Stoch Environ Res Risk Assess (2007) 21:747–764
DOI 10.1007/s00477-006-0087-9

ORIGINAL PAPER

# On selection of kernel parametes in relevance vector machines for hydrologic applications

**Shivam Tripathi · Rao S. Govindaraju**

Published online: 8 December 2006


**Abstract** Recent advances in statistical learning theory have yielded tools that are improving our capabilities for analyzing large and complex datasets. Among such tools, relevance vector machines (RVMs) are finding increasing applications in hydrology because of (1) their excellent generalization properties, and (2) the probabilistic interpretation associated with this technique that yields prediction uncertainty. RVMs combine the strengths of kernel-based methods and Bayesian theory to establish relationships between a set of input vectors and a desired output. However, a bias–variance analysis of RVM estimates revealed that a careful selection of kernel parameters is of paramount importance for achieving good performance from RVMs. In this study, several analytic methods are presented for selection of kernel parameters. These methods rely on structural properties of the data rather than expensive re-sampling approaches commonly used in RVM applications. An analytical expression for prediction risk in leave-one-out cross validation is derived. For brevity, the effectiveness of the proposed methods is assessed first by data generated from the benchmark sinc function, followed by an example involving estimation of hydraulic conductivity values over a field based on observations. It is shown that a straightforward maximization of likelihood function can lead to misleading results. The proposed methods are found to yield robust estimates of parameters for kernel functions.

S. Tripathi · R. S. Govindaraju (✉)
School of Civil Engineering, Purdue University,
West Lafayette, IN 47907, USA
e-mail: govind@ecn.purdue.edu

**Keywords** Bayesian learning · Relevance vector machines · Interpolation · Leave-one-out cross-validation · VC dimension · Bayes information criterion · Power spectrum

## 1 Introduction

In recent decades, with improved information technology and remote sensing tools, our capabilities of collecting hydrological data have increased many fold. An unprecedented investment in collecting hydrologic data has resulted in large archives of hydrologic data (e.g., USGS national water information system, the global observing systems information center database, national operational hydrologic remote sensing center datasets, agricultural research service water database, etc.). A large fraction of such data tends to be high-dimensional (i.e., exhibits large variability over a wide range of space and time scales) and severely under-constrained (sparse coverage over the input space), and is often interspersed with spurious data that confound analysis. This renders many of the previous learning algorithms such as local regression, spline interpolation, and logistic regression either inefficient or inapplicable. For hydrologic applications, the challenges are (a) achieving good generalization performance in high dimensional data where curse of dimensionality stipulates exponential increase in the amount of data for good predictions; (b) representing and effectively combining the available physical knowledge of hydrological systems into data learning algorithms; (c) characterizing and quantifying uncertainty in predictions; and finally (d) making learning algorithms computationally efficient to handle voluminous data.

Kernel methods, which have been gaining popularity in machine learning, provide a radically different approach for high dimensional learning problems. Unlike traditional learning algorithms where data are represented individually, kernel methods seek pairwise comparisons using kernel functions (Scholkopf and Smola 2002). This representation provides a simple and elegant way of capturing nonlinear relationships between input vectors and corresponding outputs using linear algorithms. Examples are support vector machines (Vapnik 1995), kernel Fisher discriminant (Mika et al. 1999), and kernel principal component analysis (Twining and Taylor 2003; Muller et al. 2001; Wu et al. 1997). These methods have demonstrated excellent generalization capabilities and good efficiency in high dimensional problems. However, most of these methods cannot provide important measures such as predictive distribution of the evidence as they are not developed with a probabilistic foundation. Thus the important issue of model uncertainty remains unsolved.

A possible way to overcome this problem is to exploit versatility of kernel methods in conjunction with Bayesian learning. Based on this idea, Tipping (2001) developed relevance vector machines (RVMs) that are kernel-based methods formulated under Bayesian construct. RVMs provide sparse solution to regression tasks by implementing Bayesian automatic relevance determination (MacKay 1994) in the transformed kernel space. RVMs have achieved excellent results for many learning problems including 3D image analysis (Agarwal and Triggs 2006), optical diagnosis of cancer cells (Majumder et al. 2005; Wei et al. 2005), prediction of chaotic time series (Quinonero-Candela and Hansen 2002) and analysis of radar data (Kovvali and Carin 2004). Recently, RVMs have found hydrologic applications in groundwater quality modeling (Khalil et al. 2005a), real time management of reservoir releases (Khalil et al. 2005b), and modeling of chaotic hydrologic time series (Khalil et al. 2006).

While RVMs are being used increasingly in hydrologic applications, there still remain some unresolved hurdles for their successful implementation. An important question that has not received the attention it deserves is the selection of kernel function parameters. It is widely acknowledged that a key factor that determines the generalization capabilities of kernel methods in general is the choice of kernel width parameter (Scholkopf and Smola 2002; Hastie et al. 2001; Vapnik 1995). The kernel function can be interpreted as a nonlinear transformation that maps the input space to a higher dimensional feature space. According to Cover's theorem (Cover 1965) a linear function can be formulated in the higher dimensional feature space to seek a nonlinear relationship between inputs and outputs in the original input space. Thus, the problem of choosing a kernel is equivalent to finding an appropriate form of data representation for learning (Evgeniou et al. 2000). Although, there is no available general solution to this problem, several empirical and theoretical studies provide insights for specific applications (Scholkopf et al. 1999; Cherkassky and Mulier 1998; Vapnik 1999, 1998). Recently, Lanckriet et al. (2004) proposed a method for extracting kernel matrix from data using semi-definite programming for a classification problem. Wang et al. (2003) proposed a method for determining the width of a RBF kernel based on scale-space theory in computer vision for support vector machines. Cherkassky and Ma (2004) suggested that width of a radial basis function (RBF) kernel for regression problems can be reasonably estimated form the range of input data vectors.

In spite of various theoretical and empirical studies on kernel methods, there is no general consensus on appropriate choice of kernel parameters because of contradictory opinions presented by authors (Cherkassky and Ma 2004). Hence, for many applications, re-sampling from data remains the only recourse for model validation. Unfortunately learning kernel parameters via traditional re-sampling methods is not only computationally expensive but requires large amount of data for successful implementation. The objective of this work is to present techniques for practical selection of kernel function parameters, and to show the effectiveness of the proposed techniques through an example of spatial interpolation of hydrologic data.

The remainder of this paper is structured as follows: Section 2 presents the mathematical formulation of RVMs. The details of the data used for the study are presented in Sect. 3. In Sect. 4, results of bias–variance analysis of RVMs are provided to gain insights into the possible consequences of inappropriate selection of kernel function parameters. In Sect. 5.1, Tipping's (2001) method of kernel function parameter selection by maximizing the likelihood is examined for the example datasets. In Sect. 5.2, the linearity in the formulation of RVMs is exploited to develop an analytical expression for leave-one-out cross-validation. In Sect. 5.3, the applicability of analytical model selection methods for obtaining optimum kernel function parameters is demonstrated. In Sect. 5.4, the possibility of selecting kernel parameters from data is explored. Thus, Sects. 4, 5.2, 5.3, and 5.4 constitute novel contributions of this paper. The proposed new kernel

parameter selection methods are then applied to the example data sets and results obtained are discussed in Sect. 6. A set of conclusions drawn from this study and recommendations for future use are presented in Sect. 7.

## 2 Mathematical formulation

Consider a finite training sample of $N$ patterns $\{(\boldsymbol{x}_i, y_i), i = 1, ..., N\}$, where $\boldsymbol{x}_i$ denotes $i^{th}$ input pattern in a $d$-dimensional space (i.e., $\boldsymbol{x}_i = [x_{i1}, ..., x_{id}] \in \Re^d$; $\mathbf{x} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^{\mathrm{T}}$) and $y_i \in \Re$ is the corresponding output. Further, let the learning machine $f(\mathbf{x}; \boldsymbol{w})$ be defined as a linear combination of kernel functions

$$f(\mathbf{x}; \boldsymbol{w}) = \hat{\boldsymbol{y}} = \sum_{j=0}^{N} w_j K(\mathbf{x}, \boldsymbol{x}_j) = \boldsymbol{\Phi}\boldsymbol{w} \tag{1}$$

such that

$$\boldsymbol{y} = f(\mathbf{x}; \boldsymbol{w}) + \boldsymbol{\varepsilon} \tag{2}$$

where, weight vector $\boldsymbol{w} = [w_0, ..., w_M]^{\mathrm{T}}$ is an adjustable or tuning parameter, $K(\mathbf{x}, \boldsymbol{x}_j)$ are the kernel functions with $K(\mathbf{x}, \boldsymbol{x}_0)=1$, $\boldsymbol{\Phi}_{N\times N+1}$ is the design matrix with elements $\Phi_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i = 1, ..., N$ ; $j = 0, ..., N$ and $\boldsymbol{\varepsilon}$ is the error term. There are several possibilities for the choice of kernel function, including linear, algebraic polynomials, trigonometric polynomials, RBFs, and sigmoid functions. In this study we have adapted a RBF kernel that is defined as

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j, \vec{\boldsymbol{\sigma}}_j) = \exp\left(-\sum_{z=1}^{d} \frac{(x_{iz} - x_{jz})^2}{\sigma_{jz}^2}\right) \tag{3}$$

where $\vec{\boldsymbol{\sigma}}_{\boldsymbol{j}} = [\sigma_{j1}, \ldots, \sigma_{jd}]$ is the width of the RBF kernel. In practice, this is held constant for all basis functions (i.e., $\vec{\boldsymbol{\sigma}}_{\mathrm{kernel}} = \vec{\boldsymbol{\sigma}}_j, j = 1, \ldots, N$) and is chosen to be radially symmetric (i.e., $\sigma_{\bullet 1} = \sigma_{\bullet 2} = ... = \sigma_{\bullet d} = \sigma_{\mathrm{kernel}}$).

The form of kernel function $K(\bullet)$ is fixed and known a priori based on domain knowledge. However, there are no general guidelines available for the selection of kernel parameters for hydrologic datasets. In Sect. 4, using bias–variance analysis, we argue that kernel width parameter ($\sigma_{\mathrm{kernel}}$) is one of the most sensitive parameters and has an important role on the performance of the learning machine. This important aspect has received little attention, and here we suggest some techniques for practical selection of $\sigma_{\mathrm{kernel}}$.

If the error term, $\boldsymbol{\varepsilon}$ in Eq. (2) is assumed to normally distributed with zero mean and unknown variance $\sigma_{\varepsilon}^2$, and the input patterns $\{(\boldsymbol{x}_i, y_i), i = 1, ..., N\}$ are generated independently, then the likelihood of the observed dataset can be written as

$$p(\boldsymbol{y}|\boldsymbol{w}, \sigma_{\varepsilon}^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{\varepsilon}^2}} \exp\left\{-\frac{1}{2\sigma_{\varepsilon}^2}\left[(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2\right]\right\} \tag{4}$$

Under a Bayesian perspective, model parameters $\boldsymbol{w}$ and $\sigma_{\varepsilon}^2$ can be estimated by first assigning prior distributions to the parameters and then estimating their posterior distribution using likelihood of the observed data. For RVM, Tipping (2001) proposed a prior conditional distribution of the form

$$p(w_j|\alpha_j) = \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{1}{2}\alpha_j w_j^2\right) = \mathcal{N}\left(0, \frac{1}{\alpha_j^2}\right), j = 0, \ldots, N \tag{5}$$

for each of the weights, where $\boldsymbol{\alpha} = [\alpha_0, \ldots, \alpha_N]^{\mathrm{T}}$ are called a hyperparameters, and a uniform uninformative prior over logarithmic scale for $\sigma_{\varepsilon}^2$ and $\alpha_j$. The choice of zero-mean-Gaussian-prior for weights expresses a preference for smaller weights, and hence a smoother estimate of the function $f(\mathbf{x}, \boldsymbol{w})$. Another advantage of this formulation is that during the process of learning, many of the hyperparameters $\alpha_j$ approach infinity, so the corresponding weights $w_j$ tend to be delta functions centered at zero, and are thus deleted from Eq. (1) leading to sparseness. The remaining patterns corresponding to non-zero weights are only deemed to be relevant for function approximation, and hence the learning machine is known as RVM.

An analytical expression for the posterior distribution of model parameters, $p(\boldsymbol{w}, \boldsymbol{\alpha}, \sigma_{\varepsilon}^2|\boldsymbol{y})$ is not available. However, it can be decomposed into two components as

$$p(\boldsymbol{w}, \boldsymbol{\alpha}, \sigma_{\varepsilon}^2|\boldsymbol{y}) = p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma_{\varepsilon}^2) p(\boldsymbol{\alpha}, \sigma_{\varepsilon}^2|\boldsymbol{y}) \tag{6}$$

The first term on the right hand side of Eq. (6) is the posterior probability of the weight $\boldsymbol{w}$ given $\sigma_{\varepsilon}^2$ and $\boldsymbol{\alpha}$, and is normally distributed.

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2) = \mathcal{N}(\boldsymbol{\mu}_w, \Sigma_w) \tag{7}$$

where mean and covariance are respectively

$$\boldsymbol{\mu}_{\boldsymbol{w}} = \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{-2} \Sigma_w \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{y} \tag{8}$$

$$\Sigma_w = \left(\sigma_{\varepsilon}^{-2} \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} + \boldsymbol{A}\right)^{-1} \tag{9}$$

with $\boldsymbol{A} = \text{diag}(\alpha_0, \alpha_1, ..., \alpha_N)$.

The second term on the right hand side of Eq. (6) is the posterior probability of $\boldsymbol{\alpha}$ and $\sigma_\varepsilon^2$. The estimation of this probability is analytically intractable, and is approximated by delta function at its mode i.e.,

$$p(\boldsymbol{\alpha}, \sigma_\varepsilon^2|\boldsymbol{y}) \sim \delta(\boldsymbol{\alpha}_{\max}, \sigma^2_{\varepsilon\max}) \tag{10}$$

where $\sigma^2_{\varepsilon\max}$ and $\boldsymbol{\alpha}_{\max}$ are values for which $p(\boldsymbol{\alpha}, \sigma_\varepsilon^2|\boldsymbol{y})$ reaches it maximum value. It turns out that maximizing $p(\boldsymbol{\alpha}, \sigma_\varepsilon^2|\boldsymbol{y})$ is equivalent to maximizing the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma_\varepsilon^2)$ which is given by

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\sigma}_\varepsilon^2) &= \int p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\sigma}_\varepsilon^2) p(\boldsymbol{w}|\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{w} \\ &= (2\pi)^{-N/2} \left|\boldsymbol{\sigma}_\varepsilon^2 \boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\right|^{1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{\sigma}_\varepsilon^2 \boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\right)^{-1}\boldsymbol{y}\right\} \end{aligned} \tag{11}$$

Closed form solutions of $\sigma^2_{\varepsilon\max}$ and $\boldsymbol{\alpha}_{\max}$ are not available. Tipping (2001) proposed an iterative estimation method based on a type of automatic relevance determination (MacKay 1994; Neal 1996). After convergence, the learning algorithm provides hyperparameter posterior mode ($\sigma^2_{\varepsilon\max}$ and $\boldsymbol{\alpha}_{\max}$) and posterior distributions of weights conditioned on $\sigma^2_{\varepsilon\max}$ and $\boldsymbol{\alpha}_{\max}$. The distribution of dependent variable $y^*$ at new location $\boldsymbol{x}^*$ can be obtained by

$$\begin{aligned} p(y^*|\boldsymbol{y}) &= \int p(y^*|\boldsymbol{w}, \sigma^2_{\varepsilon\max}) p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}_{\max}, \sigma^2_{\varepsilon\max}) \mathrm{d}\boldsymbol{w} \\ &= \mathcal{N}\left(\mu_{y^*}, \sigma^2_{y^*}\right) \end{aligned} \tag{12}$$

where the mean and variance of the predicted value are, respectively,

$$\mu_{y^*} = \boldsymbol{\mu}_{\boldsymbol{w}}^{\mathrm{T}} \boldsymbol{\Phi}(\boldsymbol{x}^*) \tag{13}$$

$$\sigma^2_{y^*} = \sigma^2_{\varepsilon\max} + \boldsymbol{\Phi}(\boldsymbol{x}^*)^{\mathrm{T}} \Sigma_w \boldsymbol{\Phi}(\boldsymbol{x}^*) \tag{14}$$

The variance of the predicted value (Eq. (14)) is the sum of the variance associated with noise in the training data and uncertainty associated in prediction of weights.

## 3 Data used in this study

We first used synthetic data generated form sinc function for evaluation of our proposed strategies. The sinc function sinc $(x) = \sin(x)/x$ has been a popular choice to illustrate the performance of kernel methods (Chalimourda et al. 2004; Cherkassky and Ma 2004; Vapnik 1998). In this work samples were randomly generated from the range $x \in [-10,10]$ and an independent Gaussian noise was added to each data point for different experiments. This was followed by a second data set of electrical conductivity measurements (surrogate for hydrologic conductivity) obtained from Zhang (1990). Surface soil samples were initially collected at 100 locations over a 1,000 by 1,000 m agricultural field near Marana, Arizona, with some additional samples collected randomly at various locations (Fig. 1). A total of 129 measurements of electrical conductivity were available for this site. One measurement was unusually large in comparison to other measurements and was not considered further in this study. The main reason of choosing a relatively small dataset is that its results are amenable to visual interpretation. Moreover, spatial interpolation of hydrologic data is often challenging due to the sparseness and presence of noisy samples in the data. Spatial interpolation of hydrologic data can be a benchmark test for statistical learning algorithms. While several other high-dimensional large data sets were also investigated, we report results from only these two data sets for clarity and brevity.

## 4 Bias–variance analysis

A key tool for understanding the effect of parameter selection in a machine-learning algorithm is the bias–variance decomposition of the approximation error. In recent years the use of this method for obtaining theoretical insights into machine-learning algorithms has grown rapidly (Berardi and Zhang 2003; Meyer et al.

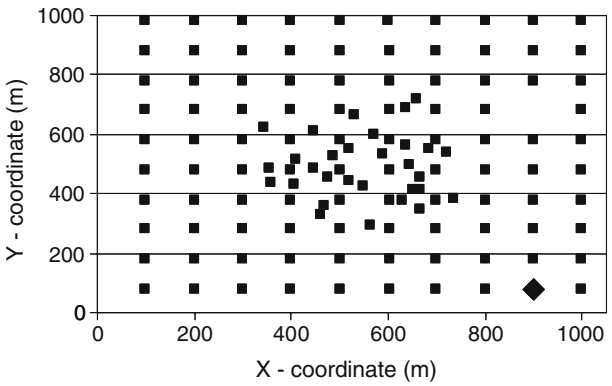


**Fig. 1** Locations of electrical conductivity measurements obtained from Zhang (1990). The *squares* show the measurement locations, while the *rhombus* shows the location of outlier in the dataset

2003; Stankovic et al. 2002; Buciu et al. 2002; Snijder et al. 1998; Twomey and Smith 1998; Geman et al. 1992) but does not appear to have been used in the context of RVMs. In bias–variance decomposition, bias measures the accuracy of the estimate given by a learning machine while variance measures the precision or specificity of the estimate. These terms are not independent but obey a form of conservation law. On one hand, a learning machine with too many degrees of freedom, will over-fit the data resulting in low bias but high variance. On the other hand, a model that contains very few degrees of freedom will not be flexible enough to approximate important features in the data resulting in high bias but low variance. The goal of learning is then to strike a right balance by identifying an appropriate tradeoff between these two terms. This tradeoff is well known in statistical literature as the bias–variance tradeoff.

For function approximation problem, the mean-square error of the learning machine can be decomposed as (Geman et al. 1992)

$$\begin{aligned} E\left[(\boldsymbol{y}-\hat{\boldsymbol{y}})^2\right] &= E\left[(\boldsymbol{y}-E[\boldsymbol{y}])^2\right] \\ &\quad + E\left[(\hat{\boldsymbol{y}}-E[\hat{\boldsymbol{y}}])^2\right] + (E[\boldsymbol{y}]-E[\hat{\boldsymbol{y}}])^2 \\ &= \text{Noise}(\boldsymbol{y}) + \text{Variance}(\hat{\boldsymbol{y}}) + \text{Bias}^2(\hat{\boldsymbol{y}}) \quad (15) \end{aligned}$$

The first term on the right hand side is the variance of the intrinsic noise present in the data. This term forms a lower bound on the error that can be obtained by any learning algorithm. The second term is the error in estimation due to random variation in selecting finite training samples, while the third term is the error due to mismatch between the target and approximating functions.

Although there are various advantages of studying bias–variance decomposition of a learning machine, there are certain limitations that arise when applying this method to real data sets. To be able to estimate the noise, variance, and bias for a particular problem (see Eq. (15)), we need to know the actual function being learned. This is not available for most real-world problems. To overcome this hurdle several alternatives have been suggested in the literature (Bauer and Kohavi 1999; Breiman 1998; Kohavi and Wolpert 1996). However, these approaches are marked by high computational cost and subjectivity in their design. Since the main purpose of studying bias–variance in this work is to understand the effect of kernel parameters on the performance of an RVM, we have used synthetic data (sinc dataset) to overcome this problem.

Following (Valentini and Dietterich 2004), 400 different small training sets of 50 samples each were generated from the sinc function followed by large test set of 1,000 samples. Further, to each training set a Gaussian noise with a standard deviation of 0.2 was added. The main idea behind selecting small training sets and a much larger test set is that small training sets show bias and variance more clearly, whereas a large test set gives reliable estimates of bias and variance. For each training set, RVMs were trained by varying the kernel width $\sigma_{\text{kernel}}$. Bias and variance decomposition of the error was then evaluated using test set as given by Eq. (15). Figure 2 depicts the estimated mean-square error, bias, and variance for different values of $\sigma_{\text{kernel}}$.

It is evident from the figure that the generalization capability of RVM is highly sensitive to the proper choice of $\sigma_{\text{kernel}}$. In particular, for high values of $\sigma_{\text{kernel}}$ the bias is very high and so is the mean-square error. Lowering the value of $\sigma_{\text{kernel}}$ results in sudden drop in bias, which then stabilizes for a range of $\sigma_{\text{kernel}}$ values before beginning to increase again. Variance, on the other hand, has maximum contribution to mean-square error for smaller values of $\sigma_{\text{kernel}}$ as expected. It generally decreases with increase in $\sigma_{\text{kernel}}$ and stabilizes for high values of $\sigma_{\text{kernel}}$. Interestingly, there exists a common region in which both bias and variance take low values. This region is marked by a sudden drop in mean-square error. In summary, the bias–variance analysis demonstrates that, for RVMs, there exists a band of suitable values of $\sigma_{\text{kernel}}$. Outside this range, an RVM will likely have poor generalization performance. The aim of model selection should therefore be to identify this region for appropriate selection of $\sigma_{\text{kernel}}$.

## 5 Methods of kernel parameter selection

In this section, we first show that the method of estimating kernel parameters by maximizing the likelihood function (Eq. (11)) may not yield satisfactory results. Subsequently, other new methods are presented to address this issue.

### 5.1 Maximizing likelihood

Tipping (2001) proposed a method for estimating kernel parameters of RVM by maximizing the likelihood function (Eq. (11)). This can be achieved by taking the derivative of log-likelihood with respect to kernel parameter $\sigma_{\text{kernel}}$, and performing gradient-based local optimization (see Tipping (2001), p. 235 for details). This approach was adopted by Khalil et al. (2005b) for developing an adaptive RVM for real-time management of water releases.

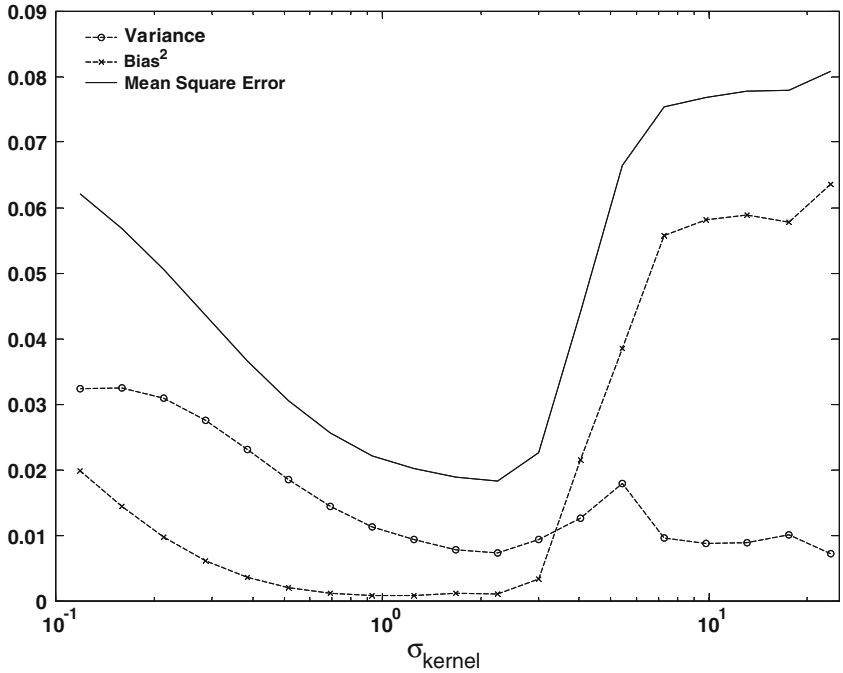


**Fig. 2** Bias–variance decomposition of error in the estimation of sinc function by RVM as a function of kernel width $\sigma_{\text{kernel}}$

It is to be noted that if kernel parameter $\sigma_{\text{kernel}}$ becomes a free variable to be determined during the learning process, the learning machine (given by Eq. (1)) will become nonlinear in the unknown parameters. Estimating parameters in such a context will need a rigorous nonlinear optimization algorithm. Tipping (2001) and Quinonero-Candela and Hansen (2002) reported significant improvement in the performance of an RVM by this technique, but emphasized the difficulties in its implementation and suggested cross-validation as a better approach for practical problems.

Besides the computational costs involved with the implementation of this method, we found that for the experimental data set at the Marana site, a naïve optimization of log-likelihood may lead to severe over-fitting. In fact, Tipping (2001) cautions that maximizing the likelihood is not guaranteed to yield optimum value of kernel parameter $\sigma_{\text{kernel}}$ in terms of model accuracy. Figure 3 shows the value of rescaled log-likelihood and number of statistically relevant vectors identified by the RVM model, as a function of $\sigma_{\text{kernel}}$, developed for the approximation of electrical conductivity values at the Marana site. Figure 4 illustrates the electrical conductivity surface generated by RVM for the value of $\sigma_{\text{kernel}}$ corresponding to maximum likelihood. The generated surface is very complex with many sharp peaks and valleys. This is an example of the RVM model over-fitting the training data. Further, it may be noted that 112 out of 128 vectors were chosen as being relevant in this RVM application. Thus, a complex model has been adopted with no advantage of sparsity.

From these figures, it is seen that smaller values of $\sigma_{\text{kernel}}$ imply that a highly complex model is needed to fit the data. Such models typically over-fit the training data resulting in higher values of likelihood function, but result in poor generalization performance. Therefore, in some cases, estimation of kernel parameter by maximizing the likelihood function alone may be undesirable.

### 5.2 Derivation of prediction risk by leave-one-out cross-validation

Estimation of parameters by re-sampling is by far the most popular method utilized in learning problems (Scholkopf and Smola 2002; Hastie et al. 2001; Haykin 1999; Cherkassky and Mulier 1998). The basic approach is to partition the available data into two sets; a training set and a validation set. The model is trained using the first set, and the validation error is measured using the second set. The validation error gives an estimate of the prediction risk or generalization error of the model. The parameters for which the generalization error is minimized are selected as best parameters. The main drawback of this approach is that it assumes that both training and validation sets are representative of the entire data. This holds true only for large data sets. When the number of samples is small, choice of partitions of training and validation data sets have an impact on the estimate of prediction risk.

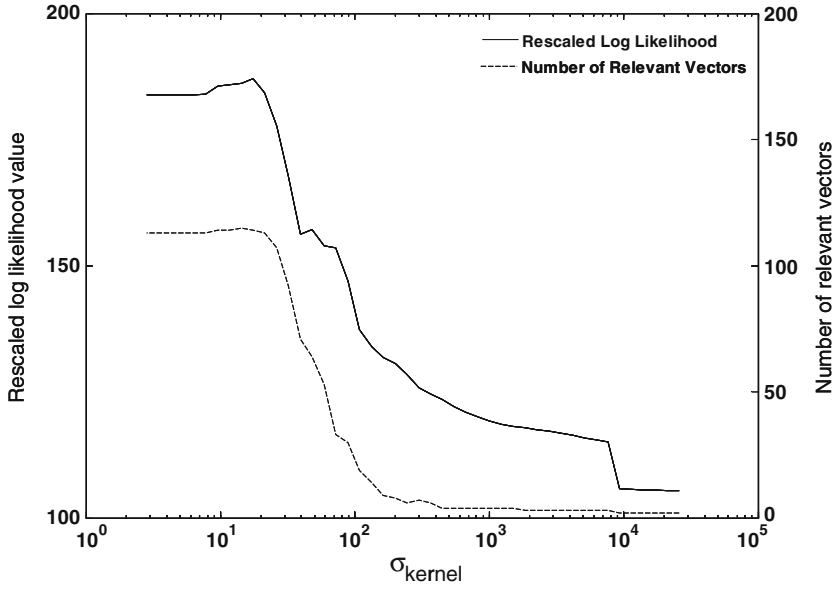


**Fig. 3** Plot showing the value of rescaled log likelihood and number of vectors deemed relevant by the RVM model developed for the approximation of electrical conductivity values at the Marana site with varying kernel width $\sigma_{\text{kernel}}$

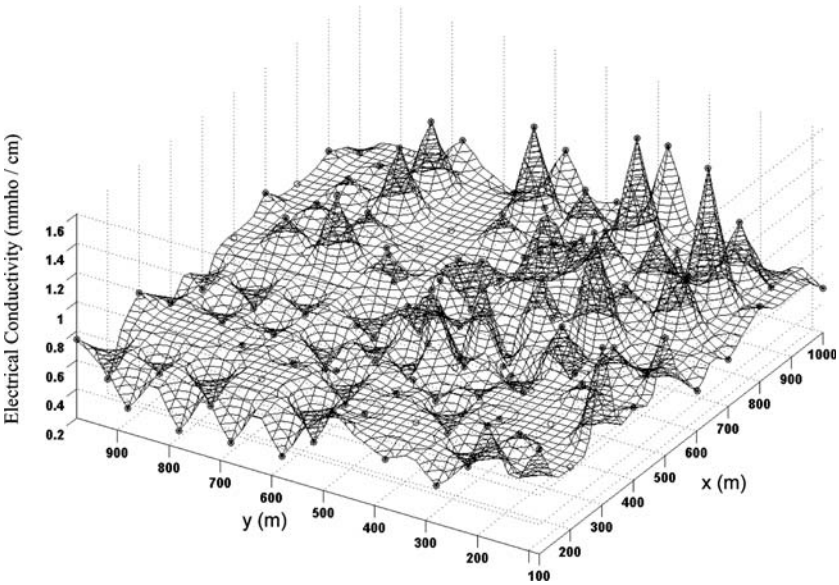


**Fig. 4** Performance of the RVM method for the approximation of electrical conductivity values at the Marana site with kernel width $\sigma_{\text{kernel}} =$ 20 as obtained from maximizing the likelihood function. One hundred and twelve vectors (shown as *solid circles*) out of a total of 128 vectors were considered statistically relevant by the RVM model

Cross-validation provides a strategy so that estimate of prediction risk is invariant to a particular partitioning of data. This method of parameter estimation has gained wide popularity in hydrologic literature (Khalil et al. 2005a; Asefa et al. 2004). One practical approach to cross-validation is *k*-fold cross-validation. In this approach available data is divided into *k* disjoint sets of approximately equal sizes. The model is trained on all the sets except for one and validation error is measured on the set left out. The procedure is repeated for a total of *k* trials, each time using a different subset for validation. Average error under validation over all trials of the experiment is used to assess the generalization performance of the model. The limitation of the *k*-fold cross-validation method is the high computational cost and uncertainty associated in the choice of *k*. Leave-one-out is a special case of *k*-fold cross-validation, where *k* equals the available number of input patterns *N*. Luntz and Brailovsky (1969) proved that leave-one-out is an almost unbiased estimator of prediction risk. However, leave-one-out estimate of prediction risk can yield a high variance.

For linear learning machines, it is possible to compute an analytical expression for leave-one-out estimate of prediction risk ($R_{\text{LOO}}$ ). This has a significant computational advantage over other re-sampling approaches in vogue in the literature for selection of parameters in RVM (Khalil et al. 2005a; Tipping 2001). The analytical expression for $R_{\text{LOO}}$ is derived next for RVMs.

The RVM is linear in parameter $\boldsymbol{w}$, once the design matrix $\boldsymbol{\Phi}$ is fixed (Eq. (1)). Therefore, the function $f(\boldsymbol{x}, \boldsymbol{w})$ learned from RVM obeys the superposition principle. Further, Tipping (2004) shows that estimation of weights in RVM formulation is identical to penalized least square estimation. It then follows from duality in least square problems (Strang 2006) that there always exists a projection matrix $\boldsymbol{P}$ that projects the observed output $\boldsymbol{y}$ into its estimate $\hat{\boldsymbol{y}}$ which in turns lies in the column space of design matrix $\boldsymbol{\Phi}$, so that

$$\hat{\boldsymbol{y}} = f(\mathbf{x}; \boldsymbol{w}) = \boldsymbol{\Phi}\boldsymbol{w} = \boldsymbol{P}\boldsymbol{y} \tag{16}$$

The projection matrix $\boldsymbol{P}$ is a $N \times N$ matrix often called as 'hat matrix' or 'smoothing matrix'. From Eqs. (1), (8) and (13)

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}\boldsymbol{w} = \boldsymbol{\Phi}\boldsymbol{\mu}_{\boldsymbol{w}} = \sigma_{\varepsilon}^{-2}\boldsymbol{\Phi}\Sigma_{w}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y} \tag{17}$$

Comparing Eqs. (16) and (17) the projection matrix $\boldsymbol{P}$ for RVM learning is given by

$$\boldsymbol{P}_{N\times N} = \sigma_{\varepsilon}^{-2}\boldsymbol{\Phi}\Sigma_{w}\boldsymbol{\Phi}^{\mathrm{T}} \tag{18}$$

Further, let the $j$ th pattern $\boldsymbol{x}_j$ be left out of training set during leave-one-out cross-validation. Then the prediction for $\boldsymbol{x}_j$ during validation operation is given as (Cherkassky and Mulier 1998)

$$\hat{y}_j = \frac{1}{1-P_{jj}}\sum_{\substack{i=1\\ i\neq j}}^{N} P_{ji}y_i \tag{19}$$

where $P_{ji}$ is the $j$ th row and $i$ th column element of the projection matrix $\boldsymbol{P}$. The square error in the prediction of $\boldsymbol{x}_j$ can be computed by

$$\begin{aligned}\left(y_j - \hat{y}_j\right)^2 &= \left(y_j - \frac{1}{1-P_{jj}}\sum_{\substack{i=1\\ i\neq j}}^{N} P_{ji}y_i\right)^2 \\ &= \left(\frac{y_j - \sum_{i=1}^{N} P_{ji}y_i}{1-P_{jj}}\right)^2 = \left(\frac{y_j - \hat{y}_j}{1-P_{jj}}\right)^2\end{aligned} \tag{20}$$

Hence, the leave-one-out estimate of prediction risk is

$$R_{\mathrm{LOO}} = \frac{1}{N}\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{y_j - \hat{y}_j}{1-P_{jj}}\right)^2 \tag{21}$$

Figures 5 and 6 illustrate the estimation of prediction risk by analytical leave-one-out cross-validation for sinc data and Marana site data, respectively. For sinc data (Fig. 5), leave-one-out cross-validation can successfully identify the region of sudden drop in prediction risk as obtained during bias–variance analysis. The prediction risk then remains approximately constant for a wide range of $\sigma_{\mathrm{kernel}}$ values before increasing again. Increase in prediction risk for smaller values of $\sigma_{\mathrm{kernel}}$ occurs at a slower rate when compared to the results of bias–variance analysis. For Marana site data (Fig. 6), there are some local minima in the estimate of prediction risk. These may be attributed to high variance that is associated with leave-one-out cross-validation. However, unlike log-likelihood maximization, the leave-one-out estimate does not show monotonic behavior with decrease in values of $\sigma_{\mathrm{kernel}}$.

### 5.3 Analytical model selection

Under regularization framework, it is widely accepted that for a given sample size, there exists a model of optimal complexity corresponding to the smallest prediction risk (Hastie et al. 2001; Haykin 1999; Cherkassky and Mulier 1998). Analytical model selection criterion uses analytical estimates of the prediction risk as a function of training error [empirical risk ($R_{\mathrm{empirical}}$)] and a penalty term based on some measure of model complexity. These estimates are grouped into two categories: classical estimates that are based on asymptotic analysis (as sample size $N \rightarrow \infty$) and the structural risk minimization (SRM) method from statistical learning theory that is based on non-asymptotic analysis (Cherkassky and Mulier 1998).

#### *5.3.1 Asymptotic analysis*

In classical estimates, the forms of prediction risk vary depending on the class of approximating functions supported by the learning machine. For linear models, a number of these estimates are available such as Akakie information criterion (AIC), Bayes information criterion (BIC), minimum description length, etc. Among them, AIC and BIC are very popular in hydrologic literature (Xu and Li 2002; Gyasi-Agyei 2001; Honjo and Kashiwagi 1999; Knotters and De Gooijer 1999; Mutua 1994; Gregory et al. 1992). The general form of the classical estimate of prediction risk for a linear model can be written as

$$\text{prediction risk} = \Psi\left(\frac{\mathrm{dof}}{N}\right)\cdot R_{\mathrm{empirical}} \tag{22}$$

where $\Psi$ is a monotonic increasing function of the ratio of degrees of freedom (dof) and training sample

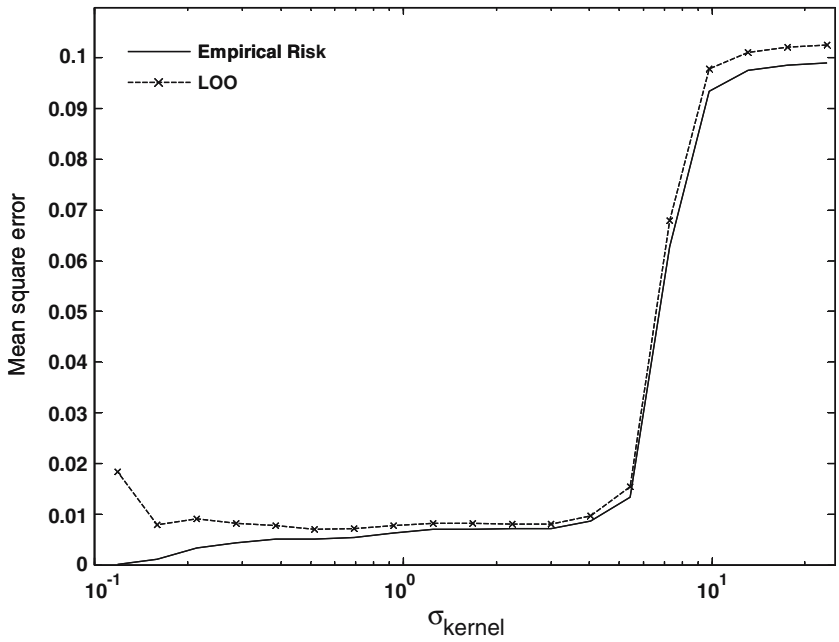


**Fig. 5** Estimation of prediction risk using analytical form of leave-one-out cross-validation (labeled LOO) for the approximation of sinc function by RVM with varying kernel width $\sigma_{\text{kernel}}$

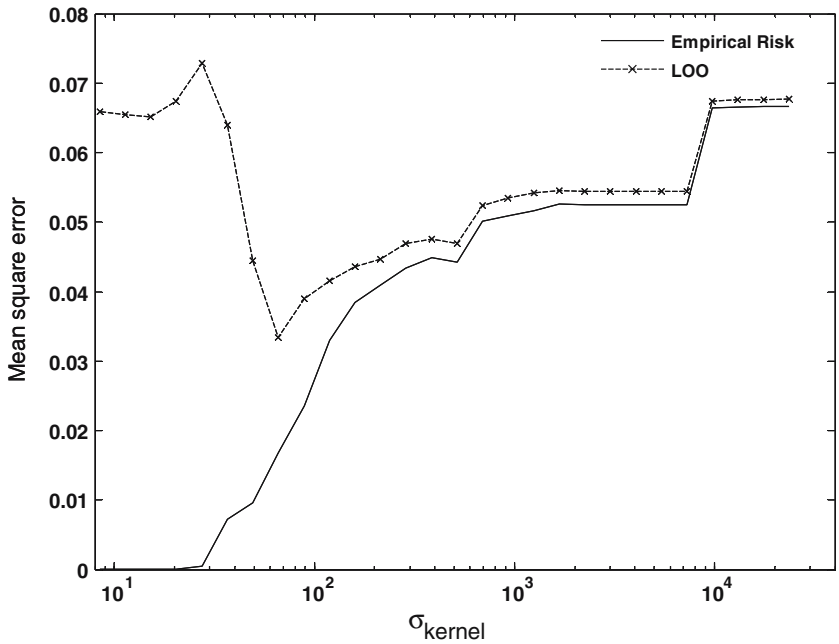


**Fig. 6** Estimation of prediction risk using analytical form of leave-one-out cross-validation (labeled LOO) for the approximation of electrical conductivity values at the Marana site by RVM with varying kernel width $\sigma_{\text{kernel}}$

size ($N$). For penalized linear models, the effective degrees of freedom can be determined from the projection matrix $\boldsymbol{P}$ (Hastie et al. 2001; Cherkassky and Mulier 1998). To the best of our knowledge, this has not been investigated for RVMs. From linear algebra (Strang 2006), it is well known that the eigen values of $\boldsymbol{P}$ for a penalized linear model are in the range [0,1]. The effective degrees of freedom is given by the number of eigen values that are proximal to unity. Determining eigen values of $\boldsymbol{P}$ is computationally intensive, hence approximations are made to determine the number of large eigen values, and hence the dof. One popular approximation (Cherkassky and Ma 2003; Cherkassky and Mulier 1998; Hastie et al. 2001) is

$$\text{dof} \approx \text{trace}\left(\boldsymbol{P}\boldsymbol{P}^{\text{T}}\right) \tag{23}$$

