Estimating the Market Share Attraction Model using Support Vector Regressions

Georgi I. Nalbantov*† Philip Hans Franses* Patrick J. F. Groenen*
Jan C. Bioch*

Econometric Institute Report EI2007-06

Abstract

We propose to estimate the parameters of the Market Share Attraction Model (Cooper & Nakanishi, 1988; Fok & Franses, 2004) in a novel way by using a non-parametric technique for function estimation called Support Vector Regressions (SVR) (Vapnik, 1995; Smola, 1996). Traditionally, the parameters of the Market Share Attraction Model are estimated via a Maximum Likelihood (ML) procedure, assuming that the data are drawn from a conditional Gaussian distribution. However, if the distribution is unknown, ML estimation may seriously fail (Vapnik, 1982). One way to tackle this problem is to introduce a linear loss function over the errors and a penalty on the magnitude of model coefficients. This leads to qualities such as robustness to outliers and avoidance of the problem of overfitting. This kind of estimation forms the basis of the SVR technique, which, as we will argue, makes it a good candidate for solving the Market Share Attraction Model. We test the SVR approach to predict (the evolution of) the market shares of 36 car brands simultaneously and report stronger results than when using a ML estimation procedure.

1 Introduction

The Market Share Attraction Model is a popular tool for analyzing market competitive structures (Cooper & Nakanishi, 1988; Fok & Franses, 2004). It is typically applied for si-

*Econometric Institute, School of Economics, Erasmus University Rotterdam
†Erasmus Research Institute of Management, Erasmus University Rotterdam

multaneously predicting the market shares of several brands within a given product category simultaneously. The model helps to evaluate the effect of marketing-mix variables on brands' performances as well as the effect of an individual brand's own efforts while conditioning on competitors' reactions. A detailed econometric analysis of the model can be found in Fok, Franses, and Paap (2002). What makes this model rather special is the requirement that the forecasted market shares are all non-negative and sum to unity.

The traditional unrestricted Market Share Attraction Model often suffers from poor predictability, especially for the relatively larger brands. The poor performance is likely to be due to various causes, including heteroscedasticity and failure to account for a trend in the data. The huge number of coefficients to be estimated is another source of concern. A common way to address those issues is to restrict the model coefficients or to aggregate brands into categories. More fundamentally, however, one can also address the commonly applied estimation procedure, which is Maximum Likelihood (ML). ML estimation is appropriate (and optimal) in cases the dependent variable has been drawn from a conditional Gaussian distribution. In cases where this is not so, the least-squares techniques are suboptimal and could lead to severely mismatched solutions for some densities (Vapnik, 1982). In cases like this, improved coefficient estimation can be obtained by using estimation methods put forward in the literature on Support Vector Machines (SVMs) and this is what we address in this paper.

SVMs are a nonparametric tool that can be used for both classification and regression estimation tasks (Vapnik, 1995; Burges, 1998; Cristianini & Shawe-Taylor, 2000). They have gained considerable popularity during the last years, following a series of successful applications in areas ranging from Bioinformatics and Optical Character Recognition to Economics and Finance (see, among others, Schölkopf, Guyon, & Weston, 2001; Schölkopf, Burges, & Vapnik, 1995; Pérez-Cruz, Afonso-Rodríguez, & Giner, 2003; Tay & Cao, 2001).

SVMs, and in particular Support Vector Regression (SVR), capitalize on two facts. The first one is the proposition that the linear loss function is the best error loss function of the worst model over any probability density function of the dependent variable given the independent variables (Huber, 1964). Thus, if the dependent variable is drawn from an unknown distribution, a linear loss function over the errors could be more appropriate than the common quadratic one. The second building block is the bound obtained on the test error (less than infinity) using the so-called Structural Minimization Principle (Vapnik, 1995).

This bound arises when a certain error-insensitive region around the predicted value of the dependent variable is introduced. The width of this region can be made arbitrarily small however. Therefore, if the dependent variable has not been sampled from a conditional Gaussian distribution, the SVR can lead to better predictions than those obtained via a least squares (ML) procedure (Pérez-Cruz et al., 2003).

