Reduced Rank Kernel Ridge Regression

Gavin C. Cawley (gcc@sys.uea.ac.uk) *
Telephone : 01603 593258
Fax : 01603 593345
School of Information Systems, University of East Anglia

Nicola L. C. Talbot

Abstract.
Ridge regression is a classical statistical technique that attempts to address the bias-variance trade-off in the design of linear regression models. A reformulation of ridge regression in dual variables permits a non-linear form of ridge regression via the well-known "kernel trick". Unfortunately, unlike support vector regression models, the resulting kernel expansion is typically fully dense. In this paper, we introduce a reduced rank kernel ridge regression (RRKRR) algorithm, capable of generating an optimally sparse kernel expansion that is functionally identical to that resulting from conventional kernel ridge regression (KRR). The proposed method is demonstrated to out-perform an alternative sparse kernel ridge regression algorithm on the Motorcycle and Boston Housing benchmarks.

Keywords: Ridge Regression, Sparse Kernel Approximation

This paper has not been submitted elsewhere in identical or similar form, nor will it be during the first three months after its submission to Neural Processing Letters.

* This work was supported in part by Royal Society research grant RSRG-22270.

1. Introduction

Ridge regression [4] is a method from classical statistics that implements a regularised form of least-squares regression. In its simplest form, given training data,

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{\ell}, \quad \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d, \quad y_i \in \mathcal{Y} \subset \mathbb{R},$$

ridge regression determines the parameter vector, $\boldsymbol{w} \in \mathbb{R}^d$, of a linear model, $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x}$, by minimising the objective function

$$W_{\mathrm{RR}}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{\ell}\sum_{i=1}^{\ell}(y_i - \boldsymbol{w} \cdot \boldsymbol{x}_i)^2. \tag{1}$$

The objective function used in ridge regression (1) implements a form of Tikhonov regularisation [12] of a sum-of-squares error metric, where γ is a regularisation parameter controlling the bias-variance trade-off [2]. This corresponds to penalised maximum likelihood estimation of $\boldsymbol{w}$, assuming the targets have been corrupted by an independent and identically distributed (i.i.d.) sample from a Gaussian noise process, with zero mean and variance σ^2, i.e.

$$y_i = \boldsymbol{w} \cdot \boldsymbol{x}_i + \epsilon_i, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

1.1. Kernel Ridge Regression

A non-linear form of ridge regression [8, 10, 11] can be obtained via the so-called "kernel trick", whereby a linear ridge regression model is constructed in a higher dimensional feature space, $\mathcal{F}$ ($\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F}$), induced by a non-linear kernel function defining the inner product

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}').$$

The kernel function, $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ may be any positive definite "Mercer" kernel, for instance the Gaussian radial basis function (RBF) kernel,

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\sigma^2}\right\}.$$

Note that the feature space $\mathcal{F}$ can be of an extremely high, or even infinite dimensionality, and so it is not generally feasible to evaluate the position of the data in feature space. Fortunately, the ridge regression algorithm can be expressed in such a way that the data, $\{\boldsymbol{x}_i\}_{i=1}^{\ell}$, appear only within inner products and can therefore be replaced by

evaluation of the kernel. The objective function minimised in kernel ridge regression can be written as,

$$W_{\mathrm{KRR}}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{\ell}\sum_{i=1}^{\ell}\xi_i^2,$$

subject to the constraints

$$\xi_i = y_i - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x_i}), \quad \forall\, i \in \{1, 2, \ldots, \ell\}.$$

The minimum of this optimisation problem coincides with the saddle-point of the primal Lagrangian

$$L_p(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{\ell}\sum_{i=1}^{\ell}\xi_i^2 + \sum_{i=1}^{\ell}\alpha_i(y_i - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) - \xi_i). \quad (2)$$

The conditions observed at the solution to this optimisation problem can be stated as follows:

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = 0 \;\Rightarrow\; \boldsymbol{w} = \sum_{i=1}^{\ell}\alpha_i\boldsymbol{\phi}(\boldsymbol{x}_i), \quad (3)$$

$$\frac{\partial L_p}{\partial \xi_i} = 0 \;\Rightarrow\; \alpha_i = 2\frac{\gamma}{\ell}\xi_i, \quad (4)$$

$$\frac{\partial L_p}{\partial \alpha_i} = 0 \;\Rightarrow\; \xi_i = y_i - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i). \quad (5)$$

Substituting (3–5) into (2) in order to eliminate $\boldsymbol{w}$ and $\boldsymbol{\xi}$, we obtain the dual Lagrangian,

$$L_d(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) - \frac{\ell}{4\gamma}\sum_{i=1}^{\ell}\alpha_i^2 + \sum_{i=1}^{\ell}y_i\alpha_i.$$

The optimal values for the Lagrange multipliers, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_\ell)^T$ are given by the minimiser of L_d. Differentiating with respect to $\boldsymbol{\alpha}$ and re-writing in matrix form,

$$\boldsymbol{\alpha} = \left(\boldsymbol{K} + \frac{\ell}{2\gamma}\boldsymbol{I}\right)^{-1}\boldsymbol{y},$$

where $\boldsymbol{K} = \{k_{ij}\}_{i,j=1}^{\ell}$, $k_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_\ell)^T$ and $\boldsymbol{I}$ is the identity matrix. From (3), we can see that the output of the kernel ridge regression model is given by

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{i=1}^{\ell}\alpha_i\boldsymbol{\phi}(\boldsymbol{x}_i)\cdot\boldsymbol{\phi}(\boldsymbol{x}), \\ &= \sum_{i=1}^{\ell}\alpha_i\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}). \end{aligned}$$

Unfortunately, unlike support vector regression models, this kernel expansion is in general fully dense, i.e. $\alpha_i \neq 0,\ \forall\ i \in \{1, 2, \ldots, \ell\}$. In this paper, we propose a reduced rank training algorithm which produces an equivalent *sparse* kernel expansion.

1.2. Reduced Rank Kernel Ridge Regression

The aim of reduced rank kernel ridge regression (RRKRR) is to identify a subset, $\{\boldsymbol{x}_i\}_{i\in\mathcal{S}} \subset \mathcal{D}$, of the training data ideally forming a basis in feature space, such that the feature space image of any element of the training data can be written as a weighted sum of the images of this subset, i.e.

$$\boldsymbol{\phi}(\boldsymbol{x}) \approx \hat{\boldsymbol{\phi}}_{\mathcal{S}}(\boldsymbol{x}) = \sum_{i\in\mathcal{S}} \zeta_i \boldsymbol{\phi}(\boldsymbol{x}_i), \qquad \forall\ \boldsymbol{x} \in \mathcal{D}.$$

The output of a kernel ridge regression model can then be written as a sparse kernel expansion involving only terms corresponding to the subset of the training data forming an approximate basis in feature space,

$$f(\boldsymbol{x}) \approx \sum_{i\in\mathcal{S}} \beta_i \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i\in\mathcal{S}} \beta_i \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}).$$

In this paper, we show that the coefficients of this expansion can be found efficiently by solving a family of only $|\mathcal{S}|$ linear equations in $|\mathcal{S}|$ unknowns (hence *reduced rank* kernel ridge regression). Furthermore, provided a sparse, but complete basis can be identified, the above approximations become equalities and so the reduced rank kernel ridge regression model is *functionally identical* to the conventional kernel ridge regression model.

The remainder of this paper is structured as follows: Section 2 describes the formation of a basis of the images of the data in feature space and goes on to present a reduced rank kernel ridge regression algorithm. Section 3 provides a comparison of standard, sparse and reduced rank kernel ridge regression algorithms. Section 4 provides some discussion of some issues raised and suggests avenues for further research; the work summarised in section 5.

2. Method

The method presented here consists of two parts, first a set of vectors forming an approximate, or better still a complete basis describing the training data in the feature space $\mathcal{F}$ is found, and then a linear ridge

regression model constructed in the sub-space of $\mathcal{F}$ spanned by these basis vectors.

2.1. Forming a Basis in Feature Space

In this work we adopt the greedy algorithm due to Baudat and Anouar [1] to construct a basis of the subspace of $\mathcal{F}$ populated by the training data, which we briefly summarise here. The normalised Euclidean distance between the position of a data item in feature space, $\boldsymbol{\phi}(\boldsymbol{x}_i)$, and $\hat{\boldsymbol{\phi}}_{\mathcal{S}}(\boldsymbol{x}_i)$, it's optimal reconstruction using the set basis vectors $\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{i\in\mathcal{S}}$, is given by

$$\delta_i = \frac{\|\boldsymbol{\phi}(\boldsymbol{x_i}) - \hat{\boldsymbol{\phi}}_{\mathcal{S}}(\boldsymbol{x}_i)\|^2}{\|\boldsymbol{\phi}(\boldsymbol{x}_i)\|^2}.$$

