Frequentist Model Averaging for ϵ-Support Vector Regression

FREQUENTIST MODEL AVERAGING FOR ϵ-SUPPORT VECTOR REGRESSION

BY

FRANCIS KIWON, B.S.F.S.

A thesis
submitted to the Department of Mathematics & Statistics
and the School of Graduate Studies
of McMaster University
in partial fulfilment of the requirements
for the degree of
Master of Science

Master of Science (2019)	McMaster University
(Mathematics & Statistics)	Hamilton, Ontario, Canada

TITLE:	Frequentist Model Averaging for ϵ-Support Vector Regression
AUTHOR:	Francis Kiwon Bachelor of Science in Foreign Service, (International Economics) Georgetown University, Al-Rayyan, Qatar
SUPERVISOR:	Dr. Jeffrey S. Racine
NUMBER OF PAGES:	viii, 68

To my one and only mother

Abstract

This thesis studies the problem of frequentist model averaging over a set of multiple ϵ-support vector regression (SVR) models, where the support vector machine (SVM) algorithm was extended to function estimation involving continuous targets, instead of categorical ones. By assigning weights to a set of candidate models instead of selecting the least misspecified one, model averaging presents a strong alternative to model selection for tackling model uncertainty. Not only do we describe the construction of smoothed BIC/AIC model averaging weights, but we also propose a Mallows model averaging procedure which selects model weights by minimizing Mallows' criterion. We conduct two studies where the set of candidate models can either include or not include the true model by making use of simulated random samples obtained from different data-generating processes of analytic form. In terms of mean squared error, we demonstrate that our proposed method outperforms other model averaging and model selection methods that were tested, and the gain is more substantial for smaller sample sizes with larger signal-to-noise ratios.

Acknowledgements

I would like to thank my supervisor Dr. Jeffrey Racine for providing support during my master's studies at McMaster University. As a mentor, Dr. Racine always opened the door of his office when I needed his help with both my thesis and coursework, and he guided me in the right direction when I got lost amidst challenging research questions. I am grateful for his constant motivation, and for encouraging me to investigate the interesting field of model averaging. Furthermore, I really appreciate Dr. Suzanna Becker and Dr. Shui Feng, who gladly accepted my invitations to be the members of my thesis defense committee and provided valuable comments for future research.

Secondly, it was an honor working with Dr. Reza Arabi Belaghi, Dr. Feng, Dr. David Lozinski, Dr. Rosario Monter, Dr. Roman Viveros-Aguilera, and Mr. Christopher McLean as their teaching assistant during my studies.

I also thank my most reassuring friends Dr. Jeongjae Lee, Mr. Matthew Brown, and Mr. Jayden Choi, as well as my spiritual teacher Fr. Abbot John Braganza OSB for endlessly encouraging me not to give up in numerous demanding moments.

Meanwhile, I am not going to forget the following people who not only inspired me to pursue graduate studies, but also reminded me of how valuable I am during my undergraduate years: Dr. Alexis Antoniades, Dr. Jose Asturias, Dr. Brendan Hill,

Dr. Patrick Laude, Dr. Patrick Meadows, Dr. Mahnaz Mousavi, Dr. Max Oidtmann, and Dr. Daniel Westbrook at Georgetown University in Qatar, and Dr. Snezhana Abarzhi at the University of Western Australia.

Lastly, I would like to express my deepest love and appreciation to my mother Yeo Soon Kim for her selfless sacrifice, which has never ceased since I joined her again in 2012 after thirteen years of unwilling separation from each other.

Contents

Chapter 1

Introduction

The issue of model uncertainty leaves practitioners unsure about which single model among a large number of candidate models to adopt for either classification or regression analysis. *Model selection* and *model averaging* are the two dominant and promising approaches entertained by practitioners who want to reduce the risks associated with model misspecification. For model selection, the user chooses only one model as the least misspecified from a set of candidate models based on any of the selection criteria, each of which may favor different models. In other words, the candidate model selected by a criterion is applied a weight of 1, while all others in the set are assigned a weight of 0. Examples of selection criteria include Akaike Information Criterion (AIC; Akaike, 1973, 1974), Mallows' C_p (Mallows, 1973), delete-one cross validation (Stone, 1974), Bayesian Information Criterion (BIC; Schwarz, 1978), and so forth.

Model averaging, an alternative to model selection, produces a weighted average of a set of candidate models with a model averaging criterion by assigning a vector of *nonnegative* weights. Barnard (1963) is one of the first papers to introduce the concept of model averaging, which was demonstrated with an analysis of airline passenger data. Within the Bayesian paradigm, the user may define the weight of each prediction based on the posterior probability for a model as long as the corresponding prior can be drawn. Hoeting *et al.* (1999) provided both methodological and theoretical foundations for the Bayesian model averaging (BMA) approach. On the other hand, since Bates and Granger (1969) proposed the forecast combination, there has been significant progress in the literature for the frequentist model averaging (FMA) where the weights are determined solely by the available data. Hjort and Claeskens (2003) offer readers a comprehensive review. Notable contributions include Buckland *et al.* (1997), Burnham and Anderson (2004), Hansen (2007), Liang *et al.* (2011), Hansen and Racine (2012), and Liu and Kuo (2016) to name but a few.

For statistical learning problems, variable and feature selection can facilitate data visualization and data understanding, reduce training times, and defy the "curse of dimensionality" to improve the prediction performance of statistical models (Guyon and Elisseeff, 2003). We are motivated by the implication that model averaging further reduces the estimation variance, and therefore can be a more promising technique, provided the bias is controlled. While the use of BMA on machine learning techniques has received much well-deserved attention, in this thesis we focus on the application of various FMA approaches to support vector regression (SVR), which is well-suited to real-world applications that make use of regression modelling, such as in the fields of biology, finance, neuroscience, and textual analysis. In particular, we adopt AIC, BIC, and Mallows' C_p criteria for selecting the model weights and allow the predictors in our models to be either categorical or continuous.

The rest of this thesis proceeds as follows. In Chapter 2, we provide the foundations of FMA approaches, which selects the model weights, asymptotic optimality of our proposed methods, and the introduction to SVR with which our model will be specified over mixed datatypes. We construct the methodology for the application of FMA to SVR in Chapter 3 and then examine the finite-sample performance of the proposed approaches relative to model selection estimators for the various data-generating processes (DGP) of the analytic form in Chapter 4. Chapter 5 considers an illustrative example and a comparison of out-of-sample data performance of our model averaging and model selection methods. Chapter 6 presents concluding remarks and brief suggestions for further research which can be carried out on the topic of our thesis. R codes are attached in the Appendix.

Chapter 2

Literature Review

2.1 Model Averaging

Consider a researcher who has gathered data concerning academic achievement of Grade 9 and 10 students in the United States. For each student, he has recorded a variety of demographic predictors such as state of residence, parents' income, gender, race, month of birth, and immigration status, along with the student's highest Preliminary SAT (PSAT) score for each of the sections. He is interested in assessing the size of each covariate's impact on the PSAT score, as well as predicting the students' performance in the actual SAT test when they are in Grade 11 and 12. He uses a simple linear model which fits the data well with reasonable parameter estimates, and decides to estimate the marginal effects of the covariates for the chosen model. However, suppose there exists another well-fitted linear model with substantively different estimates of marginal effects as well as different predictions. As all statistical models are to some extent misspecified, to rely on a single model involves risks; furthermore, selecting a specific model over several other candidate models can lead to a dilution of information about effect sizes and prediction, as observed by Hodges (1987). Model averaging suggests an alternative way around these issues, and it has the potential to provide superior results compared to model selection as detailed in several papers (see e.g., Buckland *et al.* 1997; Hoeting *et al.* 1999; Breiman 2001; Wasserman 2000; Burnham and Anderson 2003; Claeskens and Hjort 2008).

For $\mu = (\mu_1, \ldots, \mu_n)'$, a quantity of interest such as conditional mean, variance, density, or distribution function, define $\hat{\mu}_j = (\hat{\mu}_{1j}, \ldots, \hat{\mu}_{nj})' : j = 1, 2, \ldots, K$ as the estimator of μ obtained from the jth statistical model M_j, and $\mathbf{w} = (w_1, \ldots, w_K)'$ as the vector of weights such that

$$\sum_{k=1}^{K} w_k = 1; 0 \leq w_k \leq 1 \tag{2.1}$$

each of which corresponds to the jth model in the unit simplex given by

$$\mathcal{H}_n = \left\{ \mathbf{w} \in [0,1]^K : \sum_{k=1}^{K} w_k = 1 \right\} \in \mathbb{R}^K. \tag{2.2}$$

Therefore, we obtain a *model averaging estimator* which is

$$\hat{\mu}(\mathbf{w}) = \sum_{j=1}^{K} w_j \hat{\mu}_j \tag{2.3}$$

or a *model selection estimator* as a special case when we restrict the value of w_j to lie in $\{0, 1\}$.

In this section, we first introduce the framework of Bayesian model averaging (BMA), which is the most common approach for weight specification over a set of *parametric* candidate models. Then, we proceed with the principles of frequentist model averaging (FMA) based upon Kullback-Leibler information, which allows candidate models to be *nonparametric* so that practitioners can estimate an unknown data-generating process (DGP) that belongs to a rich class of functions when confronted with model misspecification.

2.1.1 Bayesian Model Averaging

Roberts (1965) first suggested a weighted combination of posterior distributions of two experts or models. Based on this idea, Leamer (1978) presented the basic paradigm for BMA, pointing out that the founding idea for BMA comes from the uncertainty associated with model selection. However, BMA was not used as a standard data analysis tool for decades due to limited theoretical investigations and lack of available computational power (Hoeting *et al.*, 1999). Draper (1995), Chatfield (1995), and Kass and Raftery (1995) all review the adverse effects of model uncertainty, and present BMA as a way of overcoming them.

For the models considered, denoted $M_1, M_2, \ldots, M_K$, and the input vector $\boldsymbol{x} = (x_1, \ldots, x_{p_j}) \in \mathcal{X} \subseteq \mathbb{R}^{p_j}$ where $p_j = \dim(M_j)$, the posterior distribution of μ given $\boldsymbol{x}$ is

$$\mathrm{P}\left(\mu|\boldsymbol{x}\right) = \sum_{j=1}^{K} \mathrm{P}\left(\mu|M_j, \boldsymbol{x}\right) \mathrm{P}\left(M_j|\boldsymbol{x}\right) \tag{2.4}$$

which is an average of the posterior distributions of μ with respect to each model under consideration, weighted by their posterior model probabilities given $\boldsymbol{x}$: $\mathrm{P}\left(M_j|\boldsymbol{x}\right) = w_j$. Define $\mathrm{P}(M_j)$ as the prior probability that the jth model is the true model, given that one of the K models is true. Then

$$w_j = \mathrm{P}\left(M_j|\boldsymbol{x}\right) = \frac{\mathrm{P}\left(\boldsymbol{x}|M_j\right)\mathrm{P}\left(M_j\right)}{\sum_{k=1}^{K}\mathrm{P}\left(\boldsymbol{x}|M_k\right)\mathrm{P}\left(M_k\right)} \tag{2.5}$$

where

$$\mathrm{P}\left(\boldsymbol{x}|M_j\right) = \int \mathrm{P}\left(\boldsymbol{x}|\boldsymbol{\theta}_j, M_j\right)\mathrm{P}\left(\boldsymbol{\theta}_j|M_j\right) d\boldsymbol{\theta}_j \tag{2.6}$$

is the marginal likelihood of the jth model, $\boldsymbol{\theta}_j$ the corresponding vector of model parameters, $\mathrm{P}\left(\boldsymbol{\theta}_j|M_j\right)$ the prior density of $\boldsymbol{\theta}_j$ under the model, and $\mathrm{P}\left(\boldsymbol{x}|\boldsymbol{\theta}_j, M_j\right)$ the likelihood function. Within the Bayesian framework, the model averaging estimator (or posterior mean of μ) is

$$\hat{\mu}(\mathbf{w}) = \mathrm{E}\left[\mu|\boldsymbol{x}\right] = \sum_{j=1}^{K} \hat{\mu}_j \mathrm{P}\left(M_j|\boldsymbol{x}\right) = \sum_{j=1}^{K} w_j \hat{\mu}_j \tag{2.7}$$

and the posterior model variance is

$$\begin{aligned}\mathrm{Var}\left[\mu|\boldsymbol{x}\right] &= \sum_{j=1}^{K} \mathrm{P}\left(M_j|\boldsymbol{x}\right)\left(\mathrm{Var}\left[\mu|\boldsymbol{x}, M_j\right] + \hat{\mu}_j^2\right) - \mathrm{E}\left[\mu|\boldsymbol{x}\right]^2 \\ &= \sum_{j=1}^{K} w_j \left(\mathrm{Var}\left[\mu|\boldsymbol{x}, M_j\right] + \hat{\mu}_j^2\right) - \hat{\mu}(\mathbf{w})^2\end{aligned} \tag{2.8}$$

where $\hat{\mu}_j = \mathrm{E}\left[\mu|\boldsymbol{x}, M_j\right]$ (Raftery, 1993; Draper, 1995; Hoeting *et al.*, 1999). We have

$$\mathrm{BIC}_j = -2\log \mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j, M_j\right) + \log(n)p_j \tag{2.9}$$

established by Schwarz (1978) such that $\mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j, M_j\right)$ is the maximum likelihood estimate of model j, $p_j = \dim\left(M_j\right)$, and therefore the marginal likelihood can be approximated by $\mathrm{P}\left(\boldsymbol{x}|M_j\right) \approx \exp\left(-\mathrm{BIC}_j/2\right)$.

For the choice of the prior $\mathrm{P}\left(M_j\right)$, George and McCulloch (1993) and Volinsky *et al.* (1997) specify a jth prior model probability as

$$\mathrm{P}\left(M_j\right) = \prod_{l=1}^{p} \pi_l^{\delta_{jl}} \left(1 - \pi_l\right)^{1-\delta_{jl}} \tag{2.10}$$

where $\pi_l \in [0, 1]$ is the prior probability that $\theta_l \neq 0$, and $\delta_{jl} = \mathbf{1}\{X_l \text{ is included in } M_j\}$. If π_l is 0.5 for all p predictors, the prior is uniform across the model space; if $\pi_l < 0.5$ for all l there is a penalty imposed on large models; if $\pi_l = 1$ all models contain the variable X_l.

On the other hand, Akaike (1978) maximizes the entropy of the distribution specified by the likelihoods with respect to $\mathrm{P}\left(M_j\right)$, defined by $\mathrm{P}\left(M_j\right) = (1 - \rho)\, \rho^j$, where the value of $\rho \in [0, 1]$ maximizes

$$\sum_{j=1}^{K} \exp\left(-\mathrm{AIC}_j/2\right) \log \mathrm{P}\left(M_j\right)$$

and we use Akaike (1974)'s classical definition of $\mathrm{AIC}_j = -2 \log \mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j, M_j\right) + 2p_j$. The marginal likelihood can be approximated by $\mathrm{P}\left(\boldsymbol{x}|M_j\right) \approx \exp\left(-\mathrm{AIC}_j/2\right)$ in the same way as was done for BIC. Thus Equation (2.5) for the Bayesian model weight can be written as

$$w_j = \frac{\exp\left(-\mathrm{BIC}_j/2\right)\mathrm{P}\left(M_j\right)}{\sum_{k=1}^{K} \exp\left(-\mathrm{BIC}_k/2\right)\mathrm{P}\left(M_k\right)}, \tag{2.11a}$$

$$\text{or } w_j = \frac{\exp\left(-\mathrm{AIC}_j/2\right)\mathrm{P}\left(M_j\right)}{\sum_{k=1}^{K} \exp\left(-\mathrm{AIC}_k/2\right)\mathrm{P}\left(M_k\right)} \tag{2.11b}$$

depending on the user's choice of information criterion.

2.1.2 Kullback-Leibler Divergence and AIC

As reviewed in Section 2.1.1, BMA methodology considers model uncertainty by setting prior probabilities for a collection of candidate models and the parameters of each. In addition, it properly accounts for the increasing estimator variability resulting from not knowing the true model *a priori*.

On the other hand, BMA typically involves conflicts between many prior opinions about the parameters of interest. In the frequentist view, we rather seek the "best approximating model" since we can never perfectly identify the "true" model which reflects full reality (Burnham and Anderson, 2003).

Kullback and Leibler (1951) provides the definition of "information" in terms of a distance between reality and its approximation. Consider M_0 with a density function $f(\boldsymbol{x})$, where the true model has unknown parameter values and the dimension of parameter space is undefined. We estimate M_0 with an approximating model j with another density function $g(\boldsymbol{x}|\boldsymbol{\theta})$. Then, the Kullback-Leibler (K-L) divergence $I(M_0, M_j)$, the information lost through approximation of M_0 using $g(\boldsymbol{x}|\boldsymbol{\theta})$, is defined as the integral

$$\begin{aligned} I(M_0, M_j) &= \int f(\boldsymbol{x}) \log\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x}|\boldsymbol{\theta})}\right) dx \\ &= \int f(\boldsymbol{x}) \log\left(f(\boldsymbol{x})\right) dx - \int f(\boldsymbol{x}) \log\left(g(\boldsymbol{x}|\boldsymbol{\theta})\right) dx \qquad (2.12) \\ &= \mathrm{E}_f\left[\log\left(f(\boldsymbol{x})\right)\right] - \mathrm{E}_f\left[\log\left(g(\boldsymbol{x}|\boldsymbol{\theta})\right)\right]. \end{aligned}$$

Note that the truth is not dependent on sample size n - the form of $f(x)$ is not assumed *a priori*, and the candidate models may not be nested. $I(M_0, M_j)$ approaches zero as the jth candidate model loses less information relative to the other candidate models. Because the truth and therefore the quantity $\mathrm{E}_f\left[\log\left(f(\boldsymbol{x})\right)\right]$ are constant and independent of the candidate models, it is enough to estimate relative expected K-L information $\mathrm{E}_f\left[\log\left(g(\boldsymbol{x}|\boldsymbol{\theta})\right)\right]$, which is unknown, to find out which approximation is the best. The maximum log-likelihood $\log \mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j, M_j\right)$ is a biased estimate of

$$\mathrm{E}_y\left[\mathrm{E}_x\left[\log\left(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}(\boldsymbol{y}))\right)\right]\right]$$

where the inner part of the double expectation is just $\mathrm{E}_f\left[\log\left(g(\boldsymbol{x}|\boldsymbol{\theta})\right)\right]$ with $\boldsymbol{\theta}$ replaced by its maximum likelihood estimator (MLE) based on model j and target data $\boldsymbol{y}$ (Akaike, 1973, 1974; Burnham and Anderson, 2004). Subtracting p_j as the asymptotic bias correction term and multiplying the result by -2, we obtain

$$\mathrm{AIC}_j = -2\log \mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j, M_j\right) + 2p_j. \tag{2.13}$$

For a least-squares (LS) estimator with homoskedastic errors $e \sim \mathcal{N}(0, \sigma^2)$ i.i.d., AIC can also be expressed as

$$\mathrm{AIC}_j = n\log\left(\frac{\sum_{i=1}^{n}\hat{e}_i^2}{n}\right) + 2p_j \tag{2.14}$$

where $\sum_{i=1}^{n}\hat{e}_i^2$ is a residual sum of squares (RSS) from the fitted model.

When n is small, AIC tends to select a model with too many parameters (see Claeskens and Hjort, 2008, Ch. 8.3). To deal with overfitting, Hurvich and Tsai (1989) suggested the corrected AIC criterion ($\mathrm{AIC_c}$) that provides a small-sample bias correction:

$$\mathrm{AIC}_{\mathrm{c},j} = \mathrm{AIC} + \frac{2p_j(p_j+1)}{n-p_j-1} \tag{2.15}$$

and $\mathrm{AIC}_{\mathrm{c},j}$ converges to AIC_j as $n \to \infty$. Interested readers may see Cavanaugh (1997) for a unified mathematical justification of AIC and AIC_c.

2.1.3 Smoothed BIC/AIC Model Averaging

Contrasted with BMA, the FMA weights are solely determined by data, and therefore priors are assumed non-informative, or uniform within the frequentist paradigm. Based on Equation (2.11a), Buckland *et al.* (1997) assign smooth BIC (sBIC) model averaging weights w_j given by:

$$w_j = \frac{\exp\left(-\mathrm{BIC}_j/2\right)}{\sum_{k=1}^{K} \exp\left(-\mathrm{BIC}_k/2\right)} \tag{2.16}$$

with $\mathrm{P}\left(M_j\right) = 1/K$.

From Equation (2.9) we have

$$\begin{aligned} \mathrm{BIC}_j &= -2\log \mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j, M_j\right) + \log(n)p_j \\ \Longleftrightarrow \exp\left(-\frac{\mathrm{BIC}_j}{2}\right) &= \mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_j, M_j\right)\exp\left(-\frac{\log(n)p_j}{2}\right). \end{aligned} \tag{2.17}$$

For two models, a and b, both of which contain p predictors, then

$$\begin{aligned} \frac{w_a}{w_b} &= \frac{\exp\left(-\mathrm{BIC}_a/2\right)}{\exp\left(-\mathrm{BIC}_b/2\right)} = \frac{\mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_a, M_a\right)\exp\left(-\log(n)p_a/2\right)}{\mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_b, M_b\right)\exp\left(-\log(n)p_b/2\right)} \\ &= \frac{\mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_a, M_a\right)}{\mathrm{P}\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_b, M_b\right)} \end{aligned} \tag{2.18}$$

is simply the likelihood ratio, or the approximation of the Bayes factor for comparing the two models (Akaike, 1981; Draper, 1995). Moreover, if the odds ratio of priors is one, then the likelihood ratio is equal to the odds ratio of posteriors of the candidate models. Similarly, "Akaike weights" for smoothed AIC (sAIC) model averaging are given by:

$$w_j = \frac{\exp\left(-\mathrm{AIC}_j/2\right)}{\sum_{k=1}^{K} \exp\left(-\mathrm{AIC}_k/2\right)}. \tag{2.19}$$

where we substitute BIC with AIC. Burnham and Anderson (2004) state that Akaike weight works, as "the weight of evidence' in favor of [model j] as being the actual K-L best model" conditional on both the data and the full collection of candidate models (see also Burnham and Anderson, 2003, Ch. 6 for further details).

2.1.4 Mallows Model Averaging

Hansen (2007) proposed an LS model averaging estimator which selects model weights by minimizing a Mallows model averaging (MMA) criterion. With a random sample $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathbb{R} : i = 1, \ldots, n; \dim(\boldsymbol{x}_i) = p$, we assume a linear model

$$\begin{aligned} & y_i = \mu_i + e_i \\ & \text{s.t. } \mu_i = \sum_{l=1}^{p} \theta_l x_{il}; \\ & \qquad \mathrm{E}\left[e_i|\boldsymbol{x}_i\right] = 0; \mathrm{E}\left[e_i^2|\boldsymbol{x}_i\right] = \sigma^2; \\ & \qquad \mathrm{E}[\mu_i^2] < \infty \end{aligned} \tag{2.20}$$

as well as a sequence of K approximating models, where the jth model contains the first p_j predictors in the ordered set such that $0 < k_1 < k_2 < \cdots < k_p$. Then candidate model j is

$$y_i = \sum_{l=1}^{p_j} \theta_l x_{il} + b_{(j)i} + e_i \tag{2.21}$$

whose approximation error is $b_{(j)i} = \sum_{l=p_j+1 \leq p}^{p} \theta_l x_{il}$.

We can express Equation (2.21) in matrix notation as $\mathbf{y} = \mathbf{X}_j \boldsymbol{\theta}_j + \mathbf{b}_j + \boldsymbol{e}$, where $\mathbf{y} = (y_1, \ldots, y_n)', \mathbf{X}_j$ is the $n \times p_j$ matrix with ilth element $x_{il}, \boldsymbol{\theta}_j = (\theta_1, \ldots, \theta_{p_j})', \mathbf{b}_j = (b_{(j)1}, \ldots, b_{(j)n})'$, and $\boldsymbol{e} = (e_1, \ldots, e_n)'$. Define the LS estimate of $\boldsymbol{\theta}_j$ in all models j as $\hat{\boldsymbol{\theta}}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{y}$, then $\mu = (\mu_1 \ldots, \mu_n)' = \mu_j + \mathbf{b}_j$ where $\mu_j = \mathbf{X}_j \boldsymbol{\theta}_j + \mathbf{b}_j$. The corresponding estimator of μ from the jth model is $\hat{\mu}_j = \mathbf{X}_j \hat{\boldsymbol{\theta}}_j = \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{y} \equiv \mathbf{H}_j\mathbf{y}$. Under Assumption (2.2), the model averaging estimators of μ and $\boldsymbol{\theta}_K$ are

$$\hat{\boldsymbol{\theta}}(\mathbf{w}) = \sum_{k=1}^{K} w_k \begin{pmatrix} \hat{\theta}_k \\ 0 \end{pmatrix} \tag{2.22a}$$

$$\begin{aligned} \hat{\mu}(\mathbf{w}) &= \sum_{k=1}^{K} w_k \mathbf{H}_k \mathbf{y} \equiv \mathbf{H}(\mathbf{w})\mathbf{y} \\ &= \sum_{k=1}^{K} w_k \mathbf{X}_k (\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{y} \equiv \sum_{k=1}^{K} w_k \hat{\mu}_k \\ &= \sum_{k=1}^{K} w_k \mathbf{X}_k \hat{\theta}_k = \mathbf{X} \sum_{k=1}^{K} w_k \begin{pmatrix} \hat{\theta}_k \\ 0 \end{pmatrix} \\ &= \mathbf{X}\hat{\boldsymbol{\theta}}(\mathbf{w}) \end{aligned} \tag{2.22b}$$

where $\mathbf{H}(\mathbf{w}) = \sum_{k=1}^{K} w_k \mathbf{H}_k$ is the “implied ‘hat’ matrix” so that $\text{tr}(\mathbf{H}_j) = p_j$ if $\mathbf{X}_j$ has full column rank.

The MMA criterion is

$$C_n(\mathbf{w}) = (\mathbf{y} - \hat{\mu}(\mathbf{w}))' (\mathbf{y} - \hat{\mu}(\mathbf{w})) + 2\sigma^2 p(\mathbf{w}) \tag{2.23}$$

where $p(\mathbf{w}) \equiv \mathrm{tr}(\mathbf{H}(\mathbf{w})) = \sum_{k=1}^{K} w_k p_k$ is the number of nontrivial parameters. We can write the formula again as

$$C_n(\mathbf{w}) = \mathbf{w}'\hat{\mathbf{E}}'\hat{\mathbf{E}}\mathbf{w} + 2\sigma^2 \boldsymbol{p}'\mathbf{w} \tag{2.24}$$

where $\hat{\mathbf{E}} = (\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_K) = ((\mathbf{y} - \hat{\mu}_1), (\mathbf{y} - \hat{\mu}_2), \ldots, (\mathbf{y} - \hat{\mu}_K)) = ((\mathbf{y} - \mathbf{H}_1\mathbf{y}), (\mathbf{y} - \mathbf{H}_2\mathbf{y}), \ldots, (\mathbf{y} - \mathbf{H}_K\mathbf{y}))$ is the $n \times K$ matrix whose column j contains the residual vector $\boldsymbol{e}_j = (e_{1j}, \ldots, e_{n_j})'$ from the jth candidate model, and $\boldsymbol{p} = (p_1, \ldots, p_K)'$. We use this criterion to select the weight vector

$$\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w} \in \mathcal{H}_n} C_n(\mathbf{w}). \tag{2.25}$$

The empirical weight vector $\hat{\mathbf{w}}$ can be obtained numerically since no closed-form solution exists for Equation (2.25). The solution minimizes $C_n(\mathbf{w})$ subject to Assumption (2.2), and can be obtained by solving a simple quadratic program.

