Reduced Support Vector Machines:  
A Statistical Theory  

Su-Yun Huang  
Institute of Statistical Science  
Academia Sinica  
Taipei 115, Taiwan  
syhuang@stat.sinica.edu.tw  

Yuh-Jye Lee  
Computer Science & Information Engineering  
National Taiwan University of Science and Technology  
Taipei 106, Taiwan  
yuh-jye@mail.ntust.edu.tw  

Abstract  
The reduced support vector machine (RSVM) was proposed for the practical objective to overcome the computational difficulties as well as to reduce the model complexity in generating a nonlinear separating surface for a massive data set. It has been also successfully applied to other kernel-based learning algorithms. Also, there were experimental studies on RSVM that showed the efficiency of RSVM. In this paper we study the RSVM from the viewpoint of robust design in model building and consider the nonlinear separating surface as a mixture of kernels. The RSVM uses a compressed model representation instead of a saturated full model. Our main result shows that the uniform random selection of a reduced set to form the compressed model in RSVM is the optimal robust selection scheme in terms of the following criteria: (1) it minimizes an intrinsic model variation measure; (2) it minimizes the maximal model bias between the compressed model and the full model; (3) it maximizes the minimal test power in distinguishing the compressed model from the full model.  

Keywords: kernel methods, maximin design, minimax design, model complexity, reduced set, robust design, support vector machines, uniform design, uniform random selection.  

1 Introduction  
In recent years support vector machines (SVMs) with linear or nonlinear kernels [3, 4, 25] have become one of the most promising learning algorithms for classification as well as regression [5, 18, 19, 12], which are the fundamental tasks in data mining [30]. Via the use of kernel mapping, the SVM allows effective and flexible nonlinear classification. There are some major problems that confront large data classification due to dealing with a fully dense nonlinear kernel matrix. To overcome these computational problems many approximation schemes have been proposed to approximate a symmetric positive semidefinite matrix [23, 29]. Lee and Mangasarian proposed as an alternative the method of reduced support vector machine (RSVM) [13]. The key ideas of the RSVM are as follows. RSVM randomly selects a small portion of data set to generate a thin rectangular kernel matrix. Then it uses this much smaller rectangular kernel matrix to replace the fully dense square kernel matrix in the nonlinear SVM formulation. Computational time, as well as memory usage, is much less demanding for RSVM than that for a conventional SVM using the full kernel matrix. As a result,
RSVM also simplifies the characterization of the nonlinear separating surface. According to Occam’s razor [22], as well as Minimum Description Length (MDL) [20] and the numerical comparisons in [13], RSVM has better generalization ability than a conventional SVM. This reduced kernel technique has been successfully applied to other kernel-based learning algorithms, such as proximal support vector machine (PSVM) [8], $\epsilon$-smooth support vector regression ($\epsilon$-SSVR) [12] and kernel Fisher’s discriminant analysis (KFDA) [10]. Also, there were experimental studies on RSVM [16, 12] that showed the efficiency of RSVM. The RSVM results reported in [16] could be further improved if a larger weight parameter, which controls the tradeoff between the training error and model complexity would have been properly chosen. Since the RSVM has reduced the model complexity by using a much smaller rectangular kernel matrix we suggest to use a larger weight parameter on the data to enforce better data fidelity. Moreover, the numerical test in [13] on the Adult data set [21] shows that the standard error of test set correctness for 50 randomly chosen (with replacement) different reduced sets with the size of 1% of original data set is less than 0.001. In fact, the smallness of standard error can be used as a guidance to determining the size of the reduced set.

In this paper we study the RSVM from the viewpoint of robust design in model building and consider the nonlinear separating surface as a mixture of kernels. The RSVM uses a compressed model representation instead of a saturated full model, where fitting data to a saturated model is deemed to be unstable and causes prediction error to rise due to over-fitting [9] unless some sort of regularization is taken. Our main result shows that the uniform random selection of a reduced set to form the compressed model in RSVM is the optimal robust selection scheme in terms of the following criteria: (1) it minimizes an intrinsic model variation measure; (2) it minimizes the maximal model bias between the compressed model and the full model; (3) it maximizes the minimal test power in distinguishing the compressed model from the full model. This uniform random subset selection in RSVM also has a link to the popular uniform design [6].

We briefly outline the contents of the paper. Section 2 provides the main ideas and formulation for RSVM. Section 3 gives the uniform randomness an optimality interpretation in terms of a model variation measure. Section 4 further provides the uniform randomness certain minimaxity and maximinity properties. Section 5 concludes the paper. All proofs are placed in the Appendix.

A word about our notation and background material is given below. All vectors will be column vectors unless otherwise specified or transposed to a row vector by a prime superscript ‘. For a vector $x$ in the $n$-dimensional real space $\mathbb{R}^n$, the plus function $x_+$ is defined as $(x_+)_i = \max \{0, x_i\}$. The scalar (inner) product of two vectors $x$ and $z$ in the $n$-dimensional real space $\mathbb{R}^n$ will be denoted by $x'z$ and the $p$-norm of $x$ will be denoted by $\|x\|_p$. For a matrix $A \in \mathbb{R}^{m \times n}$, $A_i$ is the $i$th row of $A$. A column vector of ones of arbitrary dimension will be denoted by $1$. For $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times l}$, the kernel $K(A, B)$ maps $\mathbb{R}^{m \times n} \times \mathbb{R}^{n \times l}$ into $\mathbb{R}^{m \times l}$. In particular, if $x$ and $y$ are column vectors in $\mathbb{R}^n$ then, $K(x', y)$ is a real number, $K(A, x) = K(x', A')'$ is a column vector in $\mathbb{R}^m$ and $K(A, A')$ is an $m \times m$ matrix. The base of the natural logarithm will be denoted by $e$.

2 Reduced Support Vector Machines

Consider the problem of classifying points into two classes, $A_-$ and $A_+$. We are given a training data set $\{(x_i, y_i)\}_{i=1}^m$, where $x_i \in \mathcal{X} \subset \mathbb{R}^n$ is an input vector variable and $y_i \in \{-1, 1\}$ is a class label, indicating one of the two classes, $A_-$ and $A_+$, to which the input point belongs. We represent these data points by an $m \times n$ matrix $A$, where the $i$th row $A_i$ corresponds to the $i$th input data point. We use alternately $A_+ (a row vector)$ and $x_i$ (a column vector) for the same $i$th data point depending on convenience. We use an $m \times m$ diagonal matrix $D, D_{ii} = y_i$ to specify the class membership of each input point. The main goal of the classification problem is to find a classifier that can predict correctly the unseen class labels for new data inputs. It can be achieved by constructing a linear or nonlinear separating surface, $f(x) = 0$, which is implicitly defined by a kernel function. We classify a test point $x$ to $A_+$ if $f(x) \geq 0$, otherwise, to $A_-$. We will focus on nonlinear case in this paper. In conventional SVM as well as many kernel-based learning algorithms [3, 4, 25], generating a nonlinear separating surface has to deal with a fully dense kernel matrix with the size of the number of training
examples. When training a nonlinear SVM on a massive data set, the huge and dense full kernel matrix will lead to some computational difficulties as follows [13]:

- the size of the mathematical programming problem;
- the dependence of the nonlinear separating surface on the most of entire data set, which creates unwieldy storage problems that prevents the use of nonlinear kernels for anything but a small data set.

Moreover, using a saturated full model to represent the nonlinear separating surface could lead to the over-fitting risk, which causes high variation and less stable fit.

To avoid these difficulties and the over-fitting risk, the RSVM uses a very small random subset of size \( \tilde{m} \) of the original \( m \) data points, where \( \tilde{m} \ll m \), to build up the separating surface. We denote this random subset by \( \tilde{A} \), which is used to generate a much smaller rectangular matrix \( K(\tilde{A}, \tilde{A}') \in R^{m \times \tilde{m}} \). The reduced kernel matrix is served to replace the huge and fully dense square matrix \( K(A, A') \) used in conventional SVM to cut problem size, computation time and memory usage as well as to simplify the characterization of nonlinear separating surface.

We now briefly describe the RSVM formulation, which is derived from the generalized support vector machine (GSVM) [17] and the smooth support vector machine (SSVM) [14]. Of course, this reduced kernel technique can be applied to most of kernel-based learning algorithms, in particular for those solving the problem in the primal space. The RSVM solves the following unconstrained minimization problem for a general kernel \( K(\tilde{A}, \tilde{A}') \):

\[
\min_{(\tilde{u}, \gamma) \in R^{\tilde{m}+1}} \nu \frac{1}{2} \| p(1 - D\{K(\tilde{A}, \tilde{A}')\tilde{D}\tilde{u} - 1\gamma\}, \alpha) \|_2^2 + \frac{1}{2} (\tilde{u}'\tilde{u} + \gamma^2),
\]

where the function \( p(z, \alpha) \) is a very accurate smooth approximation to \( (z)_+ \) [14], which is applied componentwise to the vector \( 1 - D\{K(\tilde{A}, \tilde{A}')\tilde{D}\tilde{u} - 1\gamma\} \) and is defined componentwise by

\[
\{\text{the } j\text{th component of } p(z, \alpha)\} = z_j + \frac{1}{\alpha} \log(1 + e^{-\alpha z_j}), \quad \alpha > 0, \quad j = 1, \ldots, m.
\]

The function \( p(x, \alpha) \) converges to \( (x)_+ \) as \( \alpha \) goes to infinity. The positive weight parameter \( \nu \) here controls the tradeoff between the classification (training) error and the suppression of \( (\tilde{u}, \gamma) \). Since RSVM has essentially reduced the model complexity via using a much smaller rectangular kernel matrix, we will suggest to use a larger weight parameter \( \nu \) in RSVM than in conventional SVM.

The diagonal matrix \( \tilde{D} \in R^{\tilde{m} \times \tilde{m}} \), with ones or minus ones along its diagonal, specifies the class membership of each point in the reduced set. If we let \( \tilde{v} = \tilde{D}\tilde{u} \), i.e., \( \tilde{v}_i = \tilde{D}_{ii}\tilde{u}_i \), then \( \tilde{v}'\tilde{v} = \tilde{u}'\tilde{u} \) and problem (1) is equivalent to

\[
\min_{(\tilde{v}, \gamma) \in R^{\tilde{m}+1}} \nu \frac{1}{2} \| p(1 - D\{K(\tilde{A}, \tilde{A}')\tilde{v} - 1\gamma\}, \alpha) \|_2^2 + \frac{1}{2} (\tilde{v}'\tilde{v} + \gamma^2).
\]

A solution to this minimization problem (3) leads to the nonlinear separating surface

\[
f(x) = \sum_{i=1}^{\tilde{m}} \tilde{v}_i K(\tilde{A}_i, x) - \gamma = 0.
\]

