Fixed-Size Least Squares Support Vector Machines: A Large Scale Application in Electrical Load Forecasting

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

Based on the Nyström approximation and the primal-dual formulation of the Least Squares Support Vector Machines (LS-SVM), it becomes possible to apply a nonlinear model to a large scale regression problem. This is done by using a sparse approximation of the nonlinear mapping induced by the kernel matrix, with an active selection of support vectors based on quadratic Renyi entropy criteria. The methodology is applied to the case of load forecasting as an example of a real-life large scale problem in industry. The results are reported for different number of initial support vectors, which cover between 1% and 4% of the entire sample, with satisfactory results.

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

1 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 [21, 31, 34]. Both techniques build a linear model in the so-called feature space where the inputs have been transformed by means of a (possibly infinite dimensional) nonlinear mapping φ. This is converted to the dual space by means of the Mercer's theorem and the use of a positive

definite kernel, without computing explicitly the mapping φ. The SVM model solves a quadratic programming problem in dual space, obtaining a sparse solution [3]. The LS-SVM formulation, on the other hand, solves a linear system under a least squares cost function [28], where the sparseness property can be obtained by sequentially pruning the support value spectrum [29]. 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 [19]. 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. Primal-dual formulations in the LS-SVM context have been given also for kernel-Fisher's discriminant (kFD), kernel-Independent Component Analysis (kICA), kernel-Canonical Correlation Analysis (kCCA) [30].

Although the LS-SVM dual formulation is quite advantageous when working with large dimensional input spaces, or when the dimension of the input space is larger than the sample size, it requires the resolution of a linear system with as many unknowns as the number of datapoints, N. This situation has an obvious drawback when N is too large, and in such case the direct application of this method becomes prohibitive. However, the primal-dual structure of the LS-SVM can be exploited further. It is possible to compute an approximation of the nonlinear mapping φ in order to perform the estimations directly in primal space; furthermore, it is possible to compute a sparse approximation by using only a subsample of selected Support Vectors from the dataset. In this case, one can estimate a large scale nonlinear regression problem in primal space.

On the other hand, large databases available for data analysis are becoming quite frequent in industry and business, and therefore the algorithms and quantitative techniques should be able to handle large datasets in a proper way. Examples are common among banking, process industry, credit card transactions, etc. As an application to an interesting real-life problem, we study the case of the short-term load forecasting problem, which is an important area of quantitative research [22, 18, 2]. Within this context, the goal of the modelling task is to generate a model that can capture all the dynamics and interaction between possible explanatory variables to explain the behavior of the load in an hourly scale. For this task, there is a broad consensus about explanatory variables: past values of the load, weather information, calendar information, and possibly some past-errors correction mechanisms. Usually the load series show important seasonal patterns (yearly, weekly, intra-daily patterns) that need to be taken into account in the modelling strategy [14]. The short-term load forecasting is typically worked out in an hourly-basis, and there are usually long time series available. In our case, the data series

comes from a local low voltage substation in Belgium and it contains about 28,000 hourly values. It is important to note that many different modelling techniques and methodologies have been proposed within the context of load forecasting, some of them with excellent results. In this setting, the goal of this paper is to show the implementation of the large-scale LS-SVM methodology to this example.

We apply a methodology based on the eigendecomposition of the kernel matrix and the use of Nyström techniques (as proposed originally in Gaussian processes [36]) with an additional entropy selection method within the framework of the fixed-size LS-SVM [30, 7] for a practical (and interesting) industrial problem. This paper is structured as follows. The description of the LS-SVM is presented in Section II. In Section III, the methodology for working in primal space is described, with the particular application to a large scale problem. Section IV presents the problem and describes the setting for the estimation, and the results are reported in Section V.

2 Function Estimation using LS-SVM

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 $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n_h}$ 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:

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

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

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

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 e.g. cross-validation techniques or Bayesian inference [19].

3 Estimation in Primal Space

In this section, the estimation in primal space is described in terms of the Nyström approximation of the nonlinear mapping $\boldsymbol{\varphi}$, and the implementation for a large scale problem.