Hansen (2007) proves that the MMA criterion $C_n(\mathbf{w})$ presents an unbiased estimate of the mean squared error (MSE) from the model averaging fit, and is asymptotically optimal in the sense of achieving the lowest MSE in a class of model averaging estimators. The R package `ma` is readily available for practitioners' use (Racine, 2017). While Hansen (2007) initially assumed that a candidate model is always nested within the larger models in sequence, the asymptotic optimality of the MMA estimator still holds even if the candidate models are non-nested, and the model weights lie within a continuous set. See Wan *et al.* (2010) for further details.

2.1.5 Jackknife Model Averaging

Hansen and Racine (2012) proposes an extended version of MMA approach called Jackknife model averaging (JMA), which allows the regression error of a candidate model to be heteroskedastic; i.e. we change the homoskedasticity assumption in Equation (2.20) to $\mathrm{E}\left[e_i^2|\boldsymbol{x}_i\right] = \sigma_i^2$ so that the conditional variance can be dependent on $\boldsymbol{x}_i$.

Define $\tilde{\mu}_j = (\hat{\mu}_{(-1),j}, \hat{\mu}_{(-2),j}, \ldots, \hat{\mu}_{(-n),j})'$ where $\hat{\mu}_{(-i),j} =: i = 1, \ldots, n$ is the jackknife estimator of μ obtained from the jth model with the ith observation deleted. We can write $\hat{\mu}_{(-i),j} = x_{ij}(\mathbf{X}'_{(-i),j}\mathbf{X}_{(-i),j})^{-1}\mathbf{X}'_{(-i),j}\mathbf{y}_{(-i)}$ where $\mathbf{X}_{(-i),j}$ and $\mathbf{y}_{(-i)}$ are the matrices $\mathbf{X}_j$ and $\mathbf{y}$ with the ith row removed; furthermore, we write $\tilde{\mu}_j = \tilde{\mathbf{H}}_j\mathbf{y}$ whose jackknife hat matrix $\tilde{\mathbf{H}}_j$ has 0's on its diagonal. The leave-one-out residual vector for $\tilde{\mu}_j$ is then $\tilde{\boldsymbol{e}}_j = \mathbf{y} - \tilde{\mu}_j$.

Now let $\mathbf{D}_j = \mathrm{diag}(1 - h_{11}^{(j)}, 1 - h_{22}^{(j)}, \ldots, 1 - h_{nn}^{(j)})$ where $h_{ii}^{(j)} = x_{ij}(\mathbf{X}_j'\mathbf{X}_j)^{-1}x_{ij}$ is the ith diagonal element of the hat matrix $\mathbf{H}_j$ for model j. From Li (1987) we show that

$$\begin{aligned}
&\tilde{\mathbf{H}}_j = \mathbf{D}_j(\mathbf{H}_j - \mathbf{I}) + \mathbf{I} \\
&\implies (\mathbf{I} - \tilde{\mathbf{H}}_j)\mathbf{y} = \mathbf{D}_j(\mathbf{I} - \mathbf{H}_J)\mathbf{y} \\
&\text{(LHS) } \mathbf{y} - \tilde{\mathbf{H}}_j\mathbf{y} = \mathbf{y} - \tilde{\mu}_j = \tilde{\boldsymbol{e}}_j \\
&\text{(RHS) } \mathbf{D}_j(\mathbf{y} - \mathbf{H}_j\mathbf{y}) = \mathbf{D}_j(\mathbf{y} - \hat{\mu}_j) = \mathbf{D}_j\hat{\boldsymbol{e}}_j \\
&\therefore \tilde{\boldsymbol{e}}_j = \mathbf{D}_j\hat{\boldsymbol{e}}_j
\end{aligned} \tag{2.26}$$

Consequently, we can immediately compute $\tilde{\boldsymbol{e}}_j$ with a simple linear operation. See also Racine (1997) for the generalization of the relationship in the first line of Equation (2.26).

The resulting JMA estimator $\tilde{\mu}(\mathbf{w})$ is

$$\begin{aligned}\tilde{\mu}(\mathbf{w}) &= \sum_{k=1}^{K} w_k \tilde{\mathbf{H}}_k \mathbf{y} \equiv \tilde{\mathbf{H}}(\mathbf{w})\mathbf{y} \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n} w_k x_{ik} (\mathbf{X}'_{(-i),k}\mathbf{X}_{(-i),k})^{-1}\mathbf{X}'_{(-i),k}\mathbf{y}_{(-i)} \\ &= \sum_{k=1}^{K} w_k \tilde{\mu}_k\end{aligned} \tag{2.27}$$

and the jackknife estimate of residual sum of squares is

$$CV_n(\mathbf{w}) = (\mathbf{y} - \tilde{\mu}(\mathbf{w}))(\mathbf{y} - \tilde{\mu}(\mathbf{w}))' = w'\tilde{\mathbf{E}}'\tilde{\mathbf{E}}w \tag{2.28}$$

where $\tilde{\mathbf{E}} = (\tilde{\boldsymbol{e}}_1, \ldots, \tilde{\boldsymbol{e}}_K) = ((\mathbf{y} - \tilde{\mu}_1), (\mathbf{y} - \tilde{\mu}_2), \ldots, (\mathbf{y} - \tilde{\mu}_K))$ is the $n \times K$ leave-one-out residual matrix, and $\tilde{\mathbf{E}}\mathbf{w} = (\mathbf{y} - \tilde{\mu}(\mathbf{w})) = \sum_{k=1}^{K} w_k \tilde{\boldsymbol{e}}_k$ is the JMA residual. Finally, the JMA or leave-one-out cross-validation choice of weight vector minimizes $CV_n(\mathbf{w})$ over $\mathbf{w} \in \mathcal{H}_n$:

$$\tilde{\mathbf{w}} = \text{argmin}_{\mathbf{w} \in \mathcal{H}_n} CV_n(\mathbf{w}). \tag{2.29}$$

Note that JMA is nearly equivalent to MMA in the presence of homoskedastic errors. The JMA estimator is asymptotically optimal in the sense of achieving the lowest possible MSE over the collection of both nested and nonnested linear models, which include but are not limited to least-squares, ridge regression, local polynomial kernel regression with fixed bandwidths, k-nearest neighbor estimators, estimators of additive interaction models, and spline estimators (Hansen and Racine, 2012). Zhang *et al.* (2013) show that this asymptotic optimality holds with serial correlation in the errors, and the method remains valid under model settings involving time-dependent data.

2.2 Support Vector Regression

In the exercise of *supervised learning* or *pattern recognition*, we find a function which predicts the values of one or more outputs, using a collection of measured or preset inputs (Hastie *et al.*, 2009). The *support vector* algorithm represents a nonlinear generalization of the *Generalized Portrait* algorithm Vapnik and Lerner (1963). It is firmly grounded in Vapnik-Chervonenkis (VC) theory developed by Vapnik (1999), a sub-branch of statistical learning theory, which provided solid theoretical foundations for controlling the generalization ability of a learning model given independent (out-of-sample) data.

Boser *et al.* (1992) and Cortes and Vapnik (1995) developed support vector machines (SVM) in the present form. SVMs have been successfully applied to classification problems in the fields of OCR (optical character recognition; Schölkopf *et al.*, 1995; Bahlmann *et al.*, 2002; Niu and Suen, 2012), financial forecasting (Van Gestel *et al.*, 2001; Kim, 2003; Huang *et al.*, 2005; Shin *et al.*, 2005), cancer prediction (Furey *et al.*, 2000; Guyon *et al.*, 2002; Huang *et al.*, 2017), EEG (electroencephalogram) signal processing (Garrett *et al.*, 2003; Thulasidas *et al.*, 2006; Subasi and Gursoy, 2010), textual analysis (Drucker *et al.*, 1999; Tong and Koller, 2002; Agarwal and Sureka, 2015), and so on. Burges (1998) published a comprehensive tutorial on SVM classifiers.

Vapnik *et al.* (1996), Drucker *et al.* (1996), and Müller *et al.* (1997) also constructed an extension of the SVM algorithm for function estimation and time series prediction involving continuous target variables. Example applications of the support vector regression (SVR) method include time series forecasting (Wu *et al.*, 2004; Chen and Wang, 2007; Lu *et al.*, 2009). It has also been widely used in the field of bioinformatics (Myasnikova *et al.*, 2002; Long *et al.*, 2011; Sun *et al.*, 2011).

Given a random sample as defined in Section 2.1.4, we aim to find a model with the corresponding function $g(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{\theta})$ which maximizes the deviation of the target variables y_i from ϵ; any errors larger than ϵ, however, will not be accepted. Consider a simple linear model that takes the form

$$g(\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{\theta}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle + b \text{ with } \boldsymbol{\theta} \in \mathcal{X}, b \in \mathbb{R} \tag{2.30}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product in $\mathcal{X}$, and b is a bias term or "threshold." We now describe the SVR method as a convex optimization problem in accordance with the tutorial by Smola and Schölkopf (2004).

2.2.1 Convex Optimization

As we seek the *flattest* model minimizing the number of $g(\boldsymbol{x_i})$ which deviate from y_i by greater than ϵ, we prefer a small $\boldsymbol{\theta}$, thereby minimizing the L^2 norm $||\boldsymbol{\theta}||^2$ which characterizes model complexity. Assuming there exists a function g which estimates all sample points within the "hard margin" ϵ, we write the following convex optimization problem given by

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\ &\text{s.t. } \begin{cases} y_i - g(\boldsymbol{x}_i) \leq \epsilon \\ g(\boldsymbol{x}_i) - y_i \leq \epsilon \end{cases} \end{aligned} \tag{2.31}$$

However, this case of maintaining a hard margin is not always feasible, so we add a pair of "slack variables" ξ_i, ξ_i^* (Smith, 1968; Bennett and Mangasarian, 1992) to the constraints in Equation (2.31) instead to adopt a "soft margin" loss function as follows (Cortes and Vapnik, 1995; Shawe-Taylor and Cristianini, 1998; Vapnik, 1999):

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &\text{s.t. } \begin{cases} y_i - g(\boldsymbol{x}_i) \leq \epsilon + \xi_i \\ g(\boldsymbol{x}_i) - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{2.32}$$

The cost hyperparameter C, which behaves like a traditional regularization parameter, trades off model complexity, or the flatness of g against how many errors larger than ϵ are tolerated in the objective function (2.32). Then, we can describe the so-called ϵ-insensitive loss function $L_\epsilon(y, g(\boldsymbol{x}))$ as

$$L_\epsilon(y, g(\boldsymbol{x})) = \max\left\{(|y - g(\boldsymbol{x})| - \epsilon), 0\right\}. \tag{2.33}$$

2.2.2 Dual Problems and Quadratic Programs

From the objective function (2.32), we construct a Lagrange function as well as the corresponding constraint where we introduce a dual set of Lagrange multipliers. Define the Lagrangian L as

$$\begin{aligned} L = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\eta_i\xi_i - \eta_i^*\xi_i^*) \\ - \sum_{i=1}^{n}\alpha_i(\epsilon + \xi_i - y_i + g(\boldsymbol{x}_i)) \\ - \sum_{i=1}^{n}\alpha_i^*(\epsilon + \xi_i^* + y_i - g(\boldsymbol{x}_i)) \\ \text{s.t. } \eta_i, \eta_i^*, \alpha_i, \alpha_i^* \geq 0 \end{aligned} \tag{2.34}$$

where $\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$ are Lagrange multipliers, with the partial derivatives of L with respect to the primal variables $(\boldsymbol{\theta}, b, \xi_i)$ equal to 0 by the saddle point conditions

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \boldsymbol{\theta} - \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\,x_i = 0 \tag{2.35a}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0 \tag{2.35b}$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i = 0; \frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0 \tag{2.35c}$$

Substituting (2.35) into (2.34) yields the following dual optimization problem given by

$$\begin{aligned} &\text{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ -\epsilon \sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) y_i \end{cases} \\ &\text{s.t.} \ \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \text{ and } \alpha_i, \alpha_i^* \in [0, C] \end{aligned} \tag{2.36}$$

We have eliminated η_i, η_i^* in Equation (2.35c) by having $\eta_i = C - \alpha_i, \eta_i^* = C - \alpha_i^*$. From Equation (2.35a) we have

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \sum_{i=1}^{n} \left(\alpha_i^* - \alpha_i\right) x_i \\ \implies g(\boldsymbol{x}) &= g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \left(\alpha_i^* - \alpha_i\right) \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b. \end{aligned} \tag{2.37}$$

The parameter estimate $\hat{\boldsymbol{\theta}}$ is a linear combination of the predictors x_i. In addition, the model complexity is independent of the number of predictors p, and is affected by the number of data points chosen as support vectors. We take advantage of these characteristics for building a nonlinear extension for our experiment in the next chapter.

2.2.3 Calculation of the bias term

We compute b by exploiting the Karush-Kuhn-Tucker (KKT) conditions, which state that the product between dual variables α_i, α_i^* and the constraints in Equation (2.32) becomes zero:

$$\alpha_i(\epsilon + \xi_i - y_i + g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})) = 0; \alpha_i^*(\epsilon + \xi_i^* + y_i - g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})) = 0 \tag{2.38a}$$

$$(C - \alpha_i)\xi_i = 0; (C - \alpha_i^*)\xi_i^* = 0; \tag{2.38b}$$

Awad and Khanna (2015) and Basak *et al.* (2007) illustrate two useful conclusions following the KKT conditions. First, only random samples whose corresponding α_i or α_i^* is nonzero are located outside the ϵ-insensitive region. Second, we have $\alpha_i\alpha_i^* = 0$ i.e., at least one of α_i and α_i^* should be zero, because it is not possible to have the data point $(\boldsymbol{x}_i, y_i)$ lie on both the lower and upper boundary. Therefore, the corresponding constraint in Equation (2.32) will be satisfied with equality, and $\xi_i = 0$ since the data point is within the ϵ-insensitive region. When $\alpha_i \in (0, C)$, we have

$$\begin{aligned} & y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\theta}} \rangle - b - \epsilon - \xi_i = 0 \\ \implies & y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\theta}} \rangle - b - \epsilon = 0 \\ \implies & \hat{b} = y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\theta}} \rangle - \epsilon \end{aligned} \tag{2.39}$$

Analogously, ξ_i^* vanishes for $\alpha_i^* \in (0, C)$, and we also have

$$\begin{aligned} & \langle \boldsymbol{x}_i, \hat{\boldsymbol{\theta}} \rangle + b - y_i - \epsilon = 0 \\ \implies & \hat{b} = y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\theta}} \rangle + \epsilon \end{aligned} \tag{2.40}$$

Alternatively, we may compute b by solving an interior point optimization problem, whose solution can converge in $O(\log n)$ operations, by searching along the central path of the feasible region. See Smola and Schölkopf (2004) and Keerthi *et al.* (2001) for further details.

Chapter 3

Methods

This chapter establishes the nonlinear SVR model, which uses the radial basis function (RBF) kernel, and the estimation of the target vector using the model selection and model averaging methods within the frequentist framework described in Section 2.1. We introduce the detailed specification of the SVR model in Section 3.1. Then we begin by imposing an equal model averaging weight on all candidate models in Section 3.2. Section 3.3 constructs smoothed AIC and BIC model averaging weights based on Vapnik's ϵ-insensitive loss function which determines the empirical risk for SVR models. Section 3.4 describes the weight choice criterion for the MMA estimator given the unknown errors of candidate models, and describes the proof of asymptotic optimality of the MMA estimator.

3.1 Model Specification

3.1.1 Nonlinear Mapping with the RBF Kernel

We make the SVR algorithm introduced in Section 2.2 nonlinear by preprocessing the input vector $\boldsymbol{x}$ onto a higher-dimensional feature space $\mathcal{F}$ via some fixed mapping $\phi : \mathcal{X} \longrightarrow \mathcal{F}$, and construct a linear model in this space given by

$$g(\boldsymbol{x}|\boldsymbol{\theta}) = \langle \phi(\boldsymbol{x}), \boldsymbol{\theta} \rangle + b = \sum_{l=1}^{p} \theta_l \phi(\boldsymbol{x}_l) + b. \tag{3.1}$$

In other words, linear regression in a higher-dimensional feature space corresponds to nonlinear regression in the low dimensional input space $\mathcal{X} \subseteq \mathbb{R}^p$. We restate the dual problems in Section 2.2.2 as

$$\begin{aligned} &\text{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^* k(\boldsymbol{x}_i, \boldsymbol{x}_j)) \\ -\epsilon \sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i \end{cases} \\ &\text{s.t.} \ \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) \text{ and } \alpha_i, \alpha_i^* \in [0, C] \end{aligned} \tag{3.2}$$

and we compute $\hat{\boldsymbol{\theta}}$ and $g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ as

$$\hat{\boldsymbol{\theta}} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi(\boldsymbol{x}_i) \text{ and } g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{3.3}$$

where $k(\boldsymbol{x}_i, \boldsymbol{x}) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \rangle$. Interested readers are referred to Aizerman *et al.* (1964) for the geometrical interpretation of the kernels as inner products in a feature space.

Without loss of generality, suppose that the kernel function is bounded in the input domain $\mathcal{X}$. We choose the radial basis function (RBF) kernel which satisfies the assumption:

$$k(\boldsymbol{x}_i, \boldsymbol{x}) = \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{x}_i\|^2\right) \tag{3.4}$$

where γ is a hyperparameter set by the user.

Unlike the formulation in Section 2.2.2, where $\hat{\boldsymbol{\theta}}$ is a linear combination of the input vector $\boldsymbol{x}$, the values of coefficients are no longer explicitly given as a result of a nonlinear transformation. The optimization problem corresponds to finding the *flattest* function in the feature space $\mathcal{F}$ instead of the input space $\mathcal{X}$.

3.1.2 Conditions for the RBF Kernel

Not only is the RBF kernel depicted in Equation (3.4) translation invariant, i.e., $k(\boldsymbol{x}_i, \boldsymbol{x}) = \kappa_\gamma(\boldsymbol{x}_i - \boldsymbol{x})$ where $\kappa_\gamma(\cdot) = \exp(-\gamma\|\cdot\|^2)$, but also it corresponds to a dot product in some feature space $\mathcal{F}$. We begin with the following theorem, which characterizes the RBF kernel as *admissible*:

Theorem 3.1 (Mercer 1909) *Suppose $k \in L_\infty$ such that the integral operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$,*

$$T_k f(\cdot) := \int_{\mathcal{X}} k(\cdot, \boldsymbol{x}) g(\boldsymbol{x}) d\nu(\boldsymbol{x}) \tag{3.5}$$

is positive, where ν denotes a measure on $\mathcal{X}$ with $\nu(\mathcal{X}) < \infty$ and $\mathcal{X}$ is the support of ν. Let $\psi_j \in L_2(\mathcal{X})$ be the eigenfunction of T_k associated with the nonzero eigenvalue λ_j and normalized such that the L_2 norm of ψ_j is one. If $\bar{\psi_j}$ is the complex conjugate of ψ_j, then $k(\boldsymbol{x}_i, \boldsymbol{x}) = \sum_{j\in\mathbb{N}} \lambda_j \psi_j \bar{(\boldsymbol{x}_i)} \psi_j(\boldsymbol{x})$ holds for almost all $(\boldsymbol{x}_i, \boldsymbol{x})$ where the series is absolutely and uniformly convergent for almost all $(\boldsymbol{x}_i, \boldsymbol{x})$.

Smola and Schölkopf (2004) explain that, according to this theorem, we can write $k(\boldsymbol{x}_i, \boldsymbol{x})$ as a dot product in $\mathcal{F}$ if the following condition holds:

$$\int_{\mathcal{X}\times\mathcal{X}} k(\boldsymbol{x}_i, \boldsymbol{x})g(\boldsymbol{x}_i)g(\boldsymbol{x})d\boldsymbol{x}_i d\boldsymbol{x} \geq 0 \text{ for all } g \in L_2(\mathcal{X}) \tag{3.6}$$

Secondly, Smola *et al.* (1998b) state a necessary and sufficient condition for a translation invariant kernel:

Theorem 3.2 (Smola *et al.*, 1998b) *A translation invariant kernel* $k(\boldsymbol{x}_i, \boldsymbol{x})$ *is admissible if and only if the Fourier transform*

$$F[k](\omega) = (\sqrt{2\pi})^{-p} \int_{\mathcal{X}} \exp(-i\langle\omega, \boldsymbol{x}\rangle)k(\boldsymbol{x})d\boldsymbol{x} \geq 0 \tag{3.7}$$

Aizerman *et al.* (1964) and Boser *et al.* (1992) also confirmed that the RBF kernel is proper, and a proof based on interpolation theory (Micchelli, 1986) as well as the theory of regularization networks (Girosi *et al.*, 1993) is given in Smola and Schölkopf (2004, see Section 7).

3.1.3 Hyperparameter Selection

The estimation accuracy of generalization performance of our SVR model with the RBF kernel is dependent on an effective setting of hyperparameters ϵ, C, and γ by the user. While the implementations of SVR using available packages usually leave the value of hyperparameters to the user's discretion, the dependence of model complexity on all three of these hyperparameters further complicates the issue of selecting optimal values.

If C is large, we accept a smaller margin, and only minimize the empirical risk function in Equation (3.11) when ignoring model complexity (low bias, high variance). On the other hand, a smaller value of C allows a larger margin and therefore a lower level of complexity, though estimation accuracy is sacrificed (high bias, low variance).

On the other hand, the hyperparameter γ determines the radius of influence of support vectors chosen in the model. If γ is too small, we have a very constrained model in which the radius of influence includes all of the input data points, so that the complexity of the data is not captured (high bias, low variance); if γ is large, the radius of influence includes only support vectors themselves, and the model overfits (low bias, high variance).

Smola *et al.* (1998a) and Kwok (2001) suggest asymptotically optimal values of ϵ which are proportional to the noise variance. The proposal does not reflect that ϵ should be smaller for larger sample sizes provided the data has the same level of noise. Mattera and Haykin (1999) selects C equal to the range of output values, but the choice does not consider the possible effects of outliers within the input vector. Cherkassky and Ma (2004) propose analytical selection of $C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|)$ directly from the input values, where $\bar{y}$ and σ_y are the mean and the standard deviation of the target vector, as well as of $\epsilon = 3\sigma\sqrt{\log n/n}$ based on both noise variance and sample sizes. In this thesis we exploit the cross-validation method suggested by Schölkopf *et al.* (1999) and Momma and Bennett (2002) for hyperparameter selection, in spite of its computational intensity compared to other suggestions. Practitioners can use the function `tune` in the R package `e1071` for selecting the best γ, C, and ϵ using cross-validation over a range of values.

3.2 Nonsmooth Model Weights

We first consider imposing the same nonsmooth model averaging weight $w = 1/K$ on all candidate models in the collection. First consider a class of the non-nested candidate models $\mathbf{y} = \mathbf{X}_s\boldsymbol{\theta}_s + \boldsymbol{e}$ where all $\mathbf{X}_s : s \in \{1, 2, \ldots, K\}$ are $n \times p_s$ disjoint subsets of $\mathbf{X}$, so we make $\mathbf{X}_s$ a sequence of independent random variables. Each $\boldsymbol{\theta}_s$ is a corresponding p_s-dimensional vector of unknown parameters. $\hat{\mu}_s$, an estimate of the quantity of interest obtained from model s, is independent because $\mathbf{X}_s$ are disjoint. If $\hat{\mu}_s$ is unbiased, the expected value of our model averaging estimator $\hat{\mu}(\mathbf{w}) = \sum_{k=1}^{K} \hat{\mu}_k/K$ is $\mathrm{E}[\hat{\mu}(\mathbf{w})] = \mu$. Define $v_K = \sum_{s=1}^{K} \mathrm{Var}[\hat{\mu}_s]$, and if for some $\delta > 0$

$$\lim_{n\to\infty} \frac{1}{v_n^{2+\delta}} \sum_{s=1}^{K} \mathrm{E}\left[|\hat{\mu}_s - \mu|^{2+\delta}\right] = 0 \tag{3.8}$$

is satisfied, by Lyapunov's central limit theorem we have

$$\hat{\mu}(\mathbf{w}) - \mu \to_d \mathcal{N}\left(0, v_K/K^2 = \sum_{s=1}^{K} \frac{\mathrm{Var}[\hat{\mu}_s]}{K^2}\right) \tag{3.9}$$

Although the choice of a uniform weight may not be optimal, we are motivated by the fact that model averaging reduces variance of the resulting estimator, thereby contributing to a smaller loss, which may be of interest to the reader.

3.3 AIC and BIC

Let $\hat{\mu}_j = g(\boldsymbol{x}|\boldsymbol{\theta_j})$ be the functional form of model j, and let $\tilde{\mu} = g(\boldsymbol{x}|\tilde{\boldsymbol{\theta}})$ represent the best candidate model. We can define the loss function, which is the squared error,

$$L(y, g(\boldsymbol{x}|\boldsymbol{\theta_j})) = (y - g(\boldsymbol{x}|\boldsymbol{\theta_j}))^2 = (y - \hat{\mu}_j)^2 \tag{3.10}$$

as the quality measure of an approximation. Since we may or may not have the true model in the collection of candidate models, the learning problem becomes finding $\tilde{\mu}$, which minimizes the prediction risk functional,

$$R(\boldsymbol{\theta_j}) = \int L(y, g(\boldsymbol{x}|\boldsymbol{\theta_j}))p(\boldsymbol{x}, y)d\boldsymbol{x}dy \tag{3.11}$$

where $p(\boldsymbol{x}, y) = p(\boldsymbol{x})p(y|\boldsymbol{x})$ is an unknown joint distribution function generating the training data such that $f(\boldsymbol{x}) = \int yp(y|\boldsymbol{x})dy$ is the conditional mean obtained from the output regression function. Equation (3.11) measures the accuracy of the prediction of the true model made by model j.

Since $p(\boldsymbol{x}, y)$ and therefore $R(\boldsymbol{\theta_j})$ are not known, we estimate the parameters by minimizing the empirical risk

$$\hat{R}(\boldsymbol{\theta}_j) = \frac{1}{n}\sum_{i=1}^{n}(y_i - g(\boldsymbol{x_i}|\boldsymbol{\theta_j}))^2 = \frac{1}{n}\sum_{i=1}^{n}L(y_i, g(\boldsymbol{x}_i|\boldsymbol{\theta_j})) \tag{3.12}$$

By substituting $L(y_i, g(\boldsymbol{x}_i|\boldsymbol{\theta_j}))$ with the ϵ-insensitive loss function $L_\epsilon(y_i, g(\boldsymbol{x}_i|\boldsymbol{\theta_j}))$, we obtain the empirical risk for SVR

$$\begin{aligned}\hat{R}_\epsilon(\boldsymbol{\theta}_j) &= \frac{1}{n}\sum_{i=1}^{n} L_\epsilon(y_i, g(\boldsymbol{x}_i|\boldsymbol{\theta_j})) \\ &\equiv \frac{1}{n}\sum_{i=1}^{n} \max\left\{(|y_i - g(\boldsymbol{x_i}|\boldsymbol{\theta_j})| - \epsilon), 0\right\}\end{aligned} \tag{3.13}$$

as well as AIC and BIC given by

$$\text{AIC}_j^* = \hat{R}_\epsilon(\boldsymbol{\theta}_j) + 2p_j\frac{\sigma^2}{n} \tag{3.14a}$$

$$\text{BIC}_j^* = \hat{R}_\epsilon(\boldsymbol{\theta}_j) + \log(n)p_j\frac{\sigma^2}{n}. \tag{3.14b}$$

Accordingly, we extend the use of these model selection criteria for the calculation of sBIC and sAIC model averaging weights according to the following formulae:

$$w_j^* = \frac{\exp\left(-\text{BIC}_j^*/2\right)}{\sum_{k=1}^{K}\exp\left(-\text{BIC}_k^*/2\right)} \tag{3.15a}$$

$$w_j^* = \frac{\exp\left(-\text{AIC}_j^*/2\right)}{\sum_{k=1}^{K}\exp\left(-\text{AIC}_k^*/2\right)}. \tag{3.15b}$$

3.4 MMA Estimation for SVR

Since σ^2 in Equations (2.23) and (3.14) is unknown, we compute $C_n(\mathbf{w})$ by substituting σ^2 with the sample estimate $\hat{\sigma}_L^2 = (Y - \hat{\mu}_L))'(Y - \hat{\mu}_L))/(n - p_L) = \hat{\boldsymbol{e}}_L'\hat{\boldsymbol{e}}_L/(n - p_L) = \sum_{i=1}^{n}\hat{e}_{iL}^2/(n - p_L)$, or residual sum of squares (RSS) obtained from the "largest" dimensional model $L \in \{1, 2, \ldots, K\}$. The R package `quadprog` solves the quadratic programming problem shown in Equation (2.23), which minimizes Equation (2.25) as discussed in Section 2.1.4.

Assumption (2.2) is based on the premise that all candidates models are equally competitive, and this restriction is plausible in terms of allowing the data to determine the relative contribution of each candidate model to the final model averaging estimator. However, if there is no prior information that all candidate models are equally competitive, relaxing the restriction that the weights sum up to 1 is likely to lower the risk (Ando and Li, 2014). Therefore we remove this restriction to admit more general settings.

Chapter 4

Simulation Studies

We investigate the finite-sample performance of the selected FMA methods applied to multiple SVR models using simulated data from several data-generating processes (DGP) of analytic form. We aim to demonstrate that our FMA methods perform better on balance than model selection techniques, as we discussed in previous chapters, and therefore model averaging can contribute to more precise prediction obtained from a specific learning framework. Section 4.1 presents the entire structure of data generation for our simulation studies, followed by a brief description of the frequentist model averaging and selection methods used for estimation in Section 4.2. The results of the analysis are provided in Section 4.3.

4.1 Data Generation

4.1.1 Scenario I

We draw $R = 500$ Monte Carlo replications from each DGP. For each replication, we consider three DGPs of sample sizes n = (50, 100, 200, 400) which include a second-order polynomial, an exponential function, and a sinusoidal function as follows:

$$\text{DGP1} : y_i = x_{i1}+x_{i2}+x_{i3}+x_{i4}+x_{i1}^2+x_{i2}^2+x_{i1}x_{i2}+x_{i1}x_{i3}+x_{i1}x_{i4}+x_{i2}x_{i3}+x_{i2}x_{i4}+x_{i3}x_{i4}+e_i$$

$$\text{DGP2} : y_i = \exp\left(x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i1}x_{i2} + x_{i1}x_{i3} + x_{i1}x_{i4} + x_{i2}x_{i4} + x_{i3}x_{i4}\right) + e_i$$

$$\text{DGP3} : y_i = \sin\left(2\pi(x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i1}x_{i2} + x_{i1}x_{i3} + x_{i1}x_{i4} + x_{i2}x_{i4} + x_{i3}x_{i4})\right)+e_i$$

where $x_{il} : i \in \{1, \ldots, n\}; l \in \{1, \ldots, p = 4\}$ are realizations of independent and identically distributed Unif$(-1, 1)$ random variables. The error e_i is distributed $\mathcal{N}(0, \sigma^2)$ and independent of x_{il}. We set σ^2, the variance of e_i, so that the expected R^2 for the true model would be $(1 + \sigma^2)^{-1} = (0.95, 0.80, 0.50, 0.20)$, which corresponds to $\sigma = (0.25, 0.50, 1.00, 2.00)$ times the standard deviation of the systematic component of the DGP.

In Scenario I, the candidate models are under-specified, i.e., the collection of candidate models does not contain the true model. We estimate the following six models: (a) $y_i = g_1(x_{i1}) + e_i, (b)$ $y_i = g_2(x_{i2}) + e_i, (c)$ $y_i = g_3(x_{i1}, x_{i2}) + e_i, (d)$ $y_i = g_4(x_{i1}, x_{i3}) + e_i, (e)$ $y_i = g_5(x_{i2}, x_{i3}) + e_i, (f)$ $y_i = g_6(x_{i1}, x_{i2}, x_{i3}) + e_i$ For each of these six models, we use k = 5-fold cross-validation to tune the hyperparameters over the ranges of $\gamma, C \in 2^{\{-3,-2,1,0,1,2,3\}}$, $\epsilon \in \{0.1, 0.25, 0.5, 1\}$, and then estimate the models by the SVR technique using the RBF kernel as outlined above. As k gets larger, there are more combinations of the training set and the held-out test set, and we obtain less biased estimates of y_i in exchange for high variance. Although there is no formal rule for choosing k, $k = 5$ has been shown empirically as a value to yield error estimates "that suffer neither from excessively high bias nor from very high variance" in general (Hastie *et al.*, 2009).

4.1.2 Scenario II

The data generation for Scenario II is similar to that for Scenario I, except that the collection of candidate models contains the true model which is unlikely to occur in practice. Thus we set x_{i4} in the three DGPs for Scenario I to zero, and the signal-to-noise ratios are set as per Scenario I. We use the same collection of models, and for each model we use the same cross-validation procedure to tune the required hyperparameters. Lastly, we average over the obtained six estimates and assign the weights that minimize the MMA objective function as per Scenario I.

4.2 Methods Used for Estimation

For the estimation of the target vector $\mathbf{y} = (y_1, \ldots, y_n)'$, we consider seven estimators: (1) Mallows model averaging ('MMA'), (2) smoothed AIC model averaging ('sAIC'), (3) smoothed BIC model averaging ('sBIC'), (4) nonsmooth model averaging ('$1/K$'), (5) AIC model selection ('AIC'), (6) BIC model selection ('BIC'), and (7) Mallows' C_p model selection ('C_p'). The sAIC and sBIC weights, which we obtain from AIC* and BIC* in Section 3.3 respectively, are given by Equation (3.15). For the MMA approach, we average the obtained six estimates for $(a) - (f)$ in Section 4.1 by assigning them the weight vector $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_6)$ using the MMA criterion outlined in Section 2.1.4. Particularly, the weights are minimizing Equation (2.23), in which σ^2 is estimated based on residuals from the largest model. Results are summarized in Tables 4.1 - 4.6, which report the mean relative MSE row normalized such that the method with the lowest mean MSE has entry 1.00. R^2 is higher for smaller values of σ. MMA, sAIC, sBIC, and $1/K$ are model averaging methods. AIC, BIC, and C_p are model selection methods. Mean MMA weights are also available in Tables 4.7 - 4.8.

4.3 Results

In Scenario I, clearly no specific method dominates over the range of sample sizes and signal-to-noise ratios considered. sAIC and sBIC have higher risk with high signal-to-noise ratio than the other methods considered. The sAIC model averaging estimator performs better with the exponential DGP than with the other two. When the noise is very high, the nonsmooth model averaging estimator can be a simple and naïve choice with MSE as low as desired for practitioners. However, the model averaging methods are shown to be no less competitive than the model selection ones. Particularly, when we consider the range of empirical risk relative to the best performing estimator in the rows of Tables 4.1-4.3, from a minimax perspective the proposed MMA estimator is competitive among its peers.

Furthermore, Tables 4.4-4.6 show that the use of MMA can outperform model selection in small sample settings as per Scenario I, even though the collection of candidate models contains the true model. sAIC, sBIC, and nonsmooth model averaging do not result in predictions as accurate as they do in Scenario I. We also summarize the mean MMA weights for both scenarios in Tables 4.7-4.12 as outlined.

Table 4.1: Median Relative MSE, Scenario I, DGP 1

n	σ	MMA	sAIC	sBIC	$1/K$	AIC	BIC	C_p
50	0.25	1.00	2.25	2.23	2.39	1.21	1.20	1.35
	0.50	1.00	1.54	1.53	1.58	1.47	1.49	1.66
	1.00	1.06	1.00	1.00	1.00	1.39	1.39	1.17
	2.00	1.77	1.00	1.03	1.00	1.29	1.34	1.13
100	0.25	1.00	1.88	1.88	1.93	1.13	1.13	1.16
	0.50	1.00	1.57	1.57	1.59	1.60	1.88	1.50
	1.00	1.00	1.11	1.12	1.11	1.51	1.65	1.20
	2.00	1.41	1.01	1.04	1.00	1.40	1.42	1.16
200	0.25	1.00	1.75	1.75	1.78	1.10	1.11	1.10
	0.50	1.00	1.57	1.57	1.59	1.66	1.89	1.55
	1.00	1.00	1.27	1.28	1.27	1.52	1.89	1.35
	2.00	1.11	1.01	1.03	1.00	1.41	1.45	1.14
400	0.25	1.00	1.69	1.69	1.72	1.04	1.04	1.04
	0.50	1.00	1.56	1.56	1.57	1.72	1.82	1.67
	1.00	1.00	1.34	1.34	1.34	1.54	1.95	1.47
	2.00	1.00	1.09	1.11	1.09	1.54	1.64	1.19

Table 4.2: Median Relative MSE, Scenario I, DGP 2

n	σ	MMA	sAIC	sBIC	$1/K$	AIC	BIC	C_p
50	0.25	1.00	1.37	1.46	1.95	1.45	1.45	1.12
	0.50	1.00	1.26	1.28	1.41	1.33	1.29	1.29
	1.00	1.01	1.05	1.12	1.00	1.19	1.19	1.21
	2.00	1.46	1.13	1.30	1.00	1.31	1.40	1.07
100	0.25	1.00	1.25	1.29	1.90	1.38	1.38	1.05
	0.50	1.00	1.33	1.37	1.51	1.43	1.43	1.34
	1.00	1.00	1.38	1.38	1.19	1.42	1.42	1.34
	2.00	1.41	1.01	1.04	1.00	1.40	1.42	1.16
200	0.25	1.00	1.44	1.49	1.60	1.46	1.47	1.11
	0.50	1.00	1.49	1.51	1.47	1.55	1.54	1.62
	1.00	1.00	1.41	1.41	1.17	1.43	1.41	1.39
	2.00	1.15	1.13	1.12	1.00	1.14	1.18	1.12
400	0.25	1.00	1.51	1.55	1.52	1.54	1.56	1.19
	0.50	1.00	1.43	1.45	1.35	1.55	1.53	1.50
	1.00	1.00	1.43	1.43	1.25	1.46	1.46	1.41
	2.00	1.04	1.04	1.08	1.00	1.11	1.11	1.08

Table 4.3: Median Relative MSE, Scenario I, DGP 3

n	σ	MMA	sAIC	sBIC	$1/K$	AIC	BIC	C_p
50	0.25	1.00	1.58	1.58	1.58	2.00	1.98	2.04
	0.50	1.00	1.43	1.43	1.43	1.79	1.79	1.80
	1.00	1.00	1.04	1.04	1.04	1.27	1.28	1.25
	2.00	1.47	1.00	1.00	1.00	1.17	1.17	1.13
100	0.25	1.00	1.24	1.24	1.24	1.45	1.45	1.45
	0.50	1.00	1.21	1.20	1.20	1.40	1.40	1.39
	1.00	1.00	1.08	1.08	1.08	1.25	1.25	1.23
	2.00	1.19	1.00	1.00	1.00	1.13	1.13	1.10
200	0.25	1.00	1.15	1.15	1.15	1.28	1.28	1.26
	0.50	1.00	1.12	1.13	1.12	1.25	1.25	1.23
	1.00	1.00	1.08	1.09	1.08	1.19	1.19	1.17
	2.00	1.03	1.00	1.00	1.00	1.09	1.09	1.07
400	0.25	1.00	1.11	1.11	1.11	1.19	1.19	1.17
	0.50	1.00	1.11	1.11	1.10	1.19	1.19	1.17
	1.00	1.00	1.09	1.09	1.08	1.16	1.16	1.14
	2.00	1.00	1.02	1.02	1.02	1.08	1.08	1.07

Table 4.4: Median Relative MSE, Scenario II, DGP 1

n	σ	MMA	sAIC	sBIC	$1/K$	AIC	BIC	C_p
50	0.25	1.01	5.10	5.04	5.49	1.03	1.00	1.14
	0.50	1.00	1.96	1.94	2.03	1.23	1.14	2.26
	1.00	1.11	1.00	1.01	1.01	1.74	1.86	1.33
	2.00	1.92	1.00	1.04	1.00	1.41	1.49	1.18
100	0.25	1.00	6.85	6.79	7.30	1.00	1.00	1.03
	0.50	1.00	2.67	2.66	2.78	1.13	1.06	1.43
	1.00	1.00	1.30	1.31	1.31	1.92	2.55	1.60
	2.00	1.47	1.01	1.04	1.00	1.66	1.75	1.24
200	0.25	1.00	9.93	9.88	10.55	1.00	1.00	1.00
	0.50	1.00	3.75	3.74	3.91	1.06	1.04	1.08
	1.00	1.00	1.75	1.76	1.77	2.05	2.92	1.27
	2.00	1.07	1.01	1.04	1.00	1.80	1.90	1.28
400	0.25	1.00	15.12	15.07	16.03	1.00	1.00	1.00
	0.50	1.00	5.49	5.49	5.74	1.03	1.01	1.04
	1.00	1.00	2.47	2.49	2.53	1.14	1.30	1.12
	2.00	1.00	1.36	1.40	1.37	2.03	2.88	1.64

Table 4.5: Median Relative MSE, Scenario II, DGP 2

n	σ	MMA	sAIC	sBIC	$1/K$	AIC	BIC	C_p
50	0.25	1.00	4.46	3.62	7.24	1.02	1.03	1.03
	0.50	1.00	2.31	2.11	2.82	1.38	1.41	1.36
	1.00	1.00	1.13	1.24	1.11	1.73	1.77	1.13
	2.00	1.70	1.13	1.23	1.00	1.38	1.49	1.13
100	0.25	1.01	5.73	4.85	8.18	1.00	1.00	1.00
	0.50	1.01	3.01	2.83	3.61	1.00	1.09	1.04
	1.00	1.00	1.58	1.78	1.52	2.69	2.69	1.38
	2.00	1.25	1.16	1.25	1.00	1.43	1.45	1.10
200	0.25	1.01	8.11	7.39	10.62	1.00	1.00	1.00
	0.50	1.01	4.57	4.53	5.37	1.00	1.01	1.02
	1.00	1.00	2.32	2.64	2.27	3.89	4.14	1.48
	2.00	1.00	1.29	1.50	1.07	1.72	1.72	1.28
400	0.25	1.00	11.25	10.71	14.31	1.00	1.00	1.00
	0.50	1.02	6.45	6.35	7.65	1.00	1.00	1.00
	1.00	1.00	3.06	3.38	3.18	1.34	6.05	1.12
	2.00	1.00	1.66	2.00	1.45	2.48	2.50	1.42

Table 4.6: Median Relative MSE, Scenario II, DGP 3

n	σ	MMA	sAIC	sBIC	$1/K$	AIC	BIC	C_p
50	0.25	1.01	5.10	5.04	5.49	1.03	1.00	1.14
	0.50	1.00	1.73	1.73	1.74	2.35	2.33	2.34
	1.00	1.00	1.03	1.03	1.03	1.30	1.30	1.28
	2.00	1.51	1.00	1.00	1.01	1.18	1.18	1.15
100	0.25	1.00	6.98	6.97	6.99	10.17	10.17	9.73
	0.50	1.00	2.31	2.30	2.31	3.16	3.17	3.12
	1.00	1.00	1.09	1.09	1.09	1.32	1.34	1.31
	2.00	1.27	1.00	1.00	1.00	1.15	1.16	1.12
200	0.25	1.00	9.72	9.71	9.74	14.73	14.78	13.84
	0.50	1.00	3.02	3.02	3.01	4.34	4.36	4.23
	1.00	1.00	1.21	1.21	1.21	1.51	1.52	1.49
	2.00	1.07	1.00	1.00	1.00	1.12	1.13	1.10
400	0.25	1.00	10.63	10.61	10.59	16.09	16.11	15.68
	0.50	1.00	3.12	3.11	3.09	4.56	4.57	4.50
	1.00	1.00	1.43	1.42	1.41	1.83	1.83	1.81
	2.00	1.00	1.00	1.00	1.00	1.12	1.13	1.11

Table 4.7: Mean MMA Weights, DGP 1, Scenario I

n	σ	$\bar{w}_1$	$\bar{w}_2$	$\bar{w}_3$	$\bar{w}_4$	$\bar{w}_5$	$\bar{w}_6$
50	0.25	0.0144	0.0224	0.1172	0.0973	0.1089	0.6398
	0.50	0.0204	0.0288	0.1411	0.1132	0.1300	0.5666
	1.00	0.0420	0.0434	0.1750	0.1610	0.2047	0.3738
	2.00	0.0625	0.0674	0.1850	0.1967	0.2125	0.2758
100	0.25	0.0053	0.0042	0.0691	0.0478	0.0516	0.8220
	0.50	0.0080	0.0072	0.0806	0.0511	0.0652	0.7879
	1.00	0.0216	0.0203	0.1421	0.1085	0.1235	0.5840
	2.00	0.0490	0.0514	0.1909	0.1625	0.1927	0.3536
200	0.25	0.0012	0.0009	0.0207	0.0215	0.0177	0.9379
	0.50	0.0027	0.0016	0.0296	0.0181	0.0244	0.9236
	1.00	0.0084	0.0010	0.0537	0.0053	0.0543	0.8200
	2.00	0.0324	0.0369	0.1438	0.1605	0.1705	0.4559
400	0.25	0.0144	0.0224	0.1172	0.0973	0.1089	0.6398
	0.50	0.0204	0.0288	0.1411	0.1132	0.1300	0.5666
	1.00	0.0420	0.0434	0.1750	0.1610	0.2047	0.3738
	2.00	0.0625	0.0067	0.1850	0.1967	0.2125	0.2758

Table 4.8: Mean MMA Weights, DGP 2, Scenario I

n	σ	$\bar{w}_1$	$\bar{w}_2$	$\bar{w}_3$	$\bar{w}_4$	$\bar{w}_5$	$\bar{w}_6$
50	0.25	0.0393	0.0384	0.1902	0.1971	0.1680	0.3670
	0.50	0.0569	0.0408	0.2005	0.1886	0.1749	0.3384
	1.00	0.0609	0.0461	0.1994	0.1908	0.1739	0.3289
	2.00	0.0618	0.0592	0.1805	0.1931	0.1851	0.3204
100	0.25	0.0182	0.0173	0.1713	0.1685	0.1409	0.4839
	0.50	0.0202	0.0241	0.1823	0.1683	0.1404	0.4647
	1.00	0.0406	0.0253	0.1905	0.1800	0.1640	0.3996
	2.00	0.0409	0.0388	0.1875	0.1872	0.1623	0.3834
200	0.25	0.0052	0.0023	0.1519	0.1300	0.0836	0.6270
	0.50	0.0095	0.0040	0.1376	0.1442	0.0709	0.6334
	1.00	0.0184	0.0014	0.1267	0.1501	0.1033	0.5874
	2.00	0.0263	0.0202	0.1426	0.1692	0.1403	0.5014
400	0.25	0.0000	0.0000	0.0975	0.0946	0.0390	0.7689
	0.50	0.0031	0.0004	0.0775	0.0878	0.0425	0.7886
	1.00	0.0077	0.0044	0.0783	0.0924	0.0608	0.7564
	2.00	0.0177	0.0140	0.1357	0.1386	0.0966	0.5974

Table 4.9: Mean MMA Weights, DGP 3, Scenario I

n	σ	$\bar{w}_1$	$\bar{w}_2$	$\bar{w}_3$	$\bar{w}_4$	$\bar{w}_5$	$\bar{w}_6$
50	0.25	0.0427	0.0257	0.2090	0.1763	0.1986	0.3476
	0.50	0.0489	0.0235	0.2037	0.2012	0.2136	0.3092
	1.00	0.0435	0.0362	0.2090	0.1880	0.2066	0.3168
	2.00	0.0581	0.0432	0.1751	0.2001	0.2340	0.2895
100	0.25	0.0291	0.0214	0.1302	0.1665	0.1733	0.4794
	0.50	0.0339	0.0235	0.1321	0.1566	0.1630	0.4910
	1.00	0.0297	0.0224	0.1345	0.1693	0.1613	0.4829
	2.00	0.0296	0.0217	0.1520	0.1999	0.1630	0.4339
200	0.25	0.0150	0.0012	0.1304	0.1082	0.1165	0.6183
	0.50	0.0206	0.0084	0.1621	0.1332	0.1199	0.5558
	1.00	0.0229	0.0097	0.1651	0.1412	0.1350	0.5261
	2.00	0.0195	0.0119	0.1456	0.1591	0.1360	0.5279
400	0.25	0.0096	0.0038	0.0691	0.0822	0.0945	0.7408
	0.50	0.0133	0.0067	0.0076	0.0926	0.0779	0.7337
	1.00	0.0126	0.0096	0.0821	0.0908	0.0791	0.7259
	2.00	0.0118	0.0068	0.1068	0.0986	0.0656	0.7104

Table 4.10: Mean MMA Weights, DGP 1, Scenario II

n	σ	$\bar{w}_1$	$\bar{w}_2$	$\bar{w}_3$	$\bar{w}_4$	$\bar{w}_5$	$\bar{w}_6$
50	0.25	0.0005	0.0003	0.0218	0.0066	0.0095	0.9613
	0.50	0.0042	0.0057	0.0577	0.0422	0.0456	0.8445
	1.00	0.0243	0.0353	0.1152	0.1281	0.1324	0.5647
	2.00	0.0625	0.0667	0.1891	0.2016	0.1991	0.2730
100	0.25	0.0004	0.0001	0.0072	0.0008	0.0022	0.9893
	0.50	0.0020	0.0010	0.0186	0.0090	0.0127	0.9567
	1.00	0.0106	0.0111	0.0778	0.0533	0.0554	0.7918
	2.00	0.0370	0.0382	0.1763	0.1427	0.1648	0.4409
200	0.25	0.0001	0.0001	0.0011	0.0010	0.0008	0.9968
	0.50	0.0005	0.0003	0.0064	0.0029	0.0039	0.9860
	1.00	0.0030	0.0039	0.0242	0.0182	0.0174	0.9333
	2.00	0.0250	0.0250	0.1143	0.1030	0.1302	0.6025
400	0.25	0.0001	0.0002	0.0004	0.0005	0.0004	0.9985
	0.50	0.0003	0.0001	0.0009	0.0011	0.0010	0.9966
	1.00	0.0005	0.0014	0.0060	0.0085	0.0043	0.9792
	2.00	0.0054	0.0103	0.0398	0.0479	0.0315	0.8650

Table 4.11: Mean MMA Weights, DGP 2, Scenario II

n	σ	$\bar{w}_1$	$\bar{w}_2$	$\bar{w}_3$	$\bar{w}_4$	$\bar{w}_5$	$\bar{w}_6$
50	0.25	0.0008	0.0003	0.0548	0.0491	0.0469	0.8481
	0.50	0.0025	0.0045	0.0491	0.1068	0.0552	0.7819
	1.00	0.0243	0.0353	0.1152	0.1281	0.1324	0.5647
	2.00	0.0551	0.0475	0.1696	0.1916	0.2006	0.3355
100	0.25	0.0000	0.0000	0.0368	0.0370	0.0207	0.9054
	0.50	0.0004	0.0010	0.0612	0.0628	0.0343	0.8402
	1.00	0.0085	0.0087	0.1064	0.1089	0.0566	0.7109
	2.00	0.0383	0.0329	0.1804	0.1757	0.1364	0.4364
200	0.25	0.0000	0.0000	0.0175	0.0158	0.0031	0.9637
	0.50	0.0001	0.0001	0.0359	0.0268	0.0145	0.9226
	1.00	0.0019	0.0017	0.0448	0.0498	0.0300	0.8718
	2.00	0.0118	0.0129	0.1011	0.1089	0.0902	0.6751
400	0.25	0.0000	0.0000	0.0051	0.0029	0.0008	0.9913
	0.50	0.0002	0.0001	0.0126	0.0091	0.0034	0.9748
	1.00	0.0009	0.0008	0.0170	0.0155	0.0046	0.9613
	2.00	0.0050	0.0055	0.0362	0.0466	0.0243	0.8825

Table 4.12: Mean MMA Weights, DGP 3, Scenario II

n	σ	$\bar{w}_1$	$\bar{w}_2$	$\bar{w}_3$	$\bar{w}_4$	$\bar{w}_5$	$\bar{w}_6$
50	0.25	0.0225	0.0174	0.1365	0.1271	0.1408	0.5557
	0.50	0.0401	0.0261	0.1308	0.1367	0.1548	0.5114
	1.00	0.0288	0.0370	0.1590	0.1658	0.1788	0.4306
	2.00	0.0464	0.0610	0.1668	0.1900	0.1818	0.3539
100	0.25	0.0097	0.0034	0.0375	0.0375	0.0465	0.8654
	0.50	0.0186	0.0036	0.0682	0.0569	0.0592	0.7934
	1.00	0.0249	0.0159	0.1071	0.0908	0.1059	0.6553
	2.00	0.0262	0.0263	0.1386	0.1404	0.1508	0.5177
200	0.25	0.0020	0.0001	0.0007	0.0017	0.0003	0.9951
	0.50	0.0024	0.0002	0.0084	0.0067	0.0082	0.9741
	1.00	0.0089	0.0021	0.0284	0.0421	0.0431	0.8754
	2.00	0.0107	0.0058	0.0849	0.1014	0.0957	0.7015
400	0.25	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
	0.50	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
	1.00	0.0000	0.0000	0.0035	0.0002	0.0056	0.9906
	2.00	0.0082	0.0020	0.0362	0.0331	0.0357	0.8848

Chapter 5

Empirical Illustration

We estimate a Mincer (earnings) equation using Wooldridge (2002)'s 'wage1' cross-sectional data with $n = 526$ observations, which he obtained from the 1976 United States Current Population Survey. We consider modelling the log of hourly wages ($y = \log(wage)$) based on a range of commonly used predictors, namely:

(1) $x_1 = educ$: years of education

(2) $x_2 = exper$: years of professional experiences

(3) $x_3 = female$: $\mathbf{1}\{i\text{th observation is female}\}$

(4) $x_4 = tenure$: years with the current employer.

5.1 Analysis

We treat the four predictors described above as belonging to X, and consider SVR models which differ in terms of the contents of X. Let d be the order of a polynomial constructed from each of $x_1, \dots, x_4$. When $d = 1$, X is a $n \times 4$ matrix, and the number of all possible combinations of the predictors should be $K = \sum_{q=1}^{4} \binom{4}{q} =$

$\binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 4 + 6 + 4 + 1 = 15$. We also consider a standard linear model ('OLS') defined with the full set of predictors for comparison. The candidate models are not completely nested in each other as were those generated in Chapter 4, and the 'wage1' data has a binary predictor.

We repeatedly shuffle and split the data into a training set of size $n_1 = 500$ and an independent test set of size $n_2 = 26$. For each training set, we fit the cross-validated SVR models, and then we apply the selected model averaging and model selection methods presented in Section 4.2, as well as the models listed above. Lastly, for each model fit we compute the mean squared prediction error (MSPE) for the independent test set given by $\text{MSPE} = \sum_{i=1}^{n_2} (y_i - \hat{\mu}_i)^2 / n_2$ where $\hat{\mu}_i$ refers to an out-of-sample prediction.

5.2 Results

Table 5.1: Median Relative MSPE

MMA	sAIC	sBIC	$1/K$	AIC	BIC	C_p	OLS
1.00	1.07	1.07	1.07	1.14	1.29	1.09	1.06

Table 5.2: Median MSPE

MMA	sAIC	sBIC	$1/K$	AIC	BIC	C_p	OLS
0.1543	0.1643	0.1643	0.1643	0.1763	0.1990	0.1675	0.1641

Table 5.1 shows us similar relative MSPE results as seen in the previous Section 4.3. In particular, the MMA estimator dominates its peers in terms of out-of-sample prediction error. The other three model averaging estimators outperform model selection methods as well. It is interesting to note that MMA not only achieves an

improvement in prediction error over the standard least-squares linear model, but also is compatible with the non-nested candidate models, as well as the binary (categorical) variable $female$ in the set of predictors.

Chapter 6

Concluding Remarks

We have applied model averaging to support vector learning models with continuous target variables within the frequentist framework. Despite its performance, FMA has not received much interest from practitioners of statistical learning methods compared to BMA. In particular, recent notable investigations in the model averaging for SVR focused on the Bayesian approach (Liao *et al.*, 2011; Wang and Liao, 2012; Wang *et al.*, 2014; Kaneko and Funatsu, 2014). We specified the SVR models with the RBF kernel which are verified as proper for estimating nonlinear DGPs. We have established: (1) the naïve nonsmooth model weight $w = 1/K$, (2) sAIC and sBIC model weights obtained from from AIC and BIC using ϵ-insensitive loss function, and (3) MMA weights, which minimizes Hansen (2007)'s Mallows-type criterion.

We investigated the finite-sample performance of the proposed methods through Monte Carlo simulation studies. In a completely nonparametric setting where we do not have the true model as one of the candidate models, the model averaging estimators under consideration performed better than model selection estimators. With large signal-to-noise ratios, our proposed MMA estimator is especially competitive among its peers, and it tends to perform better than other estimators even with smaller signal-to-noise ratios as the sample size increases. The nonsmooth model weight is also a feasible choice when we draw a small random sample with a small signal-to-noise ratio. Furthermore, even when the true model is included in the set of candidate models, the MMA estimator dominates model selection, which is not likely to be the case in general.

The construction of new model averaging weights which minimize the absolute loss, or more specifically the ϵ-insensitive loss for SVR, would be an interesting extension. Although the MMA estimator can outperform its model selection peers, asymptotic optimality on the L^1 space does yet have a strong theoretical foundation. This approach can be extended to the study of FMA for SVM classifiers where Li and Yang (2002) have already suggested model weights based on maximum likelihood and in-sample prediction accuracy. Past work on model selection criteria for SVM such as Claeskens *et al.* (2008) and Zhang *et al.* (2016) can also be utilized for sAIC- or sBIC-type model weights, whose foundations have been better established.

Lastly, we highlight that the marginal effect of each predictor in the model should be measured more accurately after model averaging, in accordance with the increasing prediction accuracy. Nevertheless, many popular learning algorithms, including but not limited to SVM, neural networks, and random forests operate as "black boxes," so practitioners usually have limited access to intuitive interpretation. Given the development of methods that present highly plausible interpretations of how a learning system arrives at a prediction (e.g., Layer-Wise Relevance Propagation by Bach *et al.* (2015); Ranked SVM by Joachims (2002)), the adoption of model averaging will be useful for further refinement of such interpretations.

Appendix A

R Codes

A.1 Assessment of the Finite-Sample Performance

We provide the R code to replicate the Monte Carlo simulations using DGP 1 under Scenario I, where the set of candidate models do not contain the true model. The DGP used, and the ranges of SVR hyperparameter values can be modified.

```
rm(list = ls())
# Initialization
library(e1071)
# library(kernlab) can be used as an alternative; codes for entertaining
# the learning method should be modified accordingly

set.seed(42)
reps <- 500 # Number of replication
n <- 50  # Sample sizes (50, 100, 200, 400)
i <- 1

y <- numeric() # Our Target Variable
dgp <- numeric()
mse.mma.0 <- numeric()
mse.aic.0 <- numeric()
mse.bic.0 <- numeric()
mse.saic.0 <- numeric()
mse.sbic.0 <- numeric()
mse.mcp.0 <- numeric()
mse.nonsmooth.0 <- numeric()

# The vector of mean model averaging weights for each candidate model
w.mean.mma <- numeric()
w.mean.saic <- numeric()
w.mean.sbic <- numeric()

# The vectors which store the model averaging weights from each replication
B <- numeric()
W.saic.col  <- numeric()
W.sbic.col  <- numeric()



# Ranges of some values of hyperparameter gamma, C, and epsilon
# Users may add or remove proper hyperparameters they want to tune,
# depending on which kernel function they use.
# Note that expanding the range of hyperparameter values may result in
# excessive computation time.
ranges <- list(gamma = 2^(-3:3), cost = 2^(-3:3),
               epsilon = c(0.1, 0.25, 0.5, 1))
write(c("MMA","sAIC", "sBIC", "w=1/K", "AIC", "BIC", "Cp"),
      file="res.out",ncol=7)

# Random samples from other continuous distributions may be entertained.
for(count in 1:reps) {
  x1 <- runif(n, -1,1)
  x2 <- runif(n, -1,1)
  x3 <- runif(n, -1,1)
  x4 <- runif(n, -1,1)

  # DGP 1
  dgp <- x1 + x2 + x3 + x4 + x1**2 + x2**2 + x1*x2
         + x1*x3 + x1*x4 + x2*x3 + x2*x4 + x3*x4
```

```
# Users may comment out DGP 1, and instead entertain one
# of the other two DGPs presented below:

# DGP 2
# dgp <- sin(2*pi*(x1 + x2 + x3 + x4 + x1*x2
#         + x1*x3 + x1*x4 + + x2*x4 + x3*x4))

# DGP 3
# dgp <- exp(x1 + x2 + x3 + x4 + x1*x2
#         + x1*x3 + x1*x4 + + x2*x4 + x3*x4)

# S-N ratios vary over (0.25, 0.50, 1.00, 2.00) times the standard deviation of
# our data-generating processes, whose corresponding R^2 should be
# (0.95, 0.80, 0.50, 0.20).
y <- dgp + rnorm(n,sd=.25*sd(dgp))

# CV for hyperparameter selection
# "cross = 10" in the argument "tune.control" leads to 10-fold CV; "cross = n"
# to leave-one-out CV
search.1 <- tune(svm, y ~ x1 + I(x1^2), ranges = ranges,
                 tunecontrol = tune.control(sampling = "cross", cross = 5))
search.2 <- tune(svm, y ~ x2 + I(x2^2), ranges = ranges,
                 tunecontrol = tune.control(sampling = "cross", cross = 5))
search.3 <- tune(svm, y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1*x2), ranges = ranges,
                 tunecontrol = tune.control(sampling = "cross", cross = 5))
search.4 <- tune(svm, y ~ x1 + x3 + I(x1^2) + I(x1*x3), ranges = ranges,
                 tunecontrol = tune.control(sampling = "cross", cross = 5))
search.5 <- tune(svm, y ~ x2 + x3 + I(x2^2) + I(x2*x3), ranges = ranges,
                 tunecontrol = tune.control(sampling = "cross", cross = 5))
search.6 <- tune(svm, y ~ x1 + x2 + x3 + I(x1^2) + I(x2^2)
                          + I(x1*x2) + I(x1*x3) + I(x2*x3), ranges = ranges,
                 tunecontrol = tune.control(sampling = "cross", cross = 5))

# Best models are specified based on the lowest generalization error
model.1.1<- search.1$best.model
model.1.2<- search.2$best.model
model.1.3<- search.3$best.model
model.1.4<- search.4$best.model
model.1.5<- search.5$best.model
model.1.6<- search.6$best.model

# Store all the candidate models in a vector, so that we can
# conduct AIC / BIC / Cp model selection over the set later.
models <- rbind(model.1.1,model.1.2,model.1.3,model.1.4,model.1.5,model.1.6)

# Gives out the number of predictors in each model using support vectors
beta.1 = t(model.1.1$coefs) %*% model.1.1$SV
beta.2 = t(model.1.2$coefs) %*% model.1.2$SV
beta.3 = t(model.1.3$coefs) %*% model.1.3$SV
beta.4 = t(model.1.4$coefs) %*% model.1.4$SV
beta.5 = t(model.1.5$coefs) %*% model.1.5$SV
beta.6 = t(model.1.6$coefs) %*% model.1.6$SV

# The number of predictors in each candidate model.
K <- c(length(beta.1),length(beta.2),length(beta.3),
       length(beta.4),length(beta.5),length(beta.6))

residual.mat <- cbind(model.1.1$residuals, model.1.2$residuals,
                      model.1.3$residuals, model.1.4$residuals,
                      model.1.5$residuals, model.1.6$residuals)

# Residual sum of squares for each candidate model
RSS <- c(sum((residual.mat[,1])^2), sum((residual.mat[,2])^2),
         sum((residual.mat[,3])^2), sum((residual.mat[,4])^2),
         sum((residual.mat[,5])^2), sum((residual.mat[,6])^2))

M <- ncol(residual.mat)

# Computing RSS from a full model
# For convenience I set the last model as a full one.
sigsq <- RSS[length(K)] / n
```

```
# Epsilon-insensitive loss function
e.1 <- (abs(dgp.1-fitted(model.1.1))-model.1.1$epsilon)
e.2 <- (abs(dgp.1-fitted(model.1.2))-model.1.2$epsilon)
e.3 <- (abs(dgp.1-fitted(model.1.3))-model.1.3$epsilon)
e.4 <- (abs(dgp.1-fitted(model.1.4))-model.1.4$epsilon)
e.5 <- (abs(dgp.1-fitted(model.1.5))-model.1.5$epsilon)
e.6 <- (abs(dgp.1-fitted(model.1.6))-model.1.6$epsilon)

# If an absolute residual is larger than the pre-defined epsilon, we do ignore it.
e.1 <- sum(e.1*(e.1>0))
e.2 <- sum(e.2*(e.2>0))
e.3 <- sum(e.3*(e.3>0))
e.4 <- sum(e.4*(e.4>0))
e.5 <- sum(e.5*(e.5>0))
e.6 <- sum(e.6*(e.6>0))

E<- c(e.1, e.2, e.3, e.4, e.5, e.6)
R<- E/n # Empirical Risk

# AIC and BIC
AIC <- R + 2*(K)*RSS/n/(n-K)
BIC <- R + log(n)*(K)*RSS/n/(n-K)

# Mallows’ Cp
Cp <- E^2 / sigsq - n + 2*K

# Model selection using AIC, BIC, and Cp
for(v in 1:length(RSS)) {
  if(Cp[v]==min(Cp)) {
    model.mcp <- models[v,]
  }
  if(AIC[v]==min(AIC)) {
    model.aic <- models[v,]
  }
  if(BIC[v]==min(BIC)) {
    model.bic <- models[v,]
  }
```

```
}

# MMA Estimator
require(quadprog)
D <- t(residual.mat)%*%residual.mat
D <- D + diag(1e-5,M,M)
A <- cbind(rep(1,M),diag(1,M,M))
b0 <- c(1,rep(0,M))
d <- -sigsq*K
b <- solve.QP(Dmat=D,dvec=d,Amat=A,bvec=b0,meq=1)$solution
b.col <- as.matrix(b, nrow = 1)
B <- cbind(B, b.col)

# sAIC, sBIC Estimators
w.saic <- exp(-AIC / 2) / sum(exp(-AIC / 2))
w.sbic <- exp(-BIC / 2) / sum(exp(-BIC / 2))
w.saic.col <- as.matrix(w.saic, nrow = 1)
w.sbic.col <- as.matrix(w.sbic, nrow = 1)

if(!anyNA(w.saic.col)) {
  W.saic.col <- cbind(W.saic.col, w.saic.col)
}
if(!anyNA(w.sbic.col)) {
  W.sbic.col <- cbind(W.sbic.col, w.sbic.col)
}

# MSEs
mse.mma.0[i] <- mean(((b[1]*fitted(model.1)+b[2]*fitted(model.2)+b[3]*fitted(model.3)
               +b[4]*fitted(model.4)+b[5]*fitted(model.5)+b[6]*fitted(model.6))-dgp)^2)
mse.saic.0[i] <- mean(((w.saic[1]*fitted(model.1)+w.saic[2]*fitted(model.2)
               +w.saic[3]*fitted(model.3)+w.saic[4]*fitted(model.4)
               +w.saic[5]*fitted(model.5)+w.saic[6]*fitted(model.6))-dgp)^2)
mse.sbic.0[i] <- mean(((w.sbic[1]*fitted(model.1)+w.sbic[2]*fitted(model.2)
               +w.sbic[3]*fitted(model.3)+w.sbic[4]*fitted(model.4)
               +w.sbic[5]*fitted(model.5)+w.sbic[6]*fitted(model.6))-dgp)^2)

mse.nonsmooth.0[i] <- mean(((fitted(model.1)+fitted(model.2)+fitted(model.3)
                    +fitted(model.4)+fitted(model.5)+fitted(model.6))
                    / nrow(models)-dgp)^2)
mse.aic.0[i] <- mean((fitted(model.aic)-dgp)^2)
mse.bic.0[i] <- mean((fitted(model.bic)-dgp)^2)
mse.mcp.0[i] <- mean((fitted(model.mcp)-dgp)^2)

i <- i+1

# We do not consider the replications where the sum of AIC/BIC is so small that the
    denominator of sAIC / sBIC model weights approaches 0.
mse.mma <- mse.mma.0[which(!is.na(mse.saic.0))]
mse.mma <- mse.mma.0[which(!is.na(mse.sbic.0))]
mse.nonsmooth <- mse.nonsmooth.0[which(!is.na(mse.saic.0))]
mse.nonsmooth <- mse.nonsmooth.0[which(!is.na(mse.sbic.0))]
mse.aic <- mse.aic.0[which(!is.na(mse.saic.0))]
mse.aic <- mse.aic.0[which(!is.na(mse.sbic.0))]
mse.bic <- mse.bic.0[which(!is.na(mse.saic.0))]
mse.bic <- mse.bic.0[which(!is.na(mse.sbic.0))]
mse.mcp <- mse.mcp.0[which(!is.na(mse.saic.0))]
mse.mcp <- mse.mcp.0[which(!is.na(mse.sbic.0))]
mse.saic <- mse.saic.0[which(!is.na(mse.saic.0))]
mse.saic <- mse.saic.0[which(!is.na(mse.sbic.0))]
mse.sbic <- mse.sbic.0[which(!is.na(mse.saic.0))]
mse.sbic <- mse.sbic.0[which(!is.na(mse.sbic.0))]

# Boxplot comparing the performance of model selection and model averaging
# methods under investigation
boxplot(data.frame(mse.mma, mse.saic, mse.sbic, mse.nonsmooth, mse.aic, mse.bic, mse.mcp
    ), notch=TRUE, outline=FALSE,
        main=paste(formatC(median(mse.mma),digits=4,format="g"),
                   formatC(median(mse.saic),digits=4,format="g"),
                   formatC(median(mse.sbic),digits=4,format="g"),
                   formatC(median(mse.nonsmooth),digits=4,format="g"),
                   formatC(median(mse.aic),digits=4,format="g"),
                   formatC(median(mse.bic),digits=4,format="g"),
                   formatC(median(mse.mcp),digits=4,format="g")),
        sub=paste("MSE for DGP, R = ",count,sep=""))
```

```
}

for (j in 1:length(K)) {
  w.mean.mma[j] <- mean(B[j,])
  w.mean.saic[j] <- mean(W.saic.col[j,])
  w.mean.sbic[j] <- mean(W.sbic.col[j,])
}

# Save the results
write(c(median(mse.mma),median(mse.saic),median(mse.sbic),median(mse.nonsmooth),
        median(mse.aic),median(mse.bic),median(mse.mcp))
      / min(median(mse.mma),median(mse.saic),median(mse.sbic),median(mse.nonsmooth),
            median(mse.aic),median(mse.bic),median(mse.mcp)),
      file="res.out",ncol=7, append = T)
write(w.mean.saic, file="res.out",ncol=6, append = T)
write(w.mean.sbic, file="res.out",ncol=6, append = T)
```

Bibliography

Agarwal, S. and Sureka, A. (2015). Using knn and svm based one-class classifier for detecting online radicalization on twitter. In R. Natarajan, G. Barua, and M. R. Patra, editors, *Distributed Computing and Internet Technology, 11th International Conference, ICDCIT 2015, Bhubaneswar, India, February 5-8, 2015. Proceedings*, ICDCIT 2015, pages 431–442, Cham, Switzerland. Springer.

Aizerman, M. A., Braverman, E. M., and Rozonoer, L. I. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, **25**(6), 821–837.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csáki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, Hungary. Akadémiai Kiadó.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Akaike, H. (1978). On the likelihood of a time series model. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **27**(3/4), 217–235.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, **16**(1), 3 – 14.

Ando, T. and Li, K.-C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, **109**(505), 254–265.

Awad, M. and Khanna, R. (2015). *Support Vector Regression*, pages 67–80. Apress, Berkeley, CA.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, **10**(7), 1–46.

Bahlmann, C., Haasdonk, B., and Burkhardt, H. (2002). Online handwriting recognition with support vector machines - a kernel approach. In *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 49–54.

Barnard, G. A. (1963). New methods of quality control. *Journal of the Royal Statistical Society. Series A (General)*, **126**(2), 255–258.

Basak, D., Pal, S., and Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, **11**(10), 203–224.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *OR*, **20**(4), 451–468.

Bennett, K. P. and Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, **1**(1), 23–34.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT ‘92, pages 144–152, New York, NY. ACM Press.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, **16**(3), 199–231.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, **53**(2), 603–618.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**(2), 121–167.

Burnham, K. P. and Anderson, D. R. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33**(2), 261–304.

Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, **33**(2), 201 – 208.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **158**(3), 419–466.

Chen, K.-Y. and Wang, C.-H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, **28**(1), 215–226.

Cherkassky, V. and Ma, Y. (2004). Practical selection of svm parameters and noise estimation for svm regression. *Neural Networks*, **17**(1), 113 – 126.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge, United Kingdom.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2008). An information criterion for variable selection in support vector machines. *Journal of Machine Learning Research*, **9**, 541–558.

Cortes, C. and Vapnik, V. N. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 45–97.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., and Vapnik, V. N. (1996). Support vector regression machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS '96, Cambridge, MA. MIT Press.

Drucker, H., Wu, D., and Vapnik, V. N. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, **10**(5), 1048–1054.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**(10), 906–914.

Garrett, D., Peterson, D. A., Anderson, C. W., and Thaut, M. H. (2003). Comparison of linear, nonlinear, and feature selection methods for EEG signal classification.

IEEE Transactions on Neural Dystems and Rehabilitation Engineering, **11**(2), 141–144.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**(423), 881–889.

Girosi, F., Jones, M., and Poggio, T. (1993). Priors stabilizers and basis functions: From regularization to radial, tensor and additive splines. Technical Report AIM 1430 / CBCL 75, Massachusetts Institute of Technology, Cambridge, MA.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, **3**, 1157–1182.

Guyon, I. M., Weston, J., Barnhill, S., and Vapnik, V. N. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**(1), 389–422.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, **75**(4), 1175–1189.

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, **167**(1), 38–46.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, New York, NY.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, **98**(464), 879–899.

Hodges, J. S. (1987). Uncertainty, policy analysis and statistics. *Statistical Science*, **2**(3), 259–275.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, **14**(4), 382–401.

Huang, M.-W., Chen, C.-W., Lin, W.-C., Ke, S.-W., and Tsai, C.-F. (2017). SVM and SVM ensembles in breast cancer prediction. *PLOS ONE*, **12**(1), 1–14.

Huang, W., Nakamori, Y., and Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, **32**(10), 2513 – 2522.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**(2), 297–307.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY. ACM.

Kaneko, H. and Funatsu, K. (2014). Adaptive soft sensor based on online support vector regression and bayesian ensemble learning for various states in chemical plants. *Chemometrics and Intelligent Laboratory Systems*, **137**, 57 – 66.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural computation*, **13**(3), 637–649.

Kim, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, **55**(1-2), 307–319.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**(1), 79–86.

Kwok, J. T. (2001). Linear dependency between ϵ and the input noise in ϵ-support vector regression. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Artificial Neural Networks — ICANN 2001, International Conference Vienna, Austria, August 21-25, 2001. Proceedings*, ICANN 2001, pages 405–410, Berlin, Germany. Springer.

Leamer, E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. A Wiley-Interscience publication. Wiley.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, **15**(3), 958–975.

Li, W. and Yang, Y. (2002). How many genes are needed for a discriminant microarray data analysis. In S. M. Lin and K. F. Johnson, editors, *Methods of Microarray Data Analysis: Papers from CAMDA '00*, pages 137–140. Springer, Boston, MA.

Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*, **106**(495), 1053–1066.

Liao, S., Zhao, N., and Zhao, Z. (2011). Bayesian-model-averaging-based model combining method on regularization path of support vector machines.

Liu, C.-A. and Kuo, B.-S. (2016). Model averaging in predictive regressions. *The Econometrics Journal*, **19**(2), 203–231.

Long, N., Gianola, D., Rosa, G. J. M., and Weigel, K. A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. *Theoretical and Applied Genetics*, **123**(7), 1065.

Lu, C.-J., Lee, T.-S., and Chiu, C.-C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, **47**(2), 115 – 125.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**(4), 661–675.

Mattera, D. and Haykin, S. (1999). Support vector machines for dynamic reconstruction of a chaotic system. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods*, pages 211–241. MIT Press, Cambridge, MA.

Micchelli, C. A. (1986). Algebraic aspects of interpolation. In *Proceedings of Symposia in Applied Mathematics*, volume 36, pages 81–102. American Mathematical Society.

Momma, M. and Bennett, K. P. (2002). A pattern search method for model selection of support vector regression. In R. Grossman, J. Han, V. Kumar, H. Mannila, and R. Motwani, editors, *Proceedings of the 2002 SIAM International Conference on Data Mining*, pages 261–274, Philadelphia, PA. SIAM.

Müller, K. R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik, V. N. (1997). Predicting time series with support vector machine. In *Artificial Neural Networks — ICANN '97, 7th International Conference Lausanne, Switzerland, October 8–10, 1997. Proceedings*, ICANN '97, pages 999–1004, Berlin, Germany. Springer.

Myasnikova, E., Samsonova, A., Samsonova, M., and Reinitz, J. (2002). Support vector regression applied to the determination of the developmental age of a drosophila embryo from its segmentation gene expression patterns. *Bioinformatics*, **18**(suppl_1), S87–S95.

Niu, X.-X. and Suen, C. Y. (2012). A novel hybrid CNN–SVM classifier for recognizing handwritten digits. *Pattern Recognition*, **45**(4), 1318 – 1325.

Racine, J. S. (1997). Feasible cross-validatory model selection for general stationary processes. *Journal of Applied Econometrics*, **12**(2), 169–179.

Racine, J. S. (2017). *ma: Model Averaging.* R package ver. 1.0-8.

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. Bollen and J. Long, editors, *Testing Structural Equation Models*, pages 163–180. SAGE, Newbury Park, CA.

Roberts, H. V. (1965). Probabilistic prediction. *Journal of the American Statistical Association*, **60**(309), 50–62.

Schölkopf, B., Burges, C. J. C., and Vapnik, V. N. (1995). Extracting support data for a given task. In U. M. Fayyad and R. Uthurusamy, editors, *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, KDD '95, pages 252–257, Menlo Park, CA. AAAI Press.

Schölkopf, B., Burges, C. J. C., and Smola, A. J. (1999). *Advances in Kernel Methods.* MIT Press, Cambridge, MA.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Shawe-Taylor, J. and Cristianini, N. (1998). Robust bounds on generalization from the margin distribution. Technical Report NC2-TR-1988-029, ESPRIT Working Group in Neural and Computational Learning II.

Shin, K.-S., Lee, T. S., and Kim, H. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, **28**(1), 127 – 135.

Smith, F. W. (1968). Pattern classifier design by linear programming. *IEEE Transactions on Computers*, **C-17**(4), 367–372.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, **14**(3), 199–222.

Smola, A. J., Murata, N., Schölkopf, B., and Müller, K.-R. (1998a). Asymptotically optimal choice of ϵ-loss for support vector machines. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, 2–4 September 1998*, ICANN 98, pages 105–110, London, United Kingdom. Springer.

Smola, A. J., Schölkopf, B., and Müller, K.-R. (1998b). General cost functions for support vector regression. In T. Downs, M. Frean, and M. Gallagher, editors, *Proceedings of the Ninth Australian Conference on Neural Networks*, ACNN '98, pages 79–83, St. Lucia, Australia. University of Queensland.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**(2), 111–147.

Subasi, A. and Gursoy, M. I. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, **37**(12), 8659–8666.

Sun, B., Zhu, Z., Li, J., and Linghu, B. (2011). Combined feature selection and cancer prognosis using support vector machine regression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **8**(6), 1671–1677.

Thulasidas, M., Guan, C., and Wu, J. (2006). Robust classification of EEG signal for brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **14**(1), 24–29.

Tong, S. and Koller, D. (2002). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, **2**, 45–66.

Van Gestel, T., Suykens, J. A. K., Baestaens, D.-E., Lambrechts, A., Lanckriet, G., Vandaele, B., De Moor, B., and Vandewalle, J. (2001). Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Transactions on Neural Networks*, **12**(4), 809–821.

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory.* Springer, Berlin, Germany.

Vapnik, V. N. and Lerner, A. Y. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, **24**(6), 774–780.

Vapnik, V. N., Golowich, S. E., and Smola, A. J. (1996). Support vector method for function approximation, regression estimation and signal processing. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Proceedings of the 9th International*

Conference on Neural Information Processing Systems, NIPS ‘96, pages 281–287, Cambridge, MA. MIT Press.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **46**(4), 433–448.

Wan, A. T., Zhang, X., and Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, **156**(2), 277 – 283.

Wang, M. and Liao, S. (2012). Model combination for support vector regression via regularization path. In P. Anthony, M. Ishizuka, and D. Lukose, editors, *PRICAI 2012: Trends in Artificial Intelligence*, pages 649–660, Berlin, Germany. Springer.

Wang, M., Song, K., Lv, H., and Liao, S. (2014). Consistent model combination for svr via regularization path. *Journal of Computational Information Systems*, **10**(22), 9609–9617.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**(1), 92–107.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.

Wu, C.-H., Ho, J.-M., and Lee, D. T. (2004). Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*, **5**(4), 276–281.

Zhang, X., Wan, A. T., and Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, **174**(2), 82 – 94.

Zhang, X., Wu, Y., Wang, L., and Li, R. (2016). A consistent information criterion for support vector machines in diverging model spaces. *Journal of Machine Learning Research*, **17**(16), 1–26.