The matrix \( \tilde{D} \) has disappeared in the formulation (3). This means that we do not need to record the class label of each data point in the reduced set. In fact, the reduced set is not necessary to be a subset of training set [11]. The minimization problem (3) retains the strong convexity and differentiability properties in the space, \( R^{\tilde{m}+1} \), of \( (\tilde{v}, \gamma) \) for any arbitrary rectangular kernel. Hence we can apply the Newton-Armijo method [14] directly to solve (3) and the existence and uniqueness of the optimal solution are also guaranteed. Moreover, the computational complexity of solving problem (3) by the Newton-Armijo method is \( O((\tilde{m} + 1)^3) \) while solving the nonlinear SSVM with
the full square kernel is $O((m + 1)^3)$ [14]. Typically, $\tilde{m} \ll m$. Also, the numerical test in [13] on the Adult data set [21] shows that the standard error of test set correctness for 50 randomly chosen (with replacement) different $A \in R^{326 \times 123}$ out of $A \in R^{32562 \times 123}$ is less than 0.001. In fact, the smallness of standard error can be used as a guidance to determining $\tilde{m}$. Finally, we note that this nonlinear separating surface (4) is a linear combination of a set of kernel functions plus a constant.

In a nutshell, the RSVM can be split into two parts. First, it selects a small random subset $\{K(\tilde{A}_1, \cdot), K(\tilde{A}_2, \cdot), \ldots, K(\tilde{A}_{\tilde{m}}, \cdot)\}$ from the likely highly correlated full-data basis set $\{K(A_i, \cdot)\}_{i=1}^m$ for building the separating function. The full-data set is inefficient with possibly heavy overlaps in function representation, but the conventional SVM has been using it. Secondly, the RSVM determines the best coefficients of the selected kernel functions by solving the unconstrained minimization problem (3) using the entire data set so that the surface will fit the whole data well. In next section, we study the RSVM from the viewpoint of model building and mixture of kernels.

3 Mixture models and optimal sampling design

The nonlinear SVM uses a saturated representation for the discriminant function:

$$f(x) = \sum_{i=1}^{m} v_i K(A_i, x) - \gamma.$$  \hspace{1cm} (5)

It is a linear combination of basis functions, $\{1\} \cup \{K(A_i, \cdot)\}_{i=1}^m$. One may regard the collection, $\{1\} \cup \{K(A_i, \cdot)\}_{i=1}^m$, as an overcomplete dictionary of functions when $m$ is large or approaching infinity. The representation (5) is said to be saturated, as there are as many parameters $v_i$ as data points. It is desirable to cut down the model complexity and fit a compressed model:

$$f(x) = \sum_{i=1}^{\tilde{m}} \tilde{v}_i K(\tilde{A}_i, x) - \gamma.$$ \hspace{1cm} (6)

We call $\{K(A_i, \cdot)\}_{i=1}^m$ the full-data basis set and $\{K(\tilde{A}_i, \cdot)\}_{i=1}^{\tilde{m}}$ a reduced basis set.

Concerning the choice of a reduced set, a simple way is the uniform random subset approach used in RSVM. The RSVM randomly selects a small portion of basis functions from the full set to generate the compressed model, while it fits the entire data set to this compressed model. Each candidate basis in the full set has equal chance of being selected. This uniform random selection is simple and straightforward without resorting to any search algorithm for optimal bases. But is it an effective way in building up an adequate compressed model for classification purpose? There have been experimental results showing its effectiveness for modeling classification surfaces [8, 13, 16, 10] as well as regression surfaces [12]. In this article, we justify and complement these experimental results with a theoretical study. We show that the uniform random selection is the optimal sampling scheme for basis recruits in terms of a few criteria stated later. Before proceed any further, we define some more notations and conditions.

- **Definition 1 (Training design)** Let $\xi_T$ be a probability measure on $X \subset R^n$. Assume that training inputs follow the probability distribution $\xi_T$. In statistical term, $\xi_T$ is called the training design. Later we will express the discriminant function as a mixture of kernels via mixing distribution $\xi_T$. Thus, $\xi_T$ is also called a mixing design for the discriminant surface.

The training inputs $x^i$, for $i = 1, \ldots, m$, are assumed i.i.d. random variables from $\xi_T$. After observing or sampling the training inputs, a discrete empirical version of the training design is given by $\xi_T m(x) = m^{-1} \sum_{i=1}^{m} \delta(x - x^i)$, where $\delta(\cdot)$ puts probability mass one at zero. In this article we do not distinguish between the generic training design and its empirical version, and use a unified notation $\xi_T$ for both unless otherwise specified.

- Let $\mathcal{H}$ denote the reproducing kernel Hilbert space generated by the kernel $K(x', z)$. Let $\mathcal{D} := \{K(\cdot, z)\}_{z \in X}$, which is an overcomplete dictionary for $\mathcal{H}$.
Definition 2 (Sampling design) Let \( \xi \) be a probability measure on \( \mathcal{X} \). Assume basis functions are selected from the dictionary \( \mathcal{D} \) according to the probability distribution \( \xi \), then we call the probability measure \( \xi \) a sampling design for basis selection. As will be seen later, \( \xi \) is also a mixing design for the discriminant surface.

Assume \( \xi_T \) and \( \xi \) are two equivalent measures, indicated by \( \xi_T = \xi \). That is, \( \xi \) is absolutely continuous with respect to \( \xi_T \), and conversely, \( \xi_T \) is also absolutely continuous with respect to \( \xi \). Denote the Radon-Nikodym derivative of \( \xi \) with respect to \( \xi_T \) by \( p(x) = d\xi(x)/d\xi_T(x) \).