3.1 Nyström Approximation 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 for a variable $\boldsymbol{x}$ with probability density $p(\boldsymbol{x})$, 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 [35, 36]. 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 $i-$th required component of $\hat{\boldsymbol{\varphi}}(\boldsymbol{x})$ simply by applying (7) if $\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{i=1}^N$ (is a training point), or for any point $\boldsymbol{x}^{(v)}$ by means of

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

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

3.2 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 $\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{i=1}^N$. However, if we have a large scale problem, it has been motivated in [30] 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. The selection of subsample of size M, the initial support vectors, is done prior to the estimation of the model, and the final performance of the model can depend on the quality of the initial selection. It is possible to take a random selection of M datapoints and use them to build the approximation of the nonlinear mapping $\boldsymbol{\varphi}$, or it is possible to use a more optimal selection. 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 [30, 11]

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

that can be approximated by

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

The use of this active selection procedure can be quite important for large scale problems, as it is related to the underlying density distribution of the sample. In this sense, the optimality of this selection is related to the final accuracy that can be obtained in the modelling exercise. It is important to stress out that the difference between the performance of a model having an initial random selection and a model having an initial entropy-based selection will depend on the characteristics of the dataset itself. A rather "simple" dataset may be well approximated by both methods; whereas in a more "complex" dataset, the models can have different performances. Intuitively, the initial selection should contain some "important regions" of the dataset, as it was shown in [7] for the case of the Santa Fe Laser example [33].

It is interesting to note that the equation (8) is related to applying kernel PCA in feature space [23]. 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 $\mathbf{\Omega}$. In the application of this paper, it is required to define the number M prior to the modelling exercise. Each fixed-size sample will lead to an approximation of the nonlinear mapping for the entire sample of size N.

3.3 Estimation Technique

It has been motivated in [7] to only use the most significant m components from the M components computed as described in the previous subsection. Originally the M-sized kernel matrix $\mathbf{\Omega}$ leads to the computation of M components, and we select only the most important $m \leq M$ by looking at the most important m terms in the eigenspectrum. Near collinearity is thus avoided, and there should be no need for the regularization parameter γ as in the original dual formulation of the LS-SVM. Once the nonlinear mapping is computed, the estimation can be carried out by

Ordinary Least Squares (OLS). For a discussion about the use of a regularization term and its properties in linear regression, the reader is referred to [26, 27, 1].

In our case, the estimation in primal space goes as follows. First, we build up to $m \leq M$ components for $\hat{\boldsymbol{\varphi}}$, such that

$$\frac{\sum_{i=1}^{m} \lambda_i^{(s)}}{\sum_{i=1}^{M} \lambda_i^{(s)}} \geq c, \tag{13}$$

where $\lambda_k^{(s)}$ are the eigenvalues of the kernel matrix $\boldsymbol{\Omega}$ computed using the subsample of size M (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 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. With this selection of m components, we can build the $\hat{\boldsymbol{\varphi}}$ approximation.

Now, let us denote by $\boldsymbol{z}_k = \hat{\boldsymbol{\varphi}}(\boldsymbol{x}_k)$ and consider the new $\boldsymbol{z}_k \in \mathbb{R}^m$ as inputs to the linear regression

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + b\mathbf{1} + \boldsymbol{\varepsilon} \tag{14}$$

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 $\boldsymbol{Z} = [\boldsymbol{z}_1^T; \boldsymbol{z}_2^T; \boldsymbol{z}_3^T; \ldots; \boldsymbol{z}_N^T] \in \mathbb{R}^{N\times m}$. 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:

$$\boldsymbol{y} = \boldsymbol{Z_F}\boldsymbol{\beta_F} + \boldsymbol{\varepsilon}. \tag{15}$$

This final regression represents the model in primal space, and it this setting it is estimated by OLS. It is important to stress that this model is built on a sparse representation, starting from the selection of $M \ll N$ support vectors to compute the m components of $\hat{\boldsymbol{\varphi}}$. The sparseness of the representation underlying this methodology is what is makes it possible to tackle large scale problems.

4 Practical Example: Short-Term Load Forecasting

In this section the practical application is described, in term of the problem context, methodological issues and results.

4.1 Description and Objective

The modelling and forecasting of the load is currently an important area of quantitative research that is receiving increasing attention [22, 18, 2]. The main goal of the modelling task is to generate a model that can capture all the dynamics and interaction between possible explanatory variables for the load. When we talk about "short-term" in this context, it is usually referred to one hour ahead, up to one day ahead, and it is a task that is used on a daily basis on every major dispatch center or by grid managers [20]. For this task, there is a broad consensus about what could be the possible important explanatory variables: past values of the load, weather information, calendar information, and possibly some past-errors correction mechanisms. Usually the load series show an important seasonal pattern in different frequencies: a yearly pattern related to the winter-summer cycle; the weekly pattern, where Mondays are similar to Mondays, weekends to weekends, etc., and finally the intra-daily pattern, where the maximum and minimum load values in a day tend to keep the same "daily shape" every day.

The short-term load forecasting is typically worked out in an hourly-basis, and there are usually long time series available. For dealing with the complexity of the task, different approaches have been proposed: linear ARIMA, state-space, expert systems, neural networks, Box-Jenkins, etc. For a comprehensive review, see for instance [25]. On each of these models, the researcher has to carefully design the variables, model specification and training-validation procedures, leading to excellent results in many cases [8]. The objective of the modelling exercise for this paper is to show the application of the large-scale methodology described above, and we have selected the load modelling problem as an interesting example. Therefore, our focus is not to beat the current state-of-the-art results within the field, but to show an starting point from where on important improvements can be obtained.

4.2 Data and Methodology

The data to be used consists of a 3.5-years load series from a particular region in Belgium, making it a dataset with more than 28,000 datapoints. We define our training set to consist of 26,000 datapoints. The model formulation to be tested is a nonlinear ARX specification, with the following structure:

- An autoregressive part of 164 lagged load values, covering a complete week.
- Temperature-related variables measuring the effect of temperature on cooling

and heating requirements [6].

- Calendar information in the form of dummy variables for month of the year, day of the week and hour of the day [13].

This leads to a set of 211 explanatory variables. The procedure for modelling with the large-scale fixed-size LS-SVM methodology can be described in the following steps:

1. Normalize inputs and outputs to have zero mean and unit variance
2. Select an initial subsample of size M
3. Build the M-size kernel matrix and compute its eigendecomposition
4. Build the nonlinear mapping approximation for the rest of the data
5. Estimate a linear regression in primal space
6. Estimate the nonlinear mapping approximation for a test data
7. Use the regression estimates with the test data nonlinear mapping to produce the out of sample forecast

As mentioned before, in this paper the linear regression in primal space is performed using OLS, as we select only those significant components of the eigendecomposition of the kernel matrix $\mathbf{\Omega}$ [7]. It is possible to perform ridge regression [15] as well, as in the original LS-SVM in dual space, with the introduction of the ridge parameter γ.

In this example, tuning of the hyperparameter σ is performed by 10 fold cross validation in the training sample. We keep the value of σ that minimizes the out-of-sample mean squared error (MSE).

To illustrate the effect of increasing sizes of M, the above methodology is tested for sizes of $M = 200, 400, 600, 800$ and 1000 support vectors, selected with the quadratic entropy criterion. It is important to stress out that we are using between 1% and 4% of the dataset to build the nonlinear mapping for the entire sample. Values of M larger than 1000 are possible, as the only constraint in this approach is the computational time that will depend on the resources at hand.

To compare the results of the nonlinear models, we estimate a linear regression with the same set of initial inputs. After estimation, the inputs that are not statistically

Estimation	MSE (CV)	Selected components m
Linear	0.043	153 (selected regressors using $t-$statistics)
M=200	0.032	166
M=400	0.022	327
M=600	0.017	416
M=800	0.016	462
M=1000	0.015	496

Table 1: Performance of the LS-SVM fixed-size models with nonlinear mapping approximation built with M support vectors, on a crossvalidation basis using the optimal σ.

significant are removed on a stepwise selection process based on the $t-$statistics, yielding a final set of 153 regressors.

5 Results

In this section the results of the fixed-size LS-SVM methodology applied to the load modelling problem are reported, for the training procedure, the selection of support vectors and the out of sample performance.