Thus if we know the dof, we can estimate prediction risk based on a particular choice of $\Psi$ using Eq. (22). For demonstration, we have used BIC or Schwarz'

criteria (Schwarz 1978). BIC estimate of prediction risk is given by

$$\text{prediction risk} = -2 \cdot \text{log-likelihood} + \log(N) \cdot \text{dof} \tag{24}$$

The estimated prediction risk obtained from BIC for sinc data and Marana site data are shown in Figs. 7 and 8, respectively. BIC estimate of sinc data (Fig. 7) could not only capture the region of sudden drop in prediction risk but closely follows the pattern obtained from bias–variance analysis. For Marana site data, the estimate of prediction risk (Fig. 8) shows a suitable range of $\sigma_{\text{kernel}}$, penalizing very simple as well as very complex models.

### 5.3.2 Structural risk minimization

Statistical learning theory provides a very general and conceptual framework for complexity control using SRM. Under SRM principle, a set of possible models are arranged in order of increasing complexity. Selection of an optimal model is based on computation of the Vapnik-Chervonenkis (VC) generalization bound that provides an upper bound on prediction risk. For a function approximation problem with $N$ samples, the following VC generalization bound holds with a probability $1 - \Theta$ (Vapnik 1998)

$$\text{prediction risk} \leqslant R_{\text{empirical}} \cdot \left\{ 1 - c\sqrt{a_1 \frac{h\left[\log\left(a_2 N/h\right)+1\right] - \log\left(\Theta/4\right)}{N}} \right\}_+^{-1} \tag{25}$$

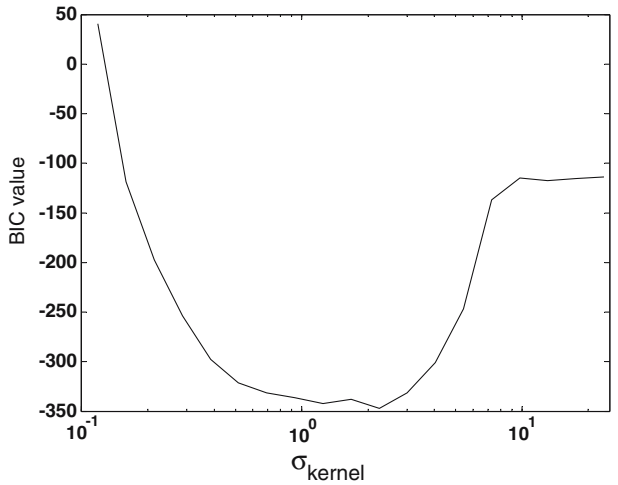


**Fig. 7** Estimation of prediction risk using Bayes information criterion (BIC) for the approximation of sinc function by RVM with varying kernel width $\sigma_{\text{kernel}}$

where $h$ is the VC dimension of the approximating function, and $c$ is a theoretical constant that depends on the choice of loss function. The quantities $a_1$ and $a_2$ are constants with values in the range $0 < a_1 \leq 4$; $0 < a_2 \leq 2$. The values of $a_1$ and $a_2$ depend upon the joint distribution of the input and output variables that is usually unknown. For worst case distribution (discontinuous density function), $a_1$ and $a_2$ have been derived to be 4 and 2 respectively (Vapnik 1995, 1999). For a function approximation problem with square error loss function, empirical studies suggest values of $a_1 = 1$, $a_2 = 1$ and $c = 1$ (Cherkassky et al. 1999; Vapnik 1998). However Cherkassky and Mulier (1998) suggested that $a_1$ and $a_2$ should be altered based on data. The quantity $\Theta$ in Eq. (25) determines the confidence interval to be described shortly.

The main difficulty in applying VC bound is to estimate VC dimension ($h$) of the approximating function. Unfortunately, it is not possible to obtain an exact analytical estimate of $h$ for most approximating functions including the form given by Eq. (1). To overcome this difficulty, experimental methods of determining $h$ are available (Cherkassky and Mulier 1999; Shao et al. 2000). Most of these methods are too complex to be applied for practical problems. However for penalized linear models like RVM, $h$ can be approximated by effective dof given by Eq. (23). This heuristic estimate of $h$ has been successfully applied in various statistical learning problems (Cherkassky and Ma 2003; Cherkassky et al. 1999; Cherkassky and Mulier 1998) and is used in this study. The upper bound on prediction risk for RVM is therefore given by

$$\text{prediction risk} \leqslant R_{\text{empirical}} \cdot \left\{ 1 - \sqrt{\frac{\text{dof}\left[\log\left(N/\text{dof}\right)+1\right] - \log\left(\Theta/4\right)}{N}} \right\}_+^{-1} \tag{26}$$

The value of confidence interval $\Theta$ should be chosen based on the number of samples. When the number of samples is small, the confidence level is set low, whereas when the number of samples is large, the confidence level is set high. Vapnik (1995) recommended the following rule for choosing the confidence interval, and was used in this study

$$\Theta = \min\left(\frac{4}{\sqrt{N}}, 1\right) \tag{27}$$

The upper bound on prediction risk as estimated by SRM principle for sinc data and Marana site data are

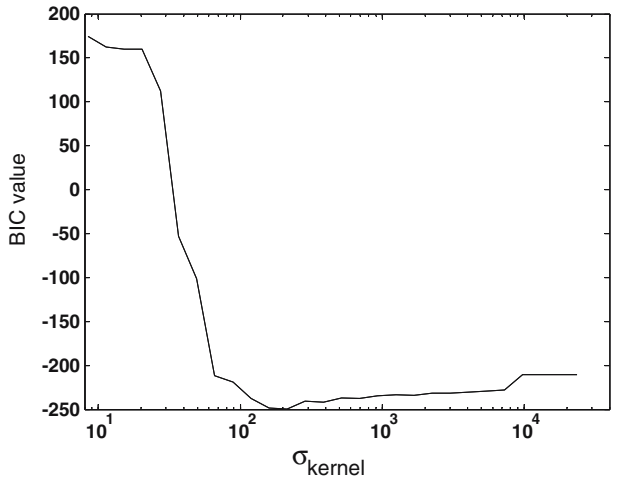


**Fig. 8** Estimation of prediction risk using Bayes information criterion (BIC) for the approximation of electrical conductivity values at the Marana site by RVM with varying kernel width $\sigma_{\text{kernel}}$

shown in Figs. 9 and 10, respectively. The results from sinc data suggest that for the value of $a_1 = 1$ and $a_2 = 1$ as suggested in literature, the estimated prediction risk could capture the region of sudden drop, but increases only for very small values of $\sigma_{\text{kernel}}$. However, we found that the estimated prediction risk closely follows the results of bias–variance analysis for $a_1 = 2$ and $a_2 = 1$. The results from Marana site data also indicate that SRM principle can be useful in finding values of $\sigma_{\text{kernel}}$ for which RVM model is neither very complex nor very simple. In fact, results obtained from SRM principle are very similar to those obtained from BIC estimate of prediction risk.

### 5.4 Structure of data

The kernel function, used in a model such as Eq. (1), can be visualized as mathematical formulation for the notion of similarity in input space of **x** (Scholkopf and Smola 2002; Vapnik 1998). In the context of Gaussian processes, kernel function is often interpreted as a covariance function that encodes our knowledge of how observations at different points of input space are related (Rasmussen and Williams 2006; Scholkopf and Smola 2002). Since RVMs also belong to the family of Gaussian processes (Tipping 2001), this interpretation can provide a way for selecting kernel width ($\sigma_{\text{kernel}}$). The Euclidean distance in the input space (**x**) beyond which the dependent variable (***y***) ceases to have any significant correlation can provide a reasonable estimate for the value of kernel width. This strategy has been reported to give satisfactory results in the context of SVM regression (Asefa et al. 2004).

The correlogram for sinc data and Marana site data are shown in Figs. 11 and 12, respectively. For sinc data, the correlation goes to zero at approximately 2.5 separation distance. This value is close to the optimum value as obtained from bias–variance analysis. It can be seen from Fig. 12 that for Marana site data, the estimated value of $\sigma_{\text{kernel}}$ from correlogram and optimum RBF kernel obtained from Bayes information criterion and SRM principle are in close agreement.

In an attempt to explain the regularization capability of kernel methods, Smola et al. (1998) argue that kernels act as a filter in frequency domain. For example an RBF kernel with a large kernel width will act as a low-pass filter in frequency domain, attenuating higher order frequencies and thus resulting in a smooth function. Alternatively, an RBF kernel with small kernel width will retain most of the higher order frequencies leading to an approximation of a complex function by the learning machine. This interpretation provides yet another way of determining kernel width. It is widely acknowledged in the field of signal processing that the frequency of optimal filter should match the frequency distribution of the signal to be constructed (Scholkopf and Smola 2002). Thus, if we know the distribution of data in frequency domain we can construct an appropriate kernel such that it filters out spurious frequencies of the data.

Power spectrum of sinc data and Marana site data along with the power spectrum of the optimal kernel as obtained from previous analyses are shown in Figs. 13 and 14, respectively. It is evident from the figure that the chosen kernels are effective in filtering out low power high frequency signals. These signals can be attributed to the noise present in data. This interpretation of kernel as filters in frequency domain can also be used for designing anisotropic kernels. Figure 14 shows the power spectra in two orthogonal directions (X and Y) for the Marana site data. The similarity in these figures suggests that the kernel function should be isotropic in this case.

## 6 Results

The procedures described in Sect. 5 were tried on many datasets of different size and dimensionality, and the RVM performance was found to be qualitatively similar. For brevity, the values of $\sigma_{\text{kernel}}$ obtained for electrical conductivity measurements at Marana site along with sinc data are presented in Table 1. It is evident from the tabulated results that there exists a range of appropriate values for RVM kernel width $\sigma_{\text{kernel}}$. This is in agreement with the findings of Wang

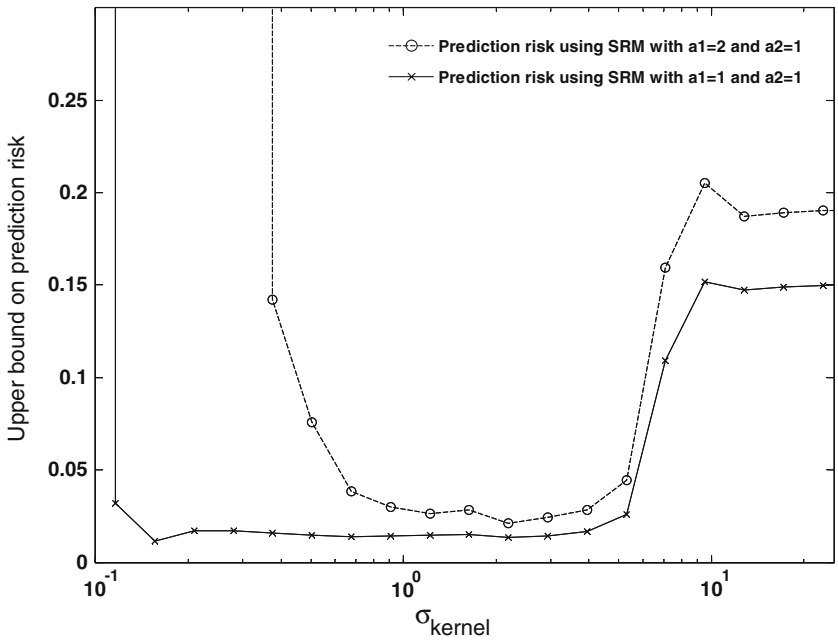


**Fig. 9** Estimation of upper bound on prediction risk using structural risk minimization (SRM) principle for the approximation of sinc function by RVM with varying kernel width $\sigma_{\text{kernel}}$. The bound holds with probability 0.6

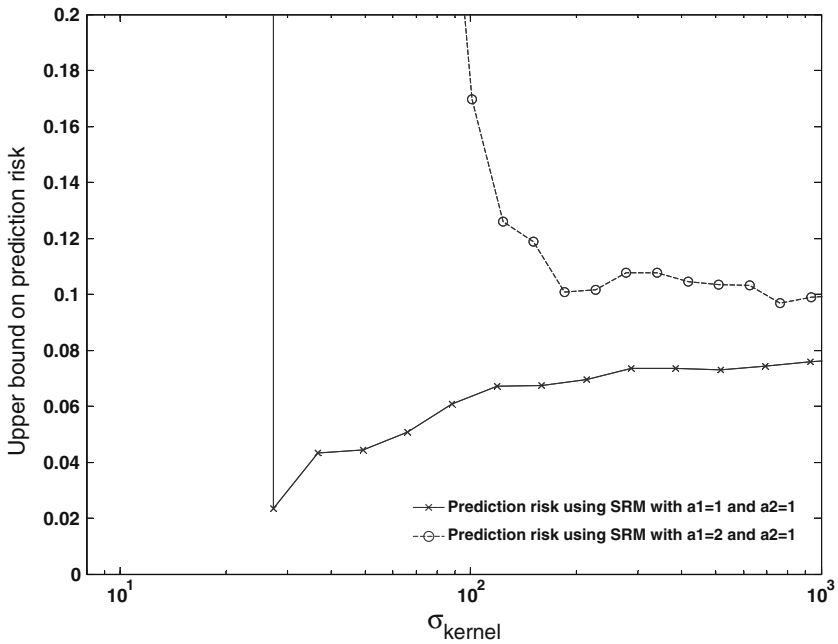


**Fig. 10** Estimation of upper bound on prediction risk using structural risk minimization (SRM) principle for the approximation of electrical conductivity values at the Marana site by RVM with varying kernel width $\sigma_{\text{kernel}}$. The bound holds with probability 0.65

et al. (2003) that the generalization performance of kernel methods (SVM in their case) remains stable within a certain range of $\sigma_{\text{kernel}}$ values. It also corroborates the findings of bias–variance analysis discussed in Sect. 4. The results further indicate that the ranges of $\sigma_{\text{kernel}}$ obtained from various methods generally tend to overlap. Based on many different datasets, the performance of BIC and SRM principles in selecting the value of $\sigma_{\text{kernel}}$ appears to be promising.

Figures 15 and 16 illustrate the fitted surface obtained from developed RVM model based on the value of $\sigma_{\text{kernel}}$ selected from above analysis. The RVM performed reasonably well in approximating sinc function (Fig. 15) and selected only 8 out of a total of 100 data points as relevant vectors. For electrical conductivity values at the Marana site, the RVM model considered only 11 vectors out of total 128 vectors as statistically relevant (Fig. 16a). The corre-

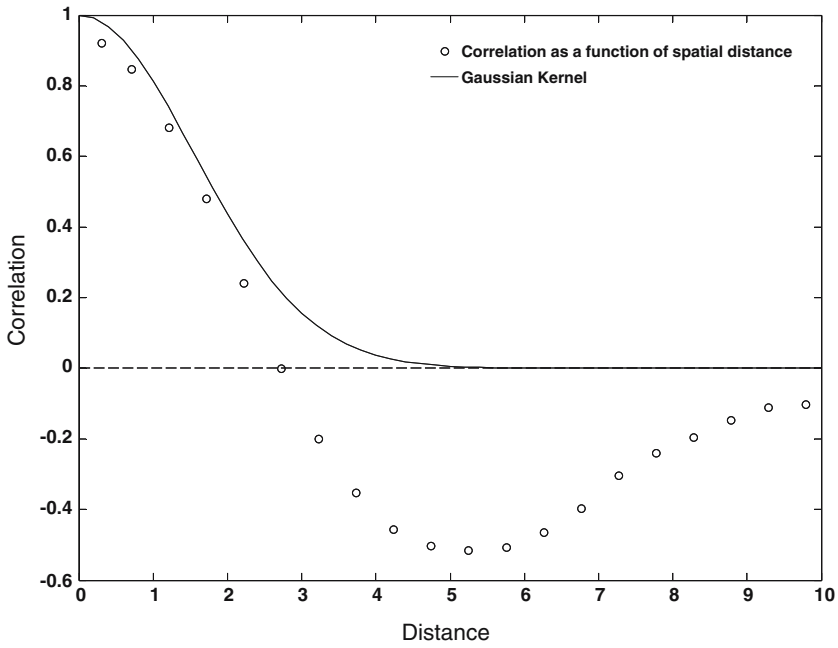


**Fig. 11** Plot showing correlation in the sinc data as a function of distance in input space along with Gaussian kernel obtained from bias–variance analysis ($\sigma_{\text{kernel}} = 2.2$)

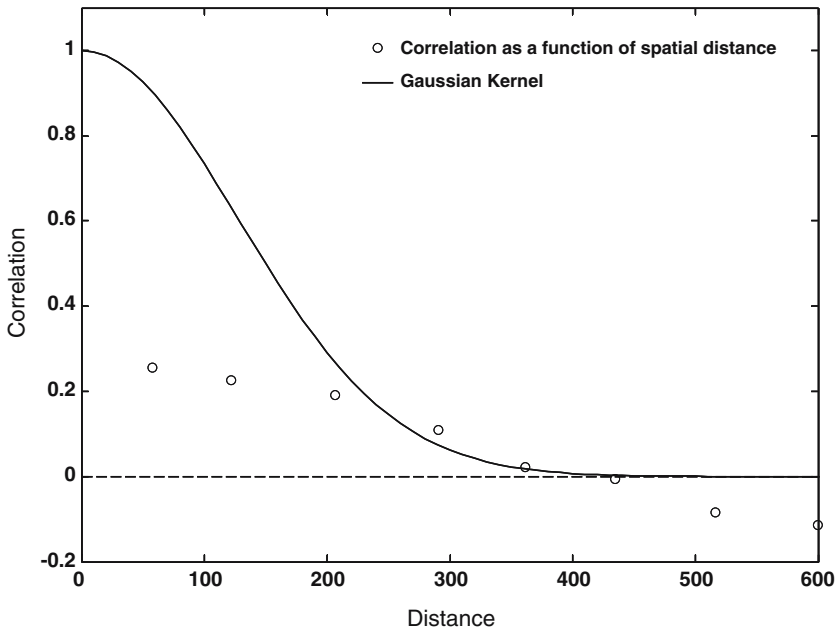


**Fig. 12** Plot showing correlation in the measurements of electrical conductivity values at the Marana site as a function of spatial distance and Gaussian kernel as obtained from BIC and SRM principle ($\sigma_{\text{kernel}} = 180$)

sponding prediction variance surface shown in Fig. 16b is quite revealing. Note that the high variance estimates are associated with locations where the separation between observed values and model prediction is large, as one would expect intuitively. Similarly, the model variances tend to be larger for locations that have little or no data support. It is worth mentioning that the number of data points deemed relevant by RVM depends on the complexity of the function to be approximated. However, for the various datasets used in this study, RVM typically chose less than 10% of available data as relevant.