This distance can be expressed in terms of inner products, so via the "kernel trick", we have

$$\delta_i = 1 - \frac{\boldsymbol{K}_{\mathcal{S}i}^T \boldsymbol{K}_{\mathcal{S}\mathcal{S}}^{-1} \boldsymbol{K}_{\mathcal{S}i}}{k_{ii}}.$$

where $\boldsymbol{K}_{\mathcal{S}\mathcal{S}}$ is a square sub-matrix of $\boldsymbol{K}$, such that $\boldsymbol{K}_{\mathcal{S}\mathcal{S}} = \{k_{ij}\}_{i,j\in\mathcal{S}}$ and $\boldsymbol{K}_{\mathcal{S}i} = (k_{ji})_{j\in\mathcal{S}}^T$ is a column vector of inner products. To form a basis, we simply minimise the mean reconstruction error δ_i over all patterns in the training set, i.e. maximise

$$J(\mathcal{S}) = \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{\boldsymbol{K}_{\mathcal{S}i}^T \boldsymbol{K}_{\mathcal{S}\mathcal{S}}^{-1} \boldsymbol{K}_{\mathcal{S}i}}{k_{ii}}.$$

Starting with $\mathcal{S} = \emptyset$, a basis is constructed in a greedy manner, adding to $\mathcal{S}$ the training vector maximising $J(\mathcal{S})$ at each iteration. The algorithm terminates when $\boldsymbol{K}_{\mathcal{S}\mathcal{S}}$ is no longer invertible, indicating that a basis has been identified.

2.2. A Reduced Rank Training Algorithm

Here we consider a formulation of kernel ridge regression including a bias term [7], known as the least-squares support vector machine (LS-SVM) [10, 11]. The mapping implemented by a least-squares support vector machine is given by

$$f(\boldsymbol{x}) = \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}) + b = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{6}$$

The optimal values for the weight vector, $\boldsymbol{w}$, and bias, b, are given by the minimum of the objective function

$$W_{\mathrm{LSSVM}}(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{\ell}\sum_{i=1}^{\ell}(y_i - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) - b)^2.$$

Equivalently, the Lagrange multipliers minimising the corresponding dual optimisation problem are given by the solution of an augmented set of linear equations

$$\begin{bmatrix} \boldsymbol{\Omega} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix},$$

where $\boldsymbol{\Omega} = \boldsymbol{K} + \ell\gamma^{-1}\boldsymbol{I}$ and $\mathbf{1} = (1, 1, \ldots, 1)^T$. If the weight vector, $\boldsymbol{w}$, can be represented as a weighted sum of basis vectors, i.e.,

$$\boldsymbol{w} = \sum_{i\in\mathcal{S}} \beta_i \boldsymbol{\phi}(\boldsymbol{x}_i),$$

then we obtain the objective function minimised in reduced rank kernel ridge regression,

$$W_{\mathrm{RRKRR}}(\boldsymbol{\beta}, b) = \frac{1}{2}\sum_{i,j\in\mathcal{S}} \beta_i\beta_j k_{ij} + \frac{\gamma}{\ell}\sum_{i=1}^{\ell}(y_i - \sum_{j\in\mathcal{S}} \beta_j k_{ij} - b)^2.$$

Setting the partial derivatives of W_{RRKRR} with respect to $\boldsymbol{\beta}$ and b to zero, and dividing through by $2\gamma/\ell$, yields:

$$\sum_{i\in\mathcal{S}} \beta_i \sum_{j=1}^{\ell} k_{ij} + \ell b = \sum_{j=1}^{\ell} y_j$$

and

$$\sum_{i\in\mathcal{S}} \beta_i \left(\frac{\ell}{2\gamma} k_{ir} + \sum_{j=1}^{\ell} k_{jr}k_{ji} \right) + b\sum_{i=1}^{\ell} k_{ir} = \sum_{i=1}^{\ell} y_i k_{ir}, \qquad \forall\ r \in \mathcal{S}$$

These equations can be expressed as a system of $|\mathcal{S}|+1$ linear equations in $|\mathcal{S}|+1$ unknowns,

$$\begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{\Phi} \\ \boldsymbol{\Phi}^T & \ell \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \sum_{k=1}^{\ell} y_k \end{bmatrix},$$

where $\boldsymbol{\Omega} = \{\omega_{ij}\}_{i,j\in\mathcal{S}}$, $\omega_{ij} = \frac{\ell}{2\gamma}k_{ij} + \sum_{r=1}^{\ell} k_{rj}k_{ri}$, $\boldsymbol{\Phi}$ is an $|\mathcal{S}|$-dimensional column vector, whose i^{th} element is given by

$$\Phi_i = \sum_{j=1}^{\ell} k_{ij}, \quad \forall\ i \in \mathcal{S},$$

and $\boldsymbol{c}$ is an $|\mathcal{S}|$-dimensional column vector, whose i^{th} element is given by

$$c_i = \sum_{j=1}^{\ell} y_j k_{ij}, \quad \forall\, i \in \mathcal{S}$$

(for notational convenience, we assume that the training data is reordered such that $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$).

3. Results

In this section, the proposed reduced rank is evaluated over two well-known benchmark datasets (Motorcycle, and Boston Housing), that reveal both the strengths and limitations of this approach. The results obtained are compared with an alternative method of imposing sparseness due to Suykens *at al.* [10]: A kernel ridge regression model is trained on the entire dataset, yielding a vector of Lagrange multipliers, $\boldsymbol{\alpha}$. A small fraction of the data (say 5%), associated with multipliers having the smallest magnitudes, are discarded and the kernel ridge regression model retrained on the remaining data. This process is repeated until a sufficiently small kernel expansion is obtained.

3.1. The Motorcycle Dataset

The Motorcycle benchmark consists of a sequence of accelerometer readings through time following a simulated motor-cycle crash during an experiment to determine the efficacy of crash-helmets (Silverman [9]). Figure 1 shows conventional and reduced rank kernel kernel ridge regression models of the Motorcycle dataset, using a Gaussian radial basis function kernel. The reduced rank model is functionally identical to the standard kernel ridge regression model with only 18 basis vectors. The difference between the output of the reduced rank and standard kernel ridge regression models is shown in figure 2; these errors are very small in comparison with the scale of the data. Figure 3 compares the 10-fold root-mean-square (RMS) cross-validation error of reduced rank and sparse kernel ridge regression algorithms as a function of the number of training patterns included in the resulting kernel expansions. The regularisation and kernel parameters were determined in each trial via minimisation of the 4-fold cross-validation error using the Nelder-Mead Simplex method [6]. The cross-validation error is consistently lower for the reduced rank model regardless of the number of patterns forming the kernel expansion, becoming greater as the number of patterns decreases.

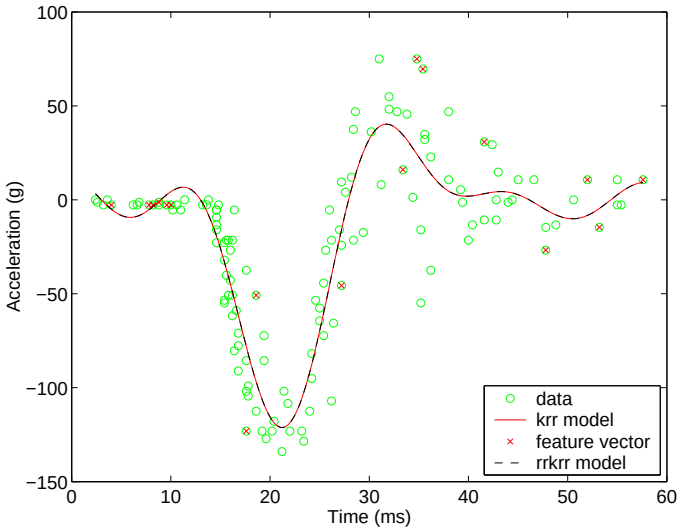

Figure 1. Kernel ridge regression (KRR) and reduced rank kernel ridge regression (RRKRR) models of the Motorcycle data set; note the standard and reduced rank kernel ridge regression models are essentially identical.