Let \( \mathcal{P} \) be the collection of all such probability measures:

\[
\mathcal{P} = \{ \xi : \text{probability measure on } \mathcal{X} \text{ satisfying } \xi \equiv \xi_T \}. 
\] (7)

Again, we will not distinguish between the generic sampling design and its empirical version and use a unified notation \( \xi \) for both.

For convenience, throughout this paper a Gaussian kernel

\[
K_\sigma(x', z) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\|x - z\|^2/(2\sigma^2)\right\},
\]

is used for the RSVM. However, the theoretical results obtained in this article can easily extend to other kernels. We often suppress the subscript \( \sigma \) in \( K_\sigma \) and let \( K(x - z) := K_\sigma(x', z) \). The value of \( K(z - x) \) represents the inner product of resulting vectors of \( x \) and \( z \) in the feature space after a nonlinear mapping implicated and defined by the Gaussian kernel. We can also interpret it as a measure of similarity between \( x \) and \( z \). In other words, \( K(A_i, A) \) records the similarity between \( A_i \) and all training inputs. In particular, if we use the Gaussian kernel in the RSVM, we can interpret RSVM as an instance-based learning algorithms \([20]\). The reduced kernel matrix arranges only the similarity between reduced set and the entire training data set. In contrast to the \( k \)-nearest neighbor algorithm using the simple voting strategy, RSVM uses the weighted voting strategy, where weights and threshold are determined by a training procedure. That is, if the weighted sum, \( \hat{v}'K(\hat{A}, x) \), is greater than a threshold, \( \gamma \), the point \( x \) is classified as a positive example.

As seen in (5) and (6), the underlying discriminant function \( f(x) \) is modeled as a mixture of kernels plus a bias term. However, as our main focus here is to assess the variability of model building due to basis sampling from the dictionary \( \mathcal{D} \), the bias term \( \gamma \) may then, without loss of generality, be assumed zero. Then we adopt the following modeling for the discriminant function via mixing distribution \( \xi_T \):

\[
f(x) = \int K(x', z)v(z)d\xi_T(z),
\] (8)

where \( v(z) \) is a coefficient function satisfying \( \int v^2(z)d\xi_T(z) < \infty \). By re-expressing the above mixture via an arbitrary mixing distribution \( \xi \in \mathcal{P} \), we have

\[
f(x) = \int K(x', z)v(z)\frac{d\xi_T(z)}{d\xi(z)}d\xi(z) = \int K(x', z)v(z)/p(z) \, d\xi(z).
\] (9)

Let \( f(x, Z) := K(x, Z)v(Z)/p(Z) \), where \( Z \in \mathcal{X} \) is a random variable following the probability distribution \( \xi \). Bases, sampled via \( K(x, Z) \) with \( Z \sim \xi \), are used to build a compressed model for the discriminant function. The integrated variance of \( f(x, Z) \), as a measure for model variation, is given below:

\[
\int \text{Var}_\xi\{f(x, Z)\}dx = \int \int K^2(x', z)v^2(z)/p^2(z) \, d\xi(z)dx - \int f^2(x)dx.
\] (10)

The integrated model variance is used to assess the model robustness. The smaller it is, the better stability (less variability) the model built from \( \xi \) will retain. Therefore, we aim to find a good sampling design \( \xi \) with small integrated model variance. As the quantity \( \int f^2(x)dx \) does not involve \( \xi \), we may drop it from equation (10) in the minimization step and use the following measure of model variation.
Definition 3 (Model variation measure) We use the following measure to assess model variation due to sampling design $\xi$:

$$V(\xi) := \int \int K^2(x', z)v^2(z)/p^2(z) \, d\xi(z) dx$$

$$= \int \int K^2(x', z)v^2(z)/p(z) \, d\xi_T(z) dx. \quad (11)$$

(If $K^2(x', z)v^2(z)/p(z)$ is not integrable, set $V(\xi) = \infty$.)

By Lemma 1 in the Appendix, it is easy to get the optimal design which minimizes $V(\xi)$ over the set $\mathcal{P}$. The resulting optimal sampling design has its pdf with respect to $\xi_T$ given by

$$d\xi_{opt}/d\xi_T := p(z) \propto |v(z)| \int |K(x', z)| dx \propto |v(z)|,$$

if $\int |K(x', z)| dx$ is a constant, e.g., the Gaussian kernel case.

Theorem 1 (Optimal sampling design) Assume the kernel employed satisfy the condition that $\int |K(x', z)| dx$ is a constant, i.e., independent of $z$. The ideal optimal sampling design for basis selection is given by $\xi_{opt}$ with pdf $d\xi_{opt}(z)/d\xi_T(z) = p(z) \propto |v(z)|$.

The coefficient function $v(z)$ is not known and has yet to be estimated in the training process. Here we use a constant as a reference function to replace $v(z)$ and the resulting $p(z)$ is uniform with respect to $\xi_T$. There are two main reasons of using a constant reference function for $v(z)$. One is to indicate no prior preference or information on $v(z)$ prior to training process. The other is to reflect the intrinsic mechanism in SVM that tends to minimize the 2-norm, $\int v^2(z) d\xi_T(z)$, of $v(z)$. For an arbitrary fixed scale $\int v(z) d\xi_T(z) = c$ for some $c \in R$, the 2-norm of $v(z)$ is minimized when $v(z)$ is a constant function. Thus, the uniformity of $p(z)$ is justified. In other words, basis functions $K(\cdot, z)$ are sampled from the dictionary $\mathcal{D}$ according to the training design.

Corollary 1 Prior to training process, a constant is used as a reference for the coefficient function $v(z)$. Then the optimal sampling design for basis selection is given by $\xi_{opt} = \xi_T$, that is, $\xi_{opt} = \xi_T$. This justifies the uniform sampling scheme (on training data) used in the original RSVM of Lee and Mangasarian [13].

This random subset approach can drastically cut down the model complexity, while the sampling design helps to guide the basis selection in terms of minimal model variation (11). However, we remind the reader the quantity $V(\xi)$ does not account for variance incurred in parameter estimation, but only for the model variation caused by bases sampling. The resulting optimal sampling design in Corollary 1 is a uniform design with respect to $\xi_T$, which seeks basis points uniformly distributed over training points. The use of uniform design has been popular since 1980. See articles [6] and [7] for a nice survey of theory and application on uniform design. The uniform design is a space filling design and it seeks to obtain maximal model robustness.

4 Minimaxity and maximinity

In this section we show that the optimal sampling design in Corollary 1 also possesses other robustness properties, namely, the minimaxity and the maximinity. The ideas are taken from robust designs [26, 27, 28, 15, 1, 2] and customized into the context of RSVM. Recall the saturated and compressed models:

saturated model: $f(x) = \sum_{i=1}^{m} v_i K(A_i, x) - \gamma$; \hfill (12)

compressed model: $f(x) = \sum_{i=1}^{n} \tilde{v}_i K(\tilde{A}_i, x) - \gamma$. \hfill (13)
Using kernel $K$ corresponds to mapping the data into a feature Hilbert space with a map: $\Phi : \mathcal{X} \rightarrow \mathcal{U}$. In the feature space $\mathcal{U}$, the normal vector of the SVM separating hyperplane, $w^Tu - \gamma = 0$, can be expanded in terms of support vectors. The saturated model has the following full expansion for the normal vector

$$\text{saturated pre-image expansion: } w = \sum_{i=1}^{m} v_i \Phi(A_i),$$

while the compressed model has the reduced expansion

$$\text{compressed pre-image expansion: } w = \sum_{i=1}^{\tilde{m}} \tilde{v}_i \Phi(\tilde{A}_i).$$

Expansions given by (14) and (15) are called pre-image expansions. See [24] for approximating the full pre-image expansion by reduced-set expansion. The error induced from approximation by the reduced-set expansion is investigated below. We define some more notations. All the $\xi$'s referred below are in $\mathcal{P}$ (7).

- $\mathcal{F} := \text{linear span}\{K(A_i, \cdot)\}_{i=1}^{m}$ and $\mathcal{R} := \text{linear span}\{K(\tilde{A}_i, \cdot)\}_{i=1}^{\tilde{m}}$.
- $\mathcal{F}_\eta^- := \{g \in \mathcal{F} : \int_X g^2(z) d\xi_T(z) \leq \eta^2, \int_X K(\tilde{A}_i, z)g(z)d\xi_T(z) = 0, \ i = 1, \ldots, \tilde{m}\}$.
- $\mathcal{F}_\eta^+ := \{g \in \mathcal{F} : \int_X g^2(z) d\xi_T(z) \geq \eta^2, \int_X K(\tilde{A}_i, z)g(z)d\xi_T(z) = 0, \ i = 1, \ldots, \tilde{m}\}$.
- $\tilde{V}_\xi := \int_X K(\tilde{A}, z)K(z', \tilde{A}')d\xi(z)$, which is an $\tilde{m} \times \tilde{m}$ matrix.
- $B(f, \xi) := \int \{f(x) - K(x', \tilde{A}')\tilde{V}_\xi - \int K(\tilde{A}, z)f(z)d\xi(z)\}^2d\xi_T(x)$, where $f \in \mathcal{F}$, $\xi \in \mathcal{P}$ and $\tilde{V}_\xi$ is a generalized inverse for $\tilde{V}_\xi$.

We introduce below a measurement of approximation error which accounts for model bias between the full model (12) and the compressed model (13). For an arbitrary function $f \in \mathcal{F}$, the $L_2(d\xi_T)$-projection of $f$ onto the reduced space $\mathcal{R}$ is given by

$$P_\mathcal{R}f(x) = K(x', \tilde{A}')\tilde{V}_\xi^{-1} \int K(\tilde{A}, z)f(z)d\xi(z).$$

We use the following integrated (with respect to $d\xi_T$) square bias to account for model bias:

$$\text{model bias } (f) := \int \{f(x) - P_\mathcal{R}f(x)\}^2d\xi_T = B(f, \xi).$$

The quantity $B(f, \xi)$ reflects a degree of inadequacy in $f$ represented by the compressed representation. There are other means of defining alternative classes $\mathcal{F}_\eta^-$ and $\mathcal{F}_\eta^+$ and bias measure $B(f, \xi)$, see, for instance, [26, 27, 28, 15, 1, 2]. The minimaxity and maximinity discussed below depend on the particular way of specifying alternative classes and also on the definition of bias measure. To guard against the maximum possible model bias (or inadequacy), we would use a sampling design that minimizes the maximum bias.

**Theorem 2 (Minimaxity)** The minimax sampling design which minimizes the maximum bias $B(f, \xi)$ for $f \in \mathcal{F}_\eta^- \oplus \mathcal{R}$ is achieved by $\xi = \xi_T$, i.e.,

$$\sup_{f \in \mathcal{F}_\eta^- \oplus \mathcal{R}} B(f, \xi_T) = \inf_{\xi \in \mathcal{P}} \sup_{f \in \mathcal{F}_\eta^- \oplus \mathcal{R}} B(f, \xi).$$

Equivalently, the uniform design with respect to $\xi_T$ is the optimal sampling design to minimize the maximum model bias between the full model (12) and the compressed model (13).
The space \( f \in \mathcal{F}_\eta^+ \oplus \mathcal{R} \) specifies a compact neighborhood of the reduced space \( \mathcal{R} \). The optimal sampling design, \( \xi = \xi_T \), is used to guard against maximum possible bias as the saturated model ranging over an arbitrary compact neighborhood of the compressed model.

Below we further provide the optimal sampling design with a maximinity property. Recall the definition of \( \mathcal{F}_\eta^+ \). The first condition specifies a class which deviates from the reduced space with at least a distance \( \eta \). When the true model is at least \( \eta \)-distance away from the reduced space \( \mathcal{R} \), i.e., it falls outside a neighborhood of \( \mathcal{R} \) and into the class specified by \( \mathcal{F}_\eta^+ \oplus \mathcal{R} \), we would hope to be able to distinguish between the true model and the compressed model in a lack of fit (LOF) test. That is, we want as large model bias as possible in order to have adequate ability (or test power) to distinguish between the two models in a LOF test. Theorem 3 tells us that sampling design \( \xi = \xi_T \) achieves the maximum test power for a worst (most indistinguishable) possible \( f \) in \( \mathcal{F}_\eta^+ \oplus \mathcal{R} \).

**Theorem 3 (Maximinity)** For a given nonnegative function \( f(x) \) as a mixture of kernels and express it as a linear combination of kernel functions sampled from the full-data basis set. We also interpret the RSVM as an instance-based learning algorithms. The reduced kernel matrix records the similarity between the reduced set functions sampled from the full-data basis set. We further provide this RSVM sampling strategy some robustness properties: the minimaxity and maximinity interpretation in terms of model bias and test power minimization problem. Our main result shows that uniformly and randomly selecting the reduced point of RSVM is the optimal sampling strategy for recruiting kernel bases in the sense of minimizing model variation measure. We further provide this RSVM sampling strategy some robustness properties: the minimaxity and maximinity interpretation in terms of model bias and test power respectively. This theoretical study provides a better understanding of RSVM learning algorithm and helps to advance the random subset approach to most kernel-based learning algorithms.

## 5 Conclusion

We study the RSVM from a robust design for model building point of view. We consider the nonlinear classifier \( f(x) \) as a mixture of kernels and express it as a linear combination of kernel functions sampled from the full-data basis set. We also interpret the RSVM as an instance-based learning algorithms. The reduced kernel matrix records the similarity between the reduced set and the entire training data set. The RSVM has used the weighted voting strategy to classify a new data point. That is, if the weighted sum, \( K(x', A')\xi \) is greater than a threshold \( \gamma \), then the point \( x \) is classified as positive example, where \( \xi \) and \( \gamma \) are determined by solving an unconstrained minimization problem. Our main result shows that uniformly and randomly selecting the reduced set of RSVM is the optimal sampling strategy for recruiting kernel bases in the sense of minimizing a model variation measure. We further provide this RSVM sampling strategy some robustness properties: the minimaxity and maximinity interpretation in terms of model bias and test power respectively. This theoretical study provides a better understanding of RSVM learning algorithm and helps to advance the random subset approach to most kernel-based learning algorithms.

## 6 Appendix

**Lemma 1** For a given nonnegative function \( t(z) \), let \( \mathcal{P}_t \) denote the collection of probability density functions \( p(z) \in \mathcal{P} \), where \( \mathcal{P} \) is as defined in (7), such that \( \int_X t(z)/p(z))d\xi_T(z) < \infty \). (Here 0/0 is defined to be zero.) Then the solution to the following optimization problem

\[
\arg \min_{p \in \mathcal{P}_t} \int_X t(z)/p(z))d\xi_T(z)
\]

is given by \( p(z) = c^{-1}\sqrt{t(z)} \), where \( c = \int_X \sqrt{t(z)} d\xi_T(z) \).

**Proof:** Since \( \int_X t(z)/p(z))d\xi_T(z) < \infty \) and \( \int_X p(z)d\xi_T(z) = 1 \), by Hölder inequality, we have

\[
\int_X t(z)/p(z))d\xi_T(z) = \int_X t(z)/p(z))d\xi_T(z) \cdot \int_X p(z)d\xi_T(z) \geq \left( \int_X \sqrt{t(z)}d\xi_T(z) \right)^2.
\]

Equality holds if and only if, there exists some nonzero constant \( \beta \) such that

\[
t(z)/p(z) = \beta p(z) \ a.e. \text{ with respect to } \xi_T.
\]

Since \( p(z) \) is a pdf, then equality holds if and only if \( p(z) = c^{-1}\sqrt{t(z)} \).

\[\square\]
Lemma 2 \( \sup_{g \in \mathcal{F}_\eta^\sim} B(g, \xi_T) = \eta^2. \)

Proof: For \( g \in \mathcal{F}_\eta^\sim \), we have \( \int K(\hat{A}, z)g(z)d\xi_T(z) = 0. \) Then
\[
B(g, \xi_T) := \int \left\{ g(x) - K(x', \hat{A}')\hat{V}_{\xi_T} \int K(\hat{A}, z)g(z)d\xi_T(z) \right\}^2 d\xi_T(x)
\]
\[
= \int g^2(x)d\xi_T \leq \eta^2, \text{ for } g \in \mathcal{F}_\eta^\sim.
\]

The proof can be completed by finding a function \( g_0 \in \mathcal{F}_\eta^\sim \) such that the above equality holds. A construction for such a function goes as follows. We first find a partition consisting of measurable sets \( \mathcal{X} = \cup_{i=1}^{\tilde{m}+1} E_i \) with \( E_i \) disjoint and \( \xi_T(E_i) > 0 \). Let \( \beta = (\beta_1, \ldots, \beta_{\tilde{m}+1})' \) be a non-null solution to the following system of \( \tilde{m} \) equations in \( \tilde{m} + 1 \) unknowns:
\[
\sum_{i=1}^{\tilde{m}+1} \beta_i \int_{E_i} K(\hat{A}_j, z)d\xi_T(z) = 0, \quad j = 1, \ldots, \tilde{m}.
\]

Let
\[
g_0(x) = \eta \left( \sum_{i=1}^{\tilde{m}+1} \beta_i^2 \xi_T(E_i) \right)^{-1/2} \cdot \sum_{i=1}^{\tilde{m}+1} \beta_i I_{E_i}(x). \tag{20}
\]

Then \( \int g_0^2(x)d\xi_T(x) = \eta^2 \) and, by (19) and (20), it is easy to see that, for \( j = 1, \ldots, \tilde{m}, \)
\[
\int g_0(x)K(\hat{A}_j, x)d\xi_T(x) = \eta \left( \sum_{i=1}^{\tilde{m}+1} \beta_i^2 \xi_T(E_i) \right)^{-1/2} \sum_{i=1}^{\tilde{m}+1} \beta_i \int_{E_i} K(\hat{A}_j, x)d\xi_T(x) = 0. \tag{21}
\]

Thus, \( g_0 \in \mathcal{F}_\eta^\sim \). Also, for such a \( g_0 \), we have \( B(g_0, \xi_T) = \int_{\mathcal{X}} g_0^2(x)d\xi_T(x) = \eta^2. \) \( \square \)

Lemma 3 \( \inf_{g \in \mathcal{F}_\eta^\sim} B(g, \xi_T) = \eta^2. \)

Proof: For \( g \in \mathcal{F}_\eta^\sim, \)
\[
B(g, \xi_T) := \int \left\{ g(x) - K(x', \hat{A}')\hat{V}_{\xi_T} \int K(\hat{A}, z)g(z)d\xi_T(z) \right\}^2 d\xi_T(x)
\]
\[
= \int g^2(x)d\xi_T \geq \eta^2.
\]

The function \( g_0 \) given in (20) is also in \( \mathcal{F}_\eta^+ \) and satisfies the equality \( B(g, \xi_T) = \eta^2. \) \( \square \)

Lemma 4 \( B(h, \xi) = 0 \) for all \( h \in \mathcal{R} \) and all \( \xi \in \mathcal{P}. \)

Proof: For any \( h(z) \in \mathcal{R}, \) there exists a vector \( \hat{h} \in R^{\tilde{m}} \) such that \( h(z) \) can be represented as \( h(z) = \hat{h}'K(\hat{A}, z) \)
\[
K(x', \hat{A}')\hat{V}_{\xi} \int K(\hat{A}, z)h(z)d\xi(z)
\]
\[
= K(x', \hat{A}')\hat{V}_{\xi} \int K(\hat{A}, z)K(z', \hat{A}')\hat{h}d\xi(z)
\]
\[
= K(x', \hat{A}')\hat{h} = h(x).
\]

Thus, we have \( B(h, \xi) = 0. \) \( \square \)

The direct sum in \( \mathcal{F}_\eta \oplus \mathcal{R} \) and \( \mathcal{F}_\eta^\sim \oplus \mathcal{R} \) is in the sense of Banach space direct sum, which is isomorphic to Hilbert space direct sum (with inner product \( <f, g> = \int fg d\xi_T \)).
Proof for Theorems 2 and 3: We will show that

1. \( \inf_{\xi \in \mathcal{P}} \sup_{g \in \mathcal{F}_\eta^+ \oplus \mathcal{R}} B(g, \xi) = \sup_{g \in \mathcal{F}_\eta^- \oplus \mathcal{R}} B(g, \xi_T) \)
   and
2. \( \sup_{\xi \in \mathcal{P}} \inf_{g \in \mathcal{F}_\eta^+ \oplus \mathcal{R}} B(g, \xi) = \inf_{g \in \mathcal{F}_\eta^- \oplus \mathcal{R}} B(g, \xi_T) \).

For an arbitrary \( f \in \mathcal{F}_\eta^+ \oplus \mathcal{R} \) (or \( f \in \mathcal{F}_\eta^- \oplus \mathcal{R} \)), \( f \) can be decomposed uniquely as \( f = g + h \), where \( g \in \mathcal{F}_\eta^- \) (or \( g \in \mathcal{F}_\eta^+ \)) and \( h \in \mathcal{R} \). Since \( B(h, \xi) = 0 \) for \( h \in \mathcal{R} \), then \( B(f, \xi) = B(g, \xi) \). Therefore, it is sufficient to prove statement (1) for \( g \) in the class \( \mathcal{F}_\eta^- \) and statement (2) for \( g \) in the class \( \mathcal{F}_\eta^+ \).

That is, instead of (1) and (2), it is sufficient to show

1'. \( \inf_{\xi \in \mathcal{P}} \sup_{g \in \mathcal{F}_\eta^-} B(g, \xi) = \sup_{g \in \mathcal{F}_\eta^-} B(g, \xi_T) \)
   and
2'. \( \sup_{\xi \in \mathcal{P}} \inf_{g \in \mathcal{F}_\eta^+} B(g, \xi) = \inf_{g \in \mathcal{F}_\eta^+} B(g, \xi_T) \).