5.1 Training Performance

The above procedure is applied for $M = 200, 400, 600, 800$ and $M = 1000$. Training using 10 fold crossvalidation is performed for each case, looking for an optimal value of the hyperparameter σ in the RBF kernel. After an initial exploratory search, the analysis is focused for the following range:

$$\sigma \in \{0.95, 1.11, 1.28, 1.49, 1.73, 2.01, 2.34, 2.72, 3.16, 3.67, 4.26, 4.95, 5.75, 6.69, 7.77, 9.03\}.$$

Figure 1 shows the evolution of the MSE in the 10 fold crossvalidation training procedure for the cases of $M = 200$ and $M = 400$, where it can be seen that the optimal value is $\sigma = 2.01$. For the cases $M = 600$, $M = 800$ and $M = 1000$ we perform the crossvalidation process using only the selected σ. The results for the computed MSE in a crossvalidation basis, and the equivalent result for the linear model, are shown in Table 1 using the selected σ. It is interesting to see the number of the components m that are sufficient for building the nonlinear mapping on each case, by setting $c = 0.99$ in expression (13).

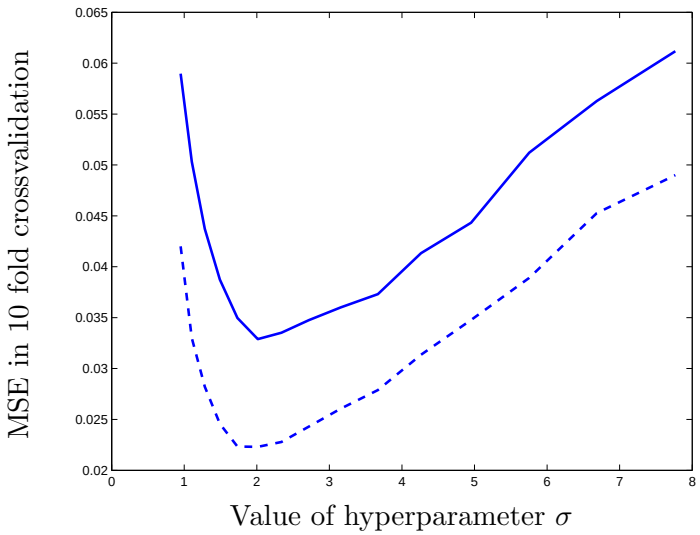


Figure 1: Performance evolution in the training procedure. The lines show the evolution of the MSE in a 10 fold crossvalidation for the cases $M = 200$ (full line) and $M = 400$ (dashed line). The optimal value for the σ hyperparameter is 2.01.

5.2 Support Vector Selection

The support vector selection has been applied using the quadratic entropy criterion. Starting from a random sample of size M, it is possible to replace elements of the selected sample by elements of the remaining sample if the entropy is minimized, and iterate this procedure until convergence. In this way, it is possible to obtain a selection of those M points that converge to a minimum value of the quadratic entropy. Figure 2 shows the evolution of the entropy value within this iterative process, for the $M = 400$ case. Figure 3 shows the position (time index) of the first element of the selected support vectors for the $M = 400$ case. It is interesting to see how the selected support vectors are those for which the output series is located in the regions of high load values (winter times), some in the lower values (summer times) and almost none of them in spring seasons. It is also clear that the output at the selected support vectors position is going through some "critical" regions.

It is possible to compare the performances for a model estimated with a random selection of support vectors against the same model estimated with a quadratic Renyi entropy selection starting from the same random selection. In other words, one can generate a random selection of support vectors and either perform a quadratic en-

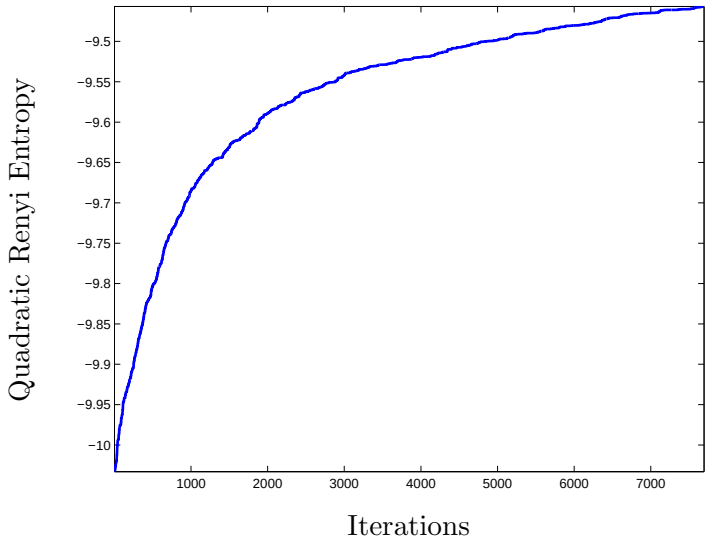


Figure 2: The evolution of the quadratic Renyi entropy within the iterative search for support vectors for the case $M = 400$