SVRs solve a quadratic programming problem to obtain a solution. Unlike competing techniques such as Neural Networks, this solution is unique and does not suffer from a local minimum. Further desirable properties of SVRs in general are their ability to avoid in-sample overfitting and their robustness against outliers in the data, and particularly those properties make them very suitable for the application to the Market Share Attraction Model.

The paper is organized as follows. The next section introduces the Market Share Attraction Model in its traditional form. Section 3 outlines the SVR technique and augments it with SVR estimation and Section 4 discusses its nonlinear extension. Section 5 presents our main findings on a data set that is used to predict the evolution of market shares of 36 car brands for a certain period. The final section gives a conclusion.

2 The Market Share Attraction Model

The purpose of the Market Share Attraction Model is to provide an overall model for the market share $M_{i,t}$ of brand i at time t for the I brands constituting the market over a period from $t = 1$ to T. An important characteristic of a market share $M_{i,t}$ is that $0 \leq M_{i,t} \leq 1$ and that it sums over all brands to one, that is, $\sum_{i=1}^{I} M_{i,t} = 1$. The typical interval between the measurements of the market shares is a week or a month. The model uses K predictor variables with nonnegative values $x_{k,i,t}$ to predict the market shares described below. Typical predictor variables are price, distribution, advertising spending, etcetera. The usefulness of the model lies in its ability to describe the competitive structures and to infer cross effects of marketing-mix instruments (Fok et al., 2002).

The so-called Multiplicative Competitive Interaction (MCI) specification of a market share $M_{i,t}$ builds on the attraction $A_{i,t}$ of brand i at time t that is defined as

$$A_{i,t} = \exp(\mu_i + \varepsilon_{i,t}) \prod_{j=1}^{I} \prod_{k=1}^{K} x_{k,j,t}^{\beta_{k,j,i}} \quad \text{for } i = 1, \ldots, I, \tag{1}$$

where $\beta_{k,j,i}$ is the unknown coefficient for brand i and μ_i is a brand-specific intercept term corresponding to the size of the brand. The vector of error terms $\boldsymbol{\varepsilon}_t = [\varepsilon_{1,t}, \ldots, \varepsilon_{I,t}]'$ is

usually assumed to be normally distributed with zero mean and some unknown covariance matrix $\mathbf{\Sigma}$. The market share of brand i at time t can be defined as the attraction of brand i at t divided by the sum of all attractions at t, that is,

$$M_{i,t} = \frac{A_{i,t}}{\sum_{j=1}^{I} A_{j,t}} \qquad \text{for } i = 1, \ldots, I. \tag{2}$$

The model in (1) with (2) is the Market Share Attraction Model. Notice that the definition of the market share of brand i at time t given in (2) implies that the attraction of the product category is the sum of the attractions of all brands and that equal attraction of two brands results in equal market shares.

In addition to the predictor variables $x_{k,i,t}$, one could also include lagged variables $x_{k,i,t-1}, x_{k,i,t-2}, \ldots, x_{k,i,t-P}$ and lagged market shares $M_{i,t-1}, M_{i,t-2}, \ldots, M_{i,t-P}$ as predictors. With these P lags, the attraction $A_{i,t}$ specification with a P-th order autoregressive structure becomes

$$A_{i,t} = \exp(\mu_i + \varepsilon_{i,t}) \prod_{j=1}^{I} \left(\prod_{k=1}^{K} x_{k,j,t}^{\beta_{k,j,i}} \prod_{p=1}^{P} \left(M_{j,t-p}^{\alpha_{p,j,i}} \prod_{k=1}^{K} x_{k,j,t-p}^{\beta_{p,k,j,i}} \right) \right), \tag{3}$$

where $\alpha_{p,j,i}$ is the effect of lagged market shares on the attraction and $\beta_{p,k,j,i}$ the effect of lagged explanatory variables. Clearly, this specification involves quite a number of parameters.