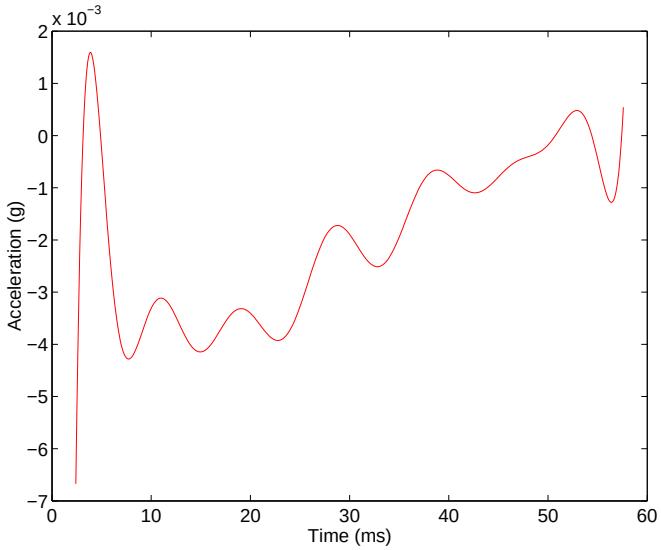

Figure 2. Difference between kernel ridge regression (KRR) and reduced rank kernel ridge regression (RRKRR) models of the Motorcycle data set.

3.2. The Boston Housing Dataset

The Boston Housing dataset describes the relationship between the median value of owner occupied homes in the suburbs of Boston and thirteen attributes representing environmental and social factors be-

lieved to be relevant [3]. In this case, for a Gaussian radial basis kernel, the feature vector selection process is unable to form a sparse basis. Figure 4 compares the 10-fold root-mean-square (RMS) cross-validation error of reduced rank and sparse kernel ridge regression algorithms as a function of the number of training patterns included in the resulting kernel expansions. Again, the regularisation and kernel parameters were determined in each trial via minimisation of the 4-fold cross-validation error. Although a sparse basis could not be found, the feature vector selection process identifies a near-optimal ranking of training patterns and so results in a good sparse approximation of the full kernel expansion.

4. Discussion

The most important aspect of the reduced rank kernel ridge regression algorithm is that at most only $|\mathcal{S}|$ columns of the kernel matrix, $\boldsymbol{K}$, need be stored in memory. The reduced rank training algorithm is therefore far better suited to large-scale applications, where the number of training patterns is sufficiently large that storage of the full kernel matrix is impractical and so the conventional training algorithm can not be used.

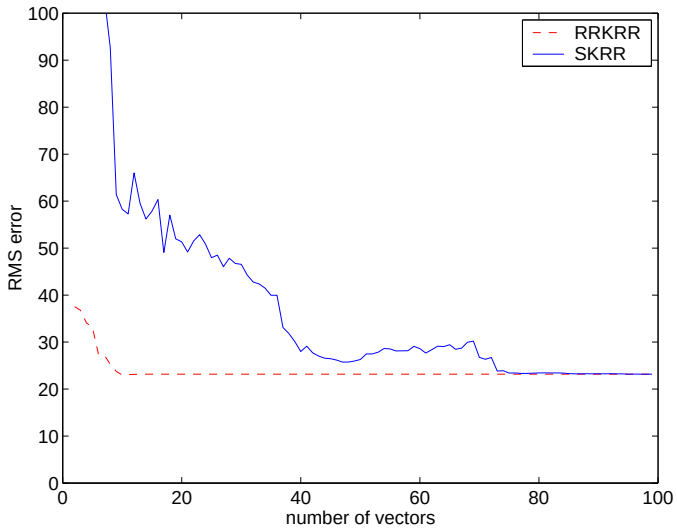

Figure 3. Cross-validation error of reduced rank and sparse kernel ridge regression models, over the motorcycle dataset, as a function of the number of training patterns included in the resulting kernel expansions.

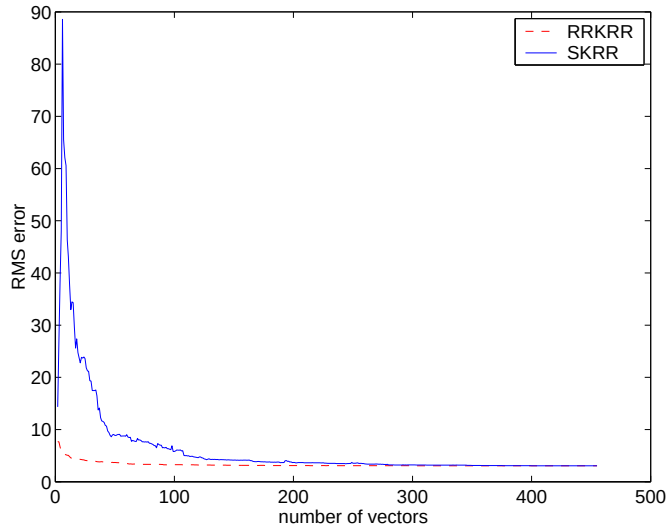

Figure 4. Cross-validation error of reduced rank and sparse kernel ridge regression models, over the Boston housing dataset, as a function of the number of training patterns included in the resulting kernel expansions.