(1') We will show that there exists a \( g_0 \in \mathcal{F}_\eta^- \) such that \( B(g_0, \xi) \geq \eta^2 \) for any \( \xi \in \mathcal{P} \). Then, combined with Lemma 2, we get

\[
B(g_0, \xi) \geq \eta^2 = \sup_{g \in \mathcal{F}_\eta^-} B(g, \xi_T).
\]

If so, statement (1') is then proved. Below we construct such a \( g_0 \).

Let \( \nu = \xi - \xi_T \), a signed measure. By Hahn decomposition theorem there exists a measurable set \( E \) such that \( \nu \) is positive on \( E \) and negative on \( E^c \). We can then find a partition \( E = \cup_{i=1}^{2\tilde{n}+1} E_i \) with \( E_i \) disjoint and \( \nu(E_i) > 0 \). Since \( \xi = \nu + \xi_T \) and \( \nu \) is positive on each \( E_i \), then \( \xi(E_i) > 0 \) and then \( \xi_T(E_i) > 0 \), too, as \( \xi \) is absolutely continuous with respect to \( \xi_T \). Let \( \beta = (\beta_1, \ldots, \beta_{2\tilde{n}+1})' \) be a non-null solution to the following system of \( 2\tilde{m} \) equations (\( \tilde{m} \) many are integrals with respect to \( d\xi_T \) and the other \( \tilde{m} \) many are integrals with respect to \( d\xi \)) in \( 2\tilde{m} + 1 \) unknowns:

\[
\sum_{i=1}^{2\tilde{n}+1} \beta_i \int_{E_i} K(\tilde{A}_j, z) d\xi_T(z) = 0, \quad j = 1, \ldots, \tilde{m}.
\]

(23)

\[
\sum_{i=1}^{2\tilde{n}+1} \beta_i \int_{E_i} K(\tilde{A}_j, z) d\xi(z) = 0, \quad j = 1, \ldots, \tilde{m}.
\]

(24)

Let

\[
g_0(x) := \eta \left( \sum_{i=1}^{2\tilde{n}+1} \beta_i^2 \xi_T(E_i) \right)^{-1/2} \cdot \sum_{i=1}^{2\tilde{n}+1} \beta_i I_{E_i}(x).
\]

(25)

Since \( \xi = \nu + \xi_T \) and \( \nu \) is positive on \( E \), then

\[
\int_E g_0^2(x)d\xi(x) \geq \int_E g_0^2(x)d\xi_T(x) = \eta^2.
\]

(26)

By (23), we get \( \int_{E_i} K(\tilde{A}_j, x) d\xi_T(x) = 0 \) for \( 1 \leq j \leq \tilde{m} \). In addition, from the equality in (26), \( g_0 \) is in \( \mathcal{F}_\eta^- \). Since \( \int_{E_i} K(\tilde{A}_j, x) d\xi(x) = 0 \) for \( 1 \leq j \leq \tilde{m} \) by (24). Thus,

\[
B(g_0, \xi) := \int \left\{ g_0(x) - K(x', \tilde{A}) \tilde{V}_\xi^{-1} \int K(\tilde{A}, z) g_0(z) d\xi(z) \right\}^2 d\xi_T(x)
\]

\[
= \int g_0^2(x)d\xi_T(x) = \eta^2.
\]

Therefore, proof for (1') is completed.

(2') We will show that there exists an \( h_0 \in \mathcal{F}_\eta^+ \) such that \( B(h_0, \xi) \leq \eta^2 \) for any \( \xi \in \mathcal{P} \). Then, combined with Lemma 3, we get

\[
B(h_0, \xi) \leq \eta^2 = \inf_{g \in \mathcal{F}_\eta^+} B(g, \xi_T).
\]

(27)

If so, statement (2') can then be proved. Below we construct such an \( h_0 \).
Let $\nu = \xi - \xi_T$ and $E^c$ be as defined earlier in this proof. We then find a partition for $E^c$ with $E^c = \bigcup_{i=1}^{\tilde{m}+1} U_i$ with $U_i$ disjoint and $\nu(U_i) < 0$. Then, $\xi_T(U_i) = \xi(U_i) - \nu(U_i) > 0$. Let $\beta = (\beta_1, \ldots, \beta_{\tilde{m}+1})^t$ be a non-null solution to the following system of $\tilde{m}$ equations in $\tilde{m}+1$ unknowns:

$$\sum_{i=1}^{\tilde{m}+1} \beta_i \int_{U_i} K(\hat{A}, z) d\xi_T(z) = 0, \ j = 1, \ldots, \tilde{m}. \quad (28)$$

Let

$$h_0(x) := \eta \left( \sum_{i=1}^{\tilde{m}+1} \beta_i^2 \xi_T(U_i) \right)^{-1/2} \cdot \sum_{i=1}^{\tilde{m}+1} \beta_i I_{U_i}(x). \quad (29)$$

Then $\int h_0(x) d\xi_T(x) = \eta^2$ and $\int h_0(x) K(\hat{A}, x) d\xi_T(x) = 0$ for $j = 1, \ldots, \tilde{m}$. Thus, $h_0 \in \mathcal{F}_1^+$. Also,

$$B(h_0, \xi) = \int_{E^c} \left\{ h_0(x) - K(x', \hat{A}) \hat{V}_\xi \int_{E^c} K(\hat{A}, z) h_0(z) d\xi_T(z) \right\}^2 d\xi_T(x) \leq \int_{E^c} h_0^2(x) d\xi_T(x) = \eta^2.$$

Therefore, proof for (2') is completed. \qed

References


