# Sparse kernel approximations for estimation of waste-heat recovery in ships

Mikael Manngård*, Jari M.Böling

*Process Control Laboratory, Faculty of Science and Engineering, Åbo Akademi University.*
*Biskopsgatan 8, FIN-20500 Åbo, Finland*

**Abstract**

Data-driven, kernel-based learning methods have been of increasing interest lately, and have shown great impact in areas such as machine learning, computer vision and pattern recognition. However, the rapidly increasing amount of available data poses new challenges that traditional methods struggle with. In this paper we present computationally tractable method for kernel regression on large data sets. The size of the kernel matrix grows rapidly as the number of data points increases, and computing and storing it may even become intractable. Hence, we have proposed a criteria based on linear independence between feature vectors for selecting the relevant input data points for building the kernel matrix. The proposed method do not require storing the full kernel matrix in memory and will result in a sparse kernel approximation. The proposed method has been applied in a case study for estimating the recovered waste-heat energy in ships based on engine load data.

*Keywords:* Kernel methods, sparse optimization, big data

## 1. Introduction

Kernel methods that use positive-definite kernel functions are used to map input data into a higher dimensional feature space, and can hence be used for non-linear modelling. A kernel is function $k : X \times X \to \mathbb{R}$ where $X \subseteq \mathbb{R}^n$. Furthermore, for a sequence of input vectors $\{x_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^n$, James Mercer showed that for any symmetric, positive semidefinite kernel matrix $K := \left(k(x_i, x_j)\right)_{i,j=1}^N$, there exist a feature map $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j) \tag{1}$$

(Cristianini and Shawe-Taylor, 2000). Inner products between feature vectors appear naturally in many problems, and kernel functions can hence sometimes be introduced to simplify them. This is often referred to as the 'kernel trick'. However, for large data sets, the kernel matrix will become very large, and there is a need for methods to reduce the size of the kernel matrix. Various sparse optimization techniques have frequently been used to find sparse representations of the kernel matrix (Bishop, 2006). However, for very large data sets, such methods will start to fail. Hence there is a need for approximative methods for reducing the kernel size that do not require storing the full kernel matrix in memory.

*Corresponding author.
Research supported by the Finnish Graduate School in Chemical Engineering (GSCE), and by the Efficient Energy Use (EFEU) research program coordinated by CLEEN Ltd..
*Email addresses:* mikael.manngard@abo.fi (Mikael Manngård), jari.boling@abo.fi (Jari M.Böling)

## 2. Problem formulation

Given a sequence of $N$ input-output pairs of data $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, and a feature mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^m$, we form the following linear regression model

$$y_i = w^\top \varphi(x_i) + b + e_i, \quad i = 1, ..., N \tag{2}$$

where $e_i$ is assumed to be Gaussian white noise, $b$ is a scalar bias and $w \in \mathbb{R}^m$. We formulate the following regularized least squares problem

$$\begin{aligned} \underset{w,b,e}{\text{minimize}} \quad & f(x) = \frac{1}{2}\gamma\|e\|_2^2 + \frac{1}{2}\|w\|_2^2 \\ \text{subject to} \quad & e_i = y_i - w^\top \varphi(x_i) - b, \qquad i = 1, ..., N, \end{aligned}$$

where $\gamma$ is a scalar weighting factor. By introducing the Lagrangian function and optimality conditions, the problem can be reformulated as a system of linear equation

$$\begin{bmatrix} 0 & \bar{1}_N^\top \\ \bar{1}_N & K + \gamma^{-1}\bar{1}_N \end{bmatrix} \begin{bmatrix} b^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{3}$$

where $b^*$ and $\lambda^*$ denote the optimal bias and Lagrange multipliers respectively and the kernel matrix $K_{i,j} = k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j)$ for $i, j = 1, ..., N$. This formulation is referred to as the least squares support vector machine (LS-SVM) (Suykens et al., 2002). Note that the problem only require computations of inner products $k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j)$, which for a properly chosen kernel, do not depend on the dimensionality of the parameter vector $w$, but only on the number of observed data points $N$. Although Equation 3 can be solved explicitly by matrix inversion or by newton iterations, computing the Kernel matrix is of size $N \times N$ may become intractable large number of data points. Hence, we propose a pre-processing method for reducing the data set and the size of the kernel matrix.

## 3. Sparse kernel approximation

Engel et al. (2004) proposed a sequential sparsification procedure for the kernel recursive least squares problem, which only admits data points that form feature vectors that are not approximately linearly dependent (ALD). In other words, given a dictionary $\mathcal{D}_{k-1} = \{x_i\}_{i\in I_{k-1}} \subseteq \{x_i\}_{i=1}^{N}$, where $I_k \subseteq \{1, ..., k\}$ is an index set of included data points at iteration step $k$. A new data point $x_k$ is appended to the dictionary if the the set of feature vectors $\{\varphi(x_i)\}_{i\in I_{k-1}}$ are approximatively linearly independent of $\{\varphi(x_k)\}$. This gives the following ALD test

$$\delta_k := \min_a \|a^\top \varphi(\tilde{x}_{k-1}) - \varphi(x_k)\|_2^2 \leq \eta, \tag{4}$$

where $\tilde{x}_k = \{x : x_i \in \mathcal{D}_k\}$ and $\eta \geq 0$ is a pre-determined accuracy parameter chosen close to zero. For $\eta = 0$, exact linear independence would be required. If $\delta_k > \eta$, the data point $x_k$ is included in the dictionary, i.e. $\mathcal{D}_k = \mathcal{D}_{k-1} \cup \{x_k\}$. Expanding Equation 4 we get

$$\begin{aligned}\delta_k = \min_a\ & a^\top \varphi(\tilde{x}_{k-1})^\top \varphi(\tilde{x}_{k-1}) a - 2a\varphi(\tilde{x}_{k-1})\varphi(x_k) \\ & + \varphi(x_k)^\top \varphi(x_k) = \\ \min_a\ & a^\top \tilde{K}_{k-1} a - 2ak(\tilde{x}_{k-1}, x_k) + k(x_k, x_k).\end{aligned} \tag{5}$$

Solving the minimization problem in Equation 5, we get the following ALD conditions

$$a_k = \tilde{K}_{k-1}^{-1} k(\tilde{x}_{k-1}, x_k), \tag{6}$$

$$\delta_k = k(x_k, x_k) - k(\tilde{x}_{k-1}, x_k) a_k. \tag{7}$$

Although the the ALD criteria were originally presented for solving kernel regression problems online, we have applied it as an effective pre-processing strategy for selecting relevant data points.

## 4. Case study: Estimation of waste-heat recovery in ships

The presented methods have been applied to a case study for estimation of waste-heat recovery in ships. Data have been gathered from a passenger cruise ship, powered by four diesel engines with a combined nominal power of 48 MW. Engine temperatures are controlled via a fresh-water cooling system, from which excess heat can be extracted trough heat exchangers. The purpose of this case study is to identify a model that describes the relation between engine electrical-load and recovered waste heat based only on measured data. Almost 650 hours of data were collected as 60 second mean measurements, which makes a total of $N = 39\,000$ data samples. Standard regression methods will start to fail for such large amounts of data, and computing and storing the kernel matrix will exceed the capacity of most computers. Hence, the proposed ALD method was used to reduce the kernel size.

The data were split into two parts, so that 2/3 of the data are training the model, and the rest are used for validation. A Gaussian radial basis kernel function $k(x_i, x_j) = \exp\left(-\|x_i - x_j\|_2^2/2\sigma^2\right)$ with $\sigma = 0.5$ was used. The ALD threshold $\eta$ can be determined by cross-validation, but was here set arbitrarily to $\eta = 10^{-3}$. A reduced kernel of size $82 \times 82$ was obtained after applying the ALD selection criteria. The ALD steps took less than 5 seconds to perform in Matlab 2014b on a 3.4 GHz Intel i7-4770 CPU, and solving Equation 3 for the reduced kernel took less than one second. In Figure 1, the fitted model output is compared to both training and validation data.

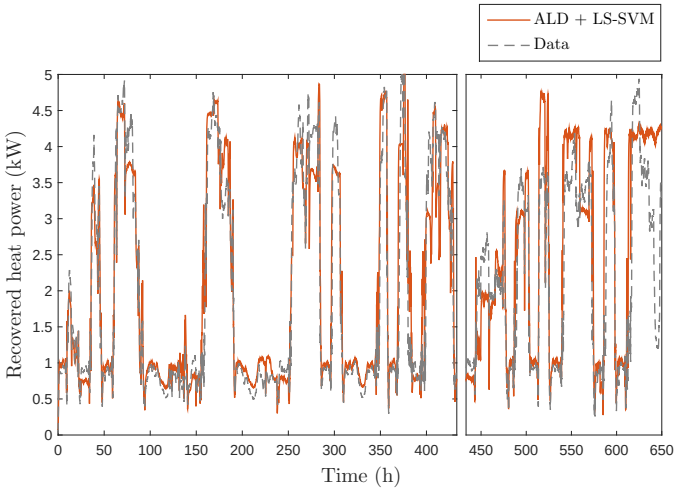


Figure 1: Estimated waste-heat recovery compared to training data (left) and validation data (right).

## 5. Conclusions

A kernel-based, non-linear regression method referred to as the least squares support vector machine has been presented. The method applies the 'kernel trick' to reformulate a regularized linear regression problem in terms of only inner products by introducing a kernel function. Solving the problem may become intractable for large data sets. Hence, a data reduction method based on approximate linear independence between feature vectors has been presented. The presented methods show promising results on a case study for estimation of the recovered waste-heat in ships.

## References

N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge university press, 2000.

C. M. Bishop, Pattern recognition and machine learning, vol. 1, Springer, 2006.

J. A. Suykens, J. De Brabanter, L. Lukas, J. Vandewalle, Weighted least squares support vector machines: robustness and sparse approximation, Neurocomputing 48 (1) (2002) 85–105.

Y. Engel, S. Mannor, R. Meir, The kernel recursive least-squares algorithm, IEEE Transactions on Signal Processing 52 (8) (2004) 2275–2285.