The reduced rank kernel ridge regression method consistently outperforms the sparse kernel ridge regression approach on all datasets investigated, the gap in performance widening as terms are dropped from the kernel expansion. The principle reason for this is that sparse kernel ridge regression ignores the residuals for patterns not used to form the kernel expansion. In effect, the sparse kernel ridge regression algorithm solves a sequence of regression problems that become progressively less and less representative of the data as a whole. The objective function used in the reduced rank training algorithm, on the other hand, includes the residuals for all training patterns regardless of the size of the kernel expansion.

For model selection via cross-validation, there is no reason why the feature vector selection should not be performed only once over the whole dataset and then multiple models trained using the extracted basis vectors. In this way the expense of basis selection can be amortised across several models in an iterative model selection process, especially if selection of kernel parameters is performed in an "outer loop", while selection of the regularisation parameter is performed in an "inner loop".

If the feature vector selection extracts a set of data forming a complete basis for the data in feature space, the resulting kernel expansion will be exactly equivalent to the full least-squares support vector machine. It is not in general possible to extract a full basis for some kernel

functions, such as the RBF kernel used here [5]. However, for datasets with a low intrinsic dimensionality (e.g. the Motorcycle dataset) an extremely close approximate sparse basis may be found. For datasets where a sparse basis cannot be found, the method of Baudat and Anouar [1] still determines a near optimal ordering of training patterns for inclusion in the kernel expansion.

5. Summary

This paper proposed a novel reduced rank training algorithm for kernel ridge regression models. The method demonstrates performance superior to that of sparse least-squares support vector machines on a range of benchmark tasks. The method also provides an plausible approach for large-scale regression problems as it is not necessary to store the entire kernel matrix.

6. Acknowledgements

The authors would like to thank Johan Suykens and Gaston Baudat for interesting conversations that led to this work. The authors also thank Rob Foxall and Danilo Mandic for their helpful comments on previous drafts of this manuscript.

References

1. Baudat, G. and F. Anouar: 2001, 'Kernel-based Methods and Function Approximation'. In: *Proceedings, International Joint Conference on Neural Networks*, Vol. 3. Washington, DC, pp. 1244–1249.
2. Geman, S., E. Bienenstock, and R. Doursat: 1992, 'Neural networks and the bias/variance dilemma'. *Neural Computation* **4**(1), 1–58.
3. Harrison, D. and D. L. Rubinfeld: 1978, 'Hedonic prices and the demand for clean air'. *Journal Environmental Economics and Management* **5**, 81–102.
4. Hoerl, A. E. and R. W. Kennard: 1970, 'Ridge Regression: Biased Estimation for Nonorthogonal Problems'. *Technometrics* **12**(1), 55–67.
5. Micchelli, C. A.: 1986, 'Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions'. *Constructive Approximation* **2**, 11–22.
6. Nelder, J. A. and R. Mead: 1965, 'A simplex method for function minimization'. *Computer Journal* **7**, 308–313.
7. Poggio, T., S. Mukherjee, R. Rifkin, A. Rakhlin, and A. Verri: 2001, '*b*'. Technical Report AI Memo 2001-011, Massachusetts Institute of Technology, Cambridge, MA.

8. Saunders, C., A. Gammerman, and V. Vovk: 1998, 'Ridge Regression Learning Algorithm in Dual Variables'. In: *Proceedings, 15th International Conference on Machine Learning*. Madison, WI, pp. 515–521.
9. Silverman, B. W.: 1985, 'Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting'. *Journal of the Royal Statistical Society, B* **47**(1), 1–52.
10. Suykens, J., L. Lukas, and J. Vandewalle: 2000, 'Sparse Approximation using Least-Squares Support Vector Machines'. In: *Proceedings, IEEE International Symposium on Circuits and Systems*. Geneva, Switzerland, pp. 11757–11760.
11. Suykens, J. A. K., J. De Brabanter, L. Lukas, and J. Vandewalle: 2001, 'Weighted Least Squares Support Vector Machines : robustness and sparse approximation'. *Neurocomputing*.
12. Tikhonov, A. N. and V. Y. Arsenin: 1977, *Solutions of ill-posed problems*. New York: John Wiley.