison.

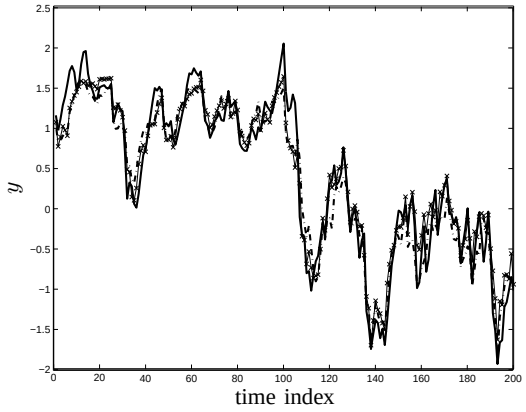

Fig. 5. Iterative prediction for the Heat-Exchanger example, for Case I ('-x' line), Case II ('-.' line), and 'true' values (full line).

VI. Conclusion

In this paper it has been shown how it is possible to make a link between LS-SVM in primal-dual formulation and traditional statistical inference techniques to perform linear regression in feature space using only significant components of the nonlinear mapping. This is done by computing an approximation of the nonlinear mapping induced by the kernel function K. The approximation allows to avoid the problem of possible multicollinearity, and therefore it is possible to use traditional OLS and statistical inference to obtain unbiased estimates of the coefficients of the linear specification.

The method was tested on some interesting examples with satisfactory results. Not only it can reproduce the results obtained with the LS-SVM in dual formulation and ridge regression, but it can also do it in a much faster way, especially when working with active selection of the support vectors for a fixed-size implementation. Although it is based on an eigendecomposition of the training kernel matrix, the required evaluations for the test sample can be computed easily by using the Nyström approximations only for those selected components of the eigenspectrum of the kernel matrix. Additionally, it is possible to apply further statistical techniques for model definition and identification in feature space, in order to improve the performance of the linear regression (with nonlinear regressors $\hat{\varphi}$).

Acknowledgements

This work was supported by grants and projects for the Research Council K.U.L (GOA-Mefisto 666, IDO, PhD/Postdocs & fellow grants), the Flemish Government (FWO: PhD/Postdocs grants, projects G.0240.99, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, ICCoS, ANMMM; AWI;IWT:PhD grants, Soft4s), the Belgian Federal Government (DWTC: IUAP IV-02, IUAP V-22; PODO-II CP/40), the EU(CAGE, ERNSI, Eureka 2063-Impact;Eureka 2419-FLiTE) and Contracts Research/Agreements (Data4s, Electrabel, Elia, LMS, IPCOS, VIB). J.Suykens is a Postdoctoral Researcher with the Fund for Scientific Research-Flanders (FWO-Vlaanderen) and a professor at the K.U.Leuven, Belgium. B. De Moor is a full professor at the K.U.Leuven, Belgium. The scientific responsibility is assumed by its authors.

VII. References

[1] Cristianini, N., and Shawe-Taylor, J., *An introduction to Support Vector Machines.* Cambridge University Press, 2000.

[2] De Moor B.L.R. (ed.) *DaISy: Database for the Identification of Systems*, Department of Electrical Engineering, ESAT-SCD-SISTA, K.U.Leuven, Belgium, URL: http://www.esat.kuleuven.ac.be/sista/daisy/, Feb-2003. Used dataset code:97-002.

[3] Davidson, R. and MacKinnon, J.G., *Estimation and Inference in Econometrics*. Oxford University Press, 1994.

[4] George, E. "The Variable Selection Problem," in Raftery E.,Tanner M. and Wells M. (Eds.) *Statistics in the 21st century*. Monographs on Statistics and Applied Probability 93. ASA. Chapman & Hall/CRC, 2003.

[5] Girolami, M. "Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem," *Neural Computation* 14(3), 669-688, 2003.

[6] Girosi, F. "An Equivalence Between Sparse Approximation and Support Vector Machines," *Neural Computation*, 10(6),1455-1480, 1998.

[7] Johnston, J. *Econometric Methods*. Third Edition, McGraw-Hill, 1991.

[8] Ljung, L. *Systems Identification: Theory for the User.* 2nd Edition, Prentice Hall, New Jersey, 1999.

[9] MacKay, D.J.C. "Probable Networks and Plausible Predictions - A Review of Practical Bayesian Methods for Supervised Neural Networks," *Networks: Computations in Neural Systems*,6, 469-505, 1995.

[10] Poggio, T. and Girosi, F. "Networks for Approximation and Learning," *Proceedings of the IEEE*, 78(9), 1481-1497, 1990.

[11] Sjöberg, J., Zhang, Q., Ljung, L,. Benveniste, A., Delyon, B., Glorennec, P.-Y., Hjalmarsson H., Juditsky, A., "Nonlinear Black-Box Modelling in Systems Identification: A Unified Overview," *Automatica* 31(12), 1691-1724, 1995.

[12] Suykens J.A.K. "Least Squares Support Vector Machines for Classification and Nonlinear Modelling," *Neural Networks World, Special Issue on PASE 2000.* 10, 29-48, 2002.

[13] Suykens J.A.K., De Brabanter J., Lukas, L. ,Vandewalle J., "Weighted Least Squares Support Vector Machines: Robustness and Sparse Approximation," *Neurocomputing*, Special issue on fundamental and information processing aspects of neurocomputing, vol. 48, no. 1-4, pp. 85-105, 2002.

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

[15] Shawe-Taylor, J. and Williams, C.K.I. "The Stability of Kernel Principal Components Analysis and its Relation to the Process Eigenspectrum," in *Advances in Neural Information Processing Systems* 15, MIT Press, 2003.

[16] Vapnik, V. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[17] Verbeek, M. *A Guide to Modern Econometrics.*Edison & Wesley, 2000.

[18] Weigend, A.S., and Gershenfeld, N.A. (Eds.) *Time Series Predictions: Forecasting the Future and Understanding the Past.* Addison-Wesley, 1994.

[19] Williams, C.K.I. "Prediction with Gaussian Processes: from Linear Regression to Linear Prediction and Beyond," in M.I. Jordan (Ed.), *Learning and Inference in Graphical Models*. Kluwer Academic Press, 1998.

[20] Williams, C.K.I and Seeger, M. "The effect of the Input Density Distribution on Kernel-Based Classifiers," in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*.

[21] Williams, C.K.I and Seeger, M. "Using the Nyström Method to Speed Up Kernel Machines," in T.Leen, T.Dietterich, V.Tresp (Eds.), *Proc. NIPS'2000*, vol 13., MIT press.