## 7 Concluding remarks

In this study the importance of kernel parameters in generalization performance of RVMs was established using bias–variance analysis. Following this, several techniques for practical selection of kernel parameters

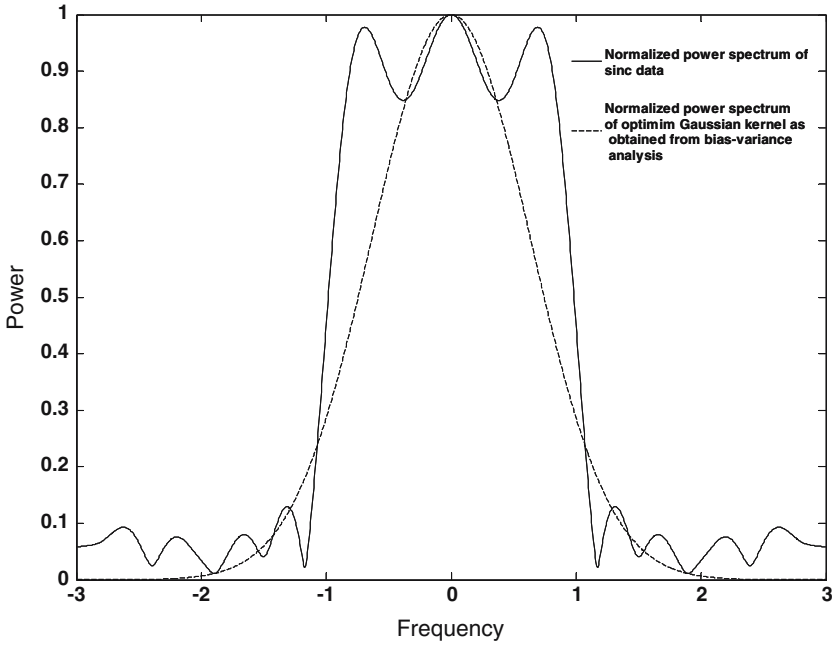


**Fig. 13** Plot showing the normalized power spectrum of sinc data and power spectrum of normalized Gaussian kernel obtained from bias–variance analysis ($\sigma_{\text{kernel}} = 2.2$)

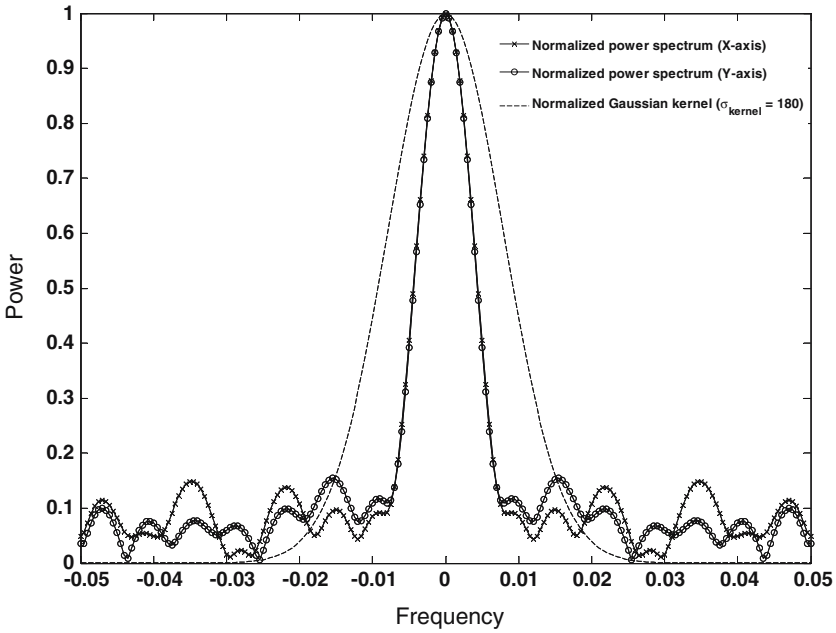


**Fig. 14** Plot showing normalized power spectrum of electrical conductivity measurements at the Marana site for two orthogonal directions (X and Y) along with power spectrum of normalized Gaussian kernel corresponding to optimal $\sigma_{\text{kernel}}$ as obtained from BIC and SRM principle ($\sigma_{\text{kernel}} = 180$)

**Table 1** Range of Gaussian kernel width parameter $\sigma_{\text{kernel}}$ for sinc data and electrical conductivity measurements at the Marana site data obtained from different model selection methods

| S. no. | Method of kernel selection | Range of $\sigma_{\text{kernel}}$ | |
|---|---|---|---|
| | | Sinc data | Marana site data |
| 1 | Bias–variance analysis | 1–3 | – |
| 2 | Analytical leave-one-out cross-validation | <3 | 50–250 |
| 3 | BIC | 1–3 | 150–300 |
| 4 | SRM | 1.5–4 | 150–250 |
| 5 | Correlogram | 1.5–2.5 | 150–200 |
| 6 | Power spectrum | 2–3 | 150–200 |

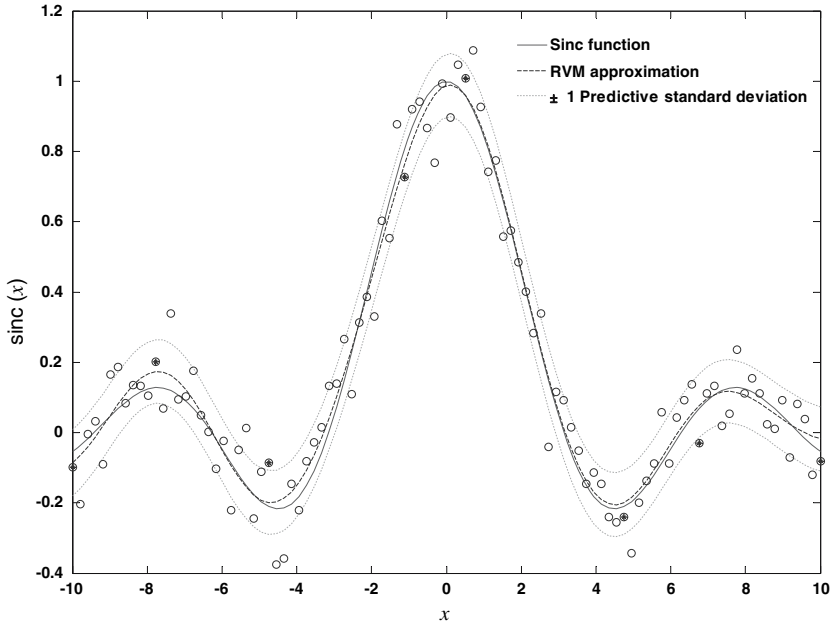


**Fig. 15** Plot showing RVM approximation of sinc function with kernel width $\sigma_{\text{kernel}} = 2.2$. Eight vectors (shown as *solid circles*) out of a total of 100 vectors were considered statistically relevant by the RVM model

were suggested. The effectiveness of the proposed techniques was illustrated through application to synthetic as well as real data sets using Gaussian kernels. For clarity, the results are presented for small data sets only. However, the proposed techniques are general, and the same approach can be easily extended to other forms of kernel functions. Further the proposed techniques are applicable to both small and large hydrologic data sets.

Bias–variance analysis of RVMs suggested that there exists a region of suitable values for RBF kernel width $\sigma_{\text{kernel}}$ over which both bias and variance have low values. Outside this region, both bias and variance increase resulting in poor generalization performance of RVMs.

Maximizing the likelihood function to estimate kernel parameters for RVMs have gained popularity in spite of the fact that its implementation requires solution of highly complex nonlinear optimization problems. Results from this study indicate that in some cases, estimation of kernel parameters by maximizing the likelihood alone may result in over-fitting. Thus, apart from maximizing the likelihood function, other methods as mentioned here should be utilized.

Linear models have been the mainstay of statistical learning for many decades. Consequently, a vast array of tools is available for parameter selection in linear models. In this work, linearity in the formulation of RVMs was used to develop an analytical expression for leave-one-out cross-validation. Further, linearity was exploited to adopt analytical model selection methods for use in RVMs. In particular BIC and Vapnik-Chervonenkis (VC) generalization bounds were demonstrated to yield satisfactory results in selection of $\sigma_{\text{kernel}}$.

It is widely accepted that the form of kernel function and its parameters depends on the structure of data. The results from this study reveal that the correlogram can be used to get a satisfactory range of $\sigma_{\text{kernel}}$ in context of RVMs also. The interpretation of kernels as filters in frequency domain can not only help in understanding the regularization property of kernels, but can also be useful for designing anisotropic kernels.

Based on empirical studies with a variety of hydrologic data sets, we found that the following steps yield reasonable value of $\sigma_{\text{kernel}}$ and are therefore recommend for practical hydrologic applications. (1) Preliminary data analysis: this includes detecting outliers and applying appropriate transformation to the data. (2) Plotting correlogram or power spectrum of the data to decide an approximate range of $\sigma_{\text{kernel}}$: for high dimensional data where, relative importance of input variables in estimating desired output varies, power spectrum can be more useful. (3) Training RVM for the chosen range of $\sigma_{\text{kernel}}$ and estimating prediction risk using BIC or SRM principle. (4) Selecting $\sigma_{\text{kernel}}$ corresponding to minimum value of prediction risk as optimal value for kernel parameter. Our experience suggests that the above steps yield satisfactory value of $\sigma_{\text{kernel}}$ and will make RVMs more amenable to hydrologic applications.

Although RVMs are new to the field of hydrology, they provide a promising alternative to many statistical hydrologic problems because of their excellent generalization properties. Besides this they have the added advantage of probabilistic interpretation that yields

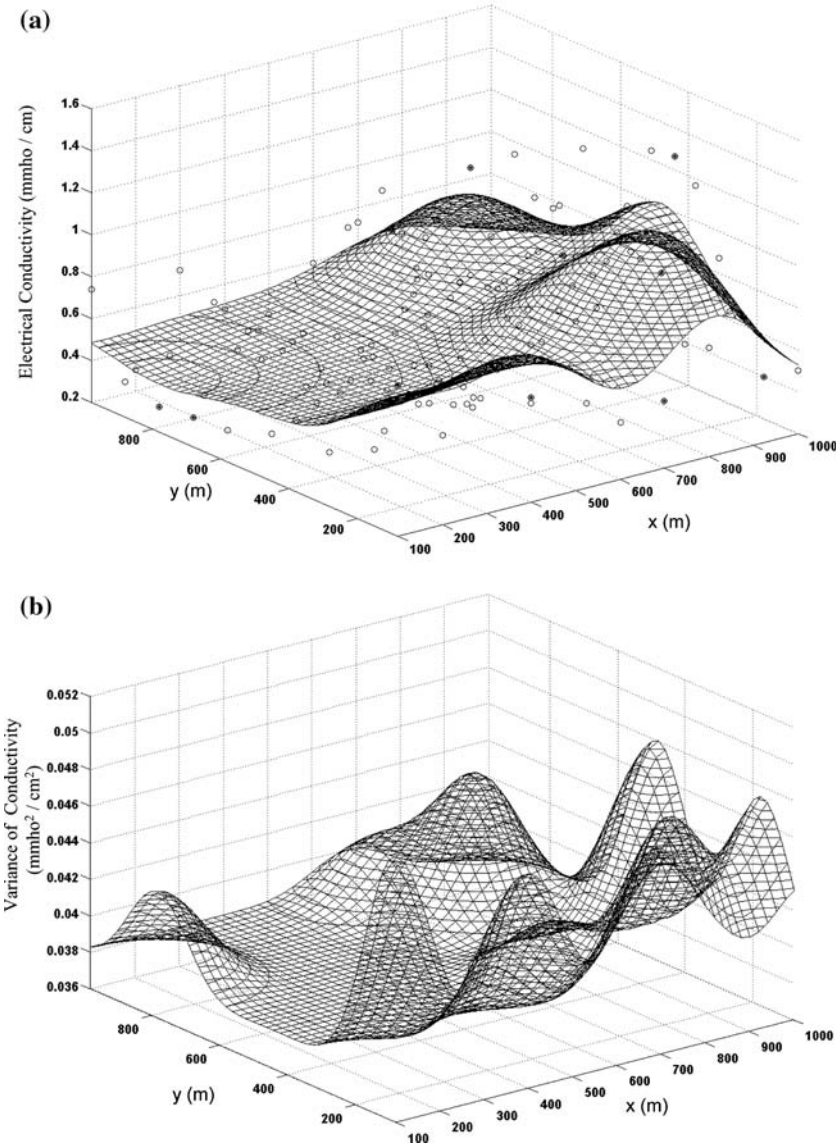


**Fig. 16** Performance of the RVM method for the approximation of electrical conductivity values at the Marana site with kernel width $\sigma_{\text{kernel}} = 180$. **a** The RVM fitted surface (*mesh*) along with measurements in *circles*. Eleven vectors (shown as *solid circles*) out of a total of 128 vectors were considered statistically relevant by the RVM model. **b** The prediction variance surface of the RVM estimates

prediction uncertainty. Therefore, several avenues should be explored in order to make this technique better-suited to a wide variety of hydrologic data. In particular, incorporating and representing physics of the hydrologic system into RVM learning, modifying the RVM formulation to handle flexible priors, and making the algorithm robust against outliers will make RVMs more amenable to hydrologic problems. Further studies by the authors will be directed at addressing some of these issues.

## References

Agarwal A, Triggs B (2006) Recovering 3D human pose from monocular images. IEEE Trans Pattern Anal Mach Intell 28(1):44–58

Asefa T, Kemblowski MW, Urroz G, McKee M, Khalil AF (2004) Support vectors-based groundwater head observation networks design. Water Resour Res 40 (11): W11509, DOI 11510.11029/12004WR003304

Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Mach Learn 36(1–2):105–139

Berardi VL, Zhang GP (2003) An empirical investigation of bias and variance in time series forecasting: modeling considerations and error evaluation. IEEE Trans Neural Netw 14(3):668–679

Berger JO (1985) Statistical decision theory and Bayesian analysis. Springer, Berlin Heidelberg New York, xvi, 617 p

Breiman L (1998) Bias–variance, regularization, instability and stabilization. In: Bishop C (ed) Proceedings of the neural networks and machine learning, Cambridge, UK, pp 27–56

Buciu I, Kotropoulos C, Pitas I (2002) On the stability of support vector machines for face detection. In: Proceedings of the international conference on image processing, Rochester, NY, pp 121–124

Chalimourda A, Scholkopf B, Smola AJ (2004) Experimentally optimal nu in support vector regression for different noise models and parameter settings. Neural Netw 17(1):127–141

Cherkassky V, Ma YQ (2003) Comparison of model selection for regression. Neural Comput 15(7):1691–1714

Cherkassky V, Ma YQ (2004) Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw 17(1):113–126

Cherkassky V, Mulier F (1998) Learning from data: concepts, theory, and methods. Wiley, New York, xviii, 441 pp

Cherkassky V, Mulier F (1999) Vapnik-Chervonenkis (VC) learning theory and its applications. IEEE Trans Neural Netw 10(5):985–987

Cherkassky V, Shao XH, Mulier FM, Vapnik VN (1999) Model complexity control for regression using VC generalization bounds. IEEE Trans Neural Netw 10(5):1075–1089

Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans Electron Comput EC-14(3):326–334

Evgeniou T, Pontil M, Poggio T (2000) Statistical learning theory: a primer. Int J Comput Vis 38(1):9–13

Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias–variance dilemma. Neural Comput 4(1):1–58

Gregory JM, Wigley TML, Jones PD (1992) Determining and interpreting the order of a 2-state Markov-Chain—application to models of daily precipitation. Water Resour Res 28(5):1443–1446

Gyasi-Agyei Y (2001) Modelling diurnal cycles in point rainfall properties. Hydrol Processes 15(4):595–608

Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics. Springer, Berlin Heidelberg New York, xvi, 533 p

Haykin SS (1999) Neural networks: a comprehensive foundation. Prentice Hall, Upper Saddle River, xxi, 842 p

Honjo Y, Kashiwagi N (1999) Matching objective and subjective information in groundwater inverse analysis by Akaike's Bayesian information criterion. Water Resour Res 35(2):435–447

Khalil AF, Almasri MN, McKee M, Kaluarachchi JJ (2005a) Applicability of statistical learning algorithms in groundwater quality modeling. Water Resour Res 41(5): W05010, DOI 05010.01029/02004WR003608

Khalil AF, McKee M, Kemblowski M, Asefa T (2005b) Sparse Bayesian learning machine for real-time management of reservoir releases. Water Resour Res 41(11): W11401, DOI 11410.11029/12004WR003891

Khalil AF, McKee M, Kemblowski M, Asefa T, Bastidas L (2006) Multiobjective analysis of chaotic dynamic systems with sparse learning machines. Adv Water Resour 29(1):72–88

Knotters M, De Gooijer JG (1999) TARSO modeling of water table depths. Water Resour Res 35(3):695–705

Kohavi R, Wolpert DH (1996) Bias plus variance decomposition for zero-one loss functions. In: Proceedings of the 13th international conference of machine learning, Bari, Italy, pp 275–283

Kovvali N, Carin L (2004) Analysis of wideband forward looking synthetic aperture radar for sensing land mines. Radio Sci 39(4):RS4S08, DOI 10.1029/2003RS002967

Lanckriet GRG, Cristianini N, Bartlett P, El Ghaoui L, Jordan MI (2004) Learning the kernel matrix with semidefinite programming. J Mach Learn Res 5:27–72

Luntz A, Brailovsky V (1969) On estimation of characters obtained in statistical procedure of recognition. Techichieskaya Kibernetica, 3 (in Russian)

MacKay DJC (1994) Bayesian methods for backpropagation networks. In: Domany E, van Hemmen JL, Schulten K (eds) Models of neural networks III. Springer, Berlin Heidelberg New York, pp 211–254

Majumder SK, Ghosh N, Gupta PK (2005) Relevance vector machine for optical diagnosis of cancer. Lasers Surg Med 36(4):323–333

Meyer D, Leisch F, Hornik K (2003) The support vector machine under test. Neurocomputing 55(1–2):169–186

Mika S, Ratsch G, Weston J, Scholkopf B, Mullers KR (1999) Fisher discriminant analysis with kernels, neural networks for signal processing IX. In: Proceedings of the 1999 IEEE signal processing society workshop, Madison, WI, USA, pp 41–48

Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B (2001) An introduction to kernel-based learning algorithms. IEEE Trans Neural Netw 12(2):181–201

Mutua FM (1994) The use of the Akaike information criterion in the identification of an optimum flood frequency model. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques 39(3):235–244

Neal RM (1996) Bayesian learning for neural networks. Springer, Berlin Heidelberg New York, xiv, 183 p

Quinonero-Candela J, Hansen LK (2002) Time series prediction based on the relevance vector machine with adaptive kernels. In: IEEE international conference on acoustics, speech and signal processing, Orlando, FL, USA, pp 985–988

Rasmussen CE, Williams CKI. (2006) Gaussian processes for machine learning. Adaptive computation and machine learning. MIT Press, Cambridge, xviii, 248 p

Scholkopf B, Smola AJ (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. Adaptive computation and machine learning. MIT Press, Cambridge, xviii, 626 pp

Scholkopf B, Burges CJC, Smola AJ (eds) (1999) Advances in kernel methods: support vector learning. MIT Press, Cambridge, vii, 376 p

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Shao XH, Cherkassky V, Li W (2000) Measuring the VC-dimension using optimized experimental design. Neural Comput 12(8):1969–1986

Smola AJ, Scholkopf B, Muller KR (1998) The connection between regularization operators and support vector kernels. Neural Netw 11(4):637–649

Snijder E, Babuska R, Verhaegen M (1998) Finding the bias–variance tradeoff during neural network training and its implication on structure selection. In: International conference on neural networks, Anchorage, AK, USA, pp 1613–1618

Stankovic S, Milosavljevic M, Buturovic L, Stankovic M, Stankovic M (2002) Statistical learning: data mining and prediction with applications to medicine and genomics. In: 6th seminar on neural network applications in electrical engineering. NEUREL 2002, Belgrade, Yugoslavia, pp 5–6

Strang G (2006) Linear algebra and its applications. Thomson, Brooks/Cole, Belmont, viii, 487 p

Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 1(3):211–244

Tipping ME (2004) Bayesian inference: an introduction to principles and practice in machine learning. Advanced

lectures on machine learning. Lecture Notes in Artificial Intelligence. Springer, Berlin Heidelberg NewYork, pp 41–62
Twining CJ, Taylor CJ (2003) The use of kernel principal component analysis to model data distributions. Pattern Recognit 36(1):217–227
Twomey JM, Smith AE (1998) Bias and variance of validation methods for function approximation neural networks under conditions of sparse data. IEEE Trans Syst Man Cybernet C Appl Rev 28(3):417–430
Valentini G, Dietterich TG (2004) Bias–variance analysis of support vector machines for the development of SVM-based ensemble methods. J Mach Learn Res 5:725–775
Vapnik VN (1995) The nature of statistical learning theory. Springer, Berlin Heidelberg New York, xv, 188 pp
Vapnik VN (1998) Statistical learning theory. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York, xxiv, 736 pp
Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10(5):988–999
Wang WJ, Xu ZB, Lu WZ, Zhang XY (2003) Determination of the spread parameter in the Gaussian kernel for classification and regression. Neurocomputing 55(3–4):643–663
Wei LY, Yang YY, Nishikawa RM, Wernick MN, Edwards A (2005) Relevance vector machine for automatic detection of clustered microcalcifications. IEEE Trans Med Imaging 24(10):1278–1285
Wu W, Massart DL, deJong S (1997) The kernel PCA algorithms for wide data. 1. Theory and algorithms. Chemometr Intell Lab Syst 36(2):165–172
Xu ZX, Li JY (2002) Short-term inflow forecasting using an artificial neural network model. Hydrol Processes 16(12):2423–2439
Zhang R (1990) Soil variability and geostatistical applications. Ph.D. thesis, The University of Arizona