tropy selection using the random selection as initial position for the iterations (and later approximate the nonlinear model with the entropy-selected support vectors), or one can just use the random selection for the nonlinear mapping approximation immediately. Both models can be compared in terms of performance on the same test set. Table 2 and Figure 4 show the comparison for the results after 20 random initial selections, in which the model is either estimated after quadratic entropy selection taking the random selection as the initial starting point (Case I), or it is estimated directly (Case II). In all tests it has been used $M = 200$.

The existence of the standard deviation in Case I accounts for the fact that the

Support Vector Selection	average MSE	Standard Deviation MSE
Entropy-based Selection (Case I)	0.0311	0.0016
Random-based Selection (Case II)	0.0317	0.0025

Table 2: Comparison of the mean and standard deviation of the MSE for a test set performance using $M = 200$ over 20 randomizations. Case I refers to the random selection of support vectors and immediate estimation of the model. Case II starts from the same random selection, performs quadratic-based selection using the random sample as starting point, and then the model is estimated.

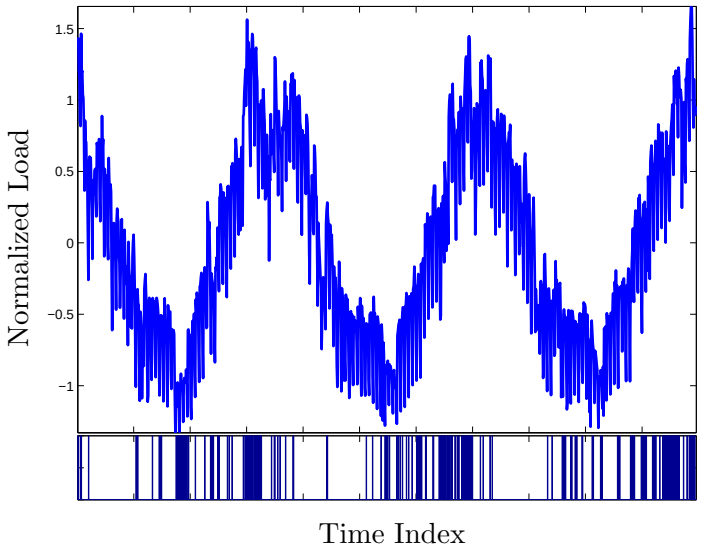


Figure 3: The normalized load used as training sample (Top), shown here only as daily averages rather than hourly values. The position of the selected support vectors corresponding to the load sample output is represented by dark bars at the bottom, showing the time index position of the first element of the support vector.

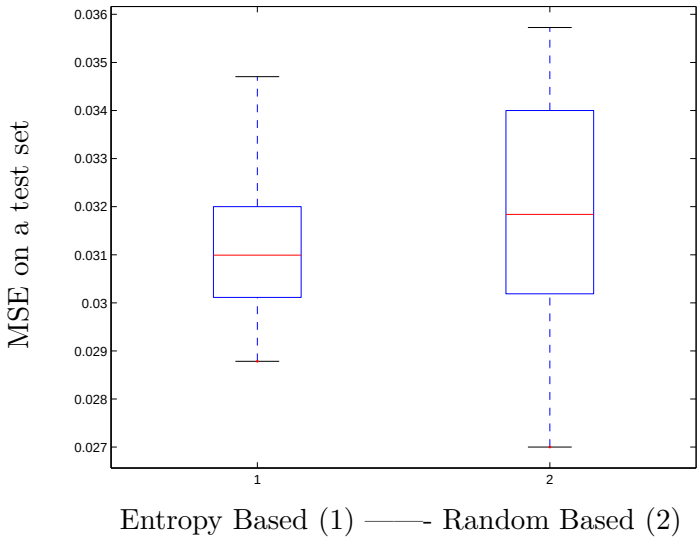


Figure 4: Box-plot of the MSE in a test set for models estimated with entropy-based (1) and random (2) selection of support vectors. Results for 20 repetitions.

convergence of the entropy selection is not unique, specially for a selection of 400 points out of 26,000 possible samples. However, starting from different random selections, the entropy-based selection yields lower dispersion in the errors. For this dataset, and after 20 repetitions, the average MSE are quite similar, but there is no guarantee that the random-selection will perform like this for a more complex dataset.

5.3 Out of Sample Performance

The estimated models are tested in a one-hour-ahead basis and a one-day-ahead prediction. In the first case, a simple one-step-ahead prediction is computed with the estimated model on each case, considering the next 30 days out of sample. No iterative predictions are required, as each hour is predicted based in the information needed to build the regressors and estimate its nonlinear mapping given the initial selection of support vectors. In the second case, an iterative prediction is performed, with a maximum of 24 hours ahead. In other words, it is the case where the operator wants to predict an entire day. Averages of the MSE for each case are reported on Table 3. An example of one-hour-ahead prediction for the case $M = 400$ and 1-day-ahead prediction for $M = 800$ are shown in Figure 5.

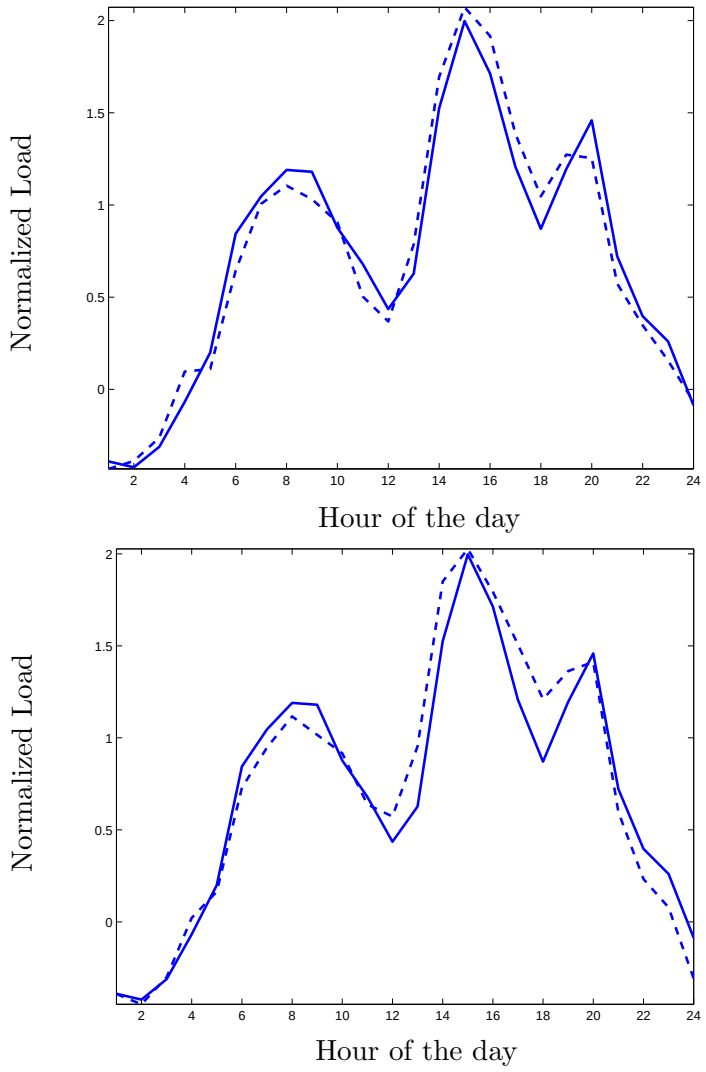


Figure 5: Out of sample performance examples for a selected 24-hours period. Top: Forecast on one-hour-ahead basis for the case $M = 400$. Bottom: Forecast on one-day-ahead basis for the case $M = 800$. Both forecasts are shown as dashed lines, and the actual values are shown as full line.

Model	MSE test (one-hour-ahead)	MSE test (one-day-ahead)
Linear	0.038	0.042
M=200	0.027	0.031
M=400	0.019	0.025
M=600	0.016	0.023
M=800	0.014	0.021
M=1000	0.014	0.021

Table 3: Comparison of the average MSE obtained in a test data for one-step-ahead (left) and 24-iterative predictions (right)

6 Conclusion

The methodology of fixed-size LS-SVM has been applied to the real-life problem of load forecasting. This paper has shown that it is possible to build a large scale nonlinear regression model from a dataset with 26,000 samples using different subsamples as support vectors, of sizes $M = 200, 400, 600, 800$ and $M = 1000$ with satisfactory results. The same nonlinear regression in the standard dual setting of the LS-SVM is hard to implement (not to say impossible) for this problem, as computing a square matrix of dimension 26,000 is prohibitive.

The results show that the nonlinear regressions in primal space improve their accuracy with the larger values of M. The maximum value of M to be used depends on the computational resources at hand, and it also depends on the underlying distributional properties of the dataset. In this context, quadratic entropy active selection of support vectors leads to performances which are less disperse as those obtained by random selection of support vectors.

Although the aim of the results of this paper is not to beat the current state of the art models developed within the load modelling and forecasting results, it is a starting point to further improve the results of the nonlinear regression. Further development in terms of input selection from a nonlinear perspective, and the adoption of ad-hoc modelling strategies for the load problem (such as modelling separate equations per hours of the day, handling of the seasonalities in a stochastic rather than deterministic manner) should lead to significant improvements.

Acknowledgments

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 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.

References

[1] Björkstrom, A. and Sundberg, R. "A Generalized View on Continuum Regression," *Scand. Journal of Statistics* 26,17-30, 1999.

[2] Bunn, D. "Forecasting Load and Prices in Competitive Power Markets,", Invited Paper, *Proceedings of the IEEE*, Vol 88, No.2, 2000.

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

[4] 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.

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

[6] Engle, R., Granger, C.J., Rice, J., and Weiss, A. "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*, Vol.81, No.394, 310-320, 1986.

[7] Espinoza, M., Suykens, J.A.K. and De Moor B.L.R. "Least Squares Support Vector Machines and Primal Space Estimation," Accepted for publication in *IEEE 42nd Conference on Decision and Control*, Maui, USA, Dec. 9-12, 2003.

[8] Fay, D., Ringwood, J., Condon, M. and Kelly, M. "24-h Electrical Load Data-A Sequential or Partitioned Time Series?," *Neurocomputing* 55, 469-498, 2003.

[9] Frank, I and Friedman, J. "A Statistical View of Some Chemometrics Regression Tools," *Technometrics* 35, 109-148, 1993.

[10] 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.

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

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

[13] Guthrie, G. and Videbeck, S. "High Frequency Electricity Spot Price Dynamics: An Intra-Day Markets Approach", New Zealand Institute for the Study of Competition and Regulation, 2002.

[14] Hylleberg, S. *Modelling Seasonality*. Oxford University Press, 1992.

[15] Hoerl, A.E. and Kennard, R.W. "Ridge Regression: Biased Estimation for Non Orthogonal Problems," *Technometrics* 8, 27-51, 1970.

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

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

[18] Lotufo, A.D.P. and Minussi, C.R. "Electric Power Systems Load Forecasting: A Survey," *IEEE Power Tech Conference*, Budapest, Hungary, 1999.

[19] 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.

[20] Mariani, E. and Murthy, S.S. *Advanced Load Dispatch for Power Systems*. Advances in Industrial Control, Springer-Verlag, 1997.

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

[22] Ramanathan, R., Engle, R., Granger, C.W.J., Vahid-Aragui, F., Brace, C. "Short-run Forecasts of Electricity Load and Peaks," *International Journal of Forecasting* 13, 161-174, 1997.

[23] 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.

[24] 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.

[25] Steinherz, H., Pedreira and Castro, R. "Neural Networks for Short-Term Load Forecasting: A Review and Evaluation",*IEEE Transactions on Power Systems*, Vol.16. No.1,2001.

[26] Stone, M. and Brooks, R.J. "Continuum Regression: Cross-Validated Sequentially Contructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression," *J. R. Statist. Soc. B*, 52, 237-269, 1990.

[27] Sundberg, R. "Continuum Regression and Ridge Regression," *J. R. Statist. Soc. B*, 55, 653-659, 1993.

[28] Suykens, J.A.K., Vandewalle, J. "Least Squares Support Vector Machines Classifiers," *Neural Processing Letters* 9, 293-300, (1999).

[29] 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.

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

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

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

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

[34] 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.

[35] 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).*

[36] 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.