To estimate the parameters, the model is linearized in two steps. First, we choose brand I as a benchmark brand and express the market share of each of the remaining brands as a fraction of this benchmark brand, that is,

$$\frac{M_{i,t}}{M_{I,t}} = \frac{\exp(\mu_i + \varepsilon_{i,t}) \prod_{j=1}^{I} \left(\prod_{k=1}^{K} x_{k,j,t}^{\beta_{k,j,i}} \prod_{p=1}^{P} \left(M_{j,t-p}^{\alpha_{p,j,i}} \prod_{k=1}^{K} x_{k,j,t-p}^{\beta_{p,k,j,i}} \right) \right)}{\exp(\mu_I + \varepsilon_{I,t}) \prod_{j=1}^{I} \left(\prod_{k=1}^{K} x_{k,j,t}^{\beta_{k,j,I}} \prod_{p=1}^{P} \left(M_{j,t-p}^{\alpha_{p,j,I}} \prod_{k=1}^{K} x_{k,j,t-p}^{\beta_{p,k,j,I}} \right) \right)}. \tag{4}$$

The second step is to take the natural logarithm (denoted by log) of both sides of (4). Together, these two steps result in the $(I-1)$-dimensional set of equations given by

$$\begin{aligned} \log M_{i,t} - \log M_{I,t} \quad = \quad & (\mu_i - \mu_I) + \sum_{j=1}^{I} \sum_{k=1}^{K} (\beta_{k,j,i} - \beta_{k,j,I}) \log x_{k,j,t} \\ & + \sum_{p=1}^{P} \sum_{j=1}^{I} (\alpha_{p,j,i} - \alpha_{p,j,I}) \log M_{j,t-p} \\ & + \sum_{p=1}^{P} \sum_{j=1}^{I} \sum_{k=1}^{K} (\beta_{p,k,j,i} - \beta_{p,k,j,I}) \log x_{k,j,t-p} + \eta_{i,t}. \end{aligned} \tag{5}$$

Because the μ_i parameters only appear as the difference $\mu_i - \mu_I$ with the benchmark parameter μ_I, they are not uniquely identified. However, the parameters $\tilde{\mu}_i = \mu_i - \mu_I$

are uniquely identified. Similarly, $\tilde{\beta}_{k,j,i} = \beta_{k,j,i} - \beta_{k,j,I}$, $\tilde{\beta}_{p,k,j,i} = \beta_{p,k,j,i} - \beta_{p,k,j,I}$, and $\tilde{\alpha}_{p,j,i} = \alpha_{p,j,i} - \alpha_{p,j,I}$ can also be uniquely identified. Therefore, for estimation we use $\tilde{\mu}_i, \tilde{\beta}_{k,j,i}, \tilde{\beta}_{p,k,j,i}$, and $\tilde{\alpha}_{p,j,i}$.

The errors $\eta_{i,t}$ in (5) are equal to $\eta_{i,t} = \varepsilon_{i,t} - \varepsilon_{I,t}$, or, equivalently, $\boldsymbol{\eta}_t = \mathbf{L}\boldsymbol{\varepsilon}_t$ with the $(I-1) \times I$ matrix $\mathbf{L} = [\mathbf{I} \mid -\mathbf{1}]$ where $\mathbf{I}$ an $(I-1)$-dimensional identity matrix and $\mathbf{1}$ is an $(I-1)$-vector of ones. Hence, given the earlier assumptions that $\boldsymbol{\varepsilon}_t$ is normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\eta}_t$ is also normally distributed with mean $\mathbf{0}$ and a $(I-1) \times (I-1)$ covariance matrix equal to $\tilde{\boldsymbol{\Sigma}} = \mathbf{L}\boldsymbol{\Sigma}\mathbf{L}'$. As a consequence, out of the $I(I+1)/2$ unknown (co)variances in $\boldsymbol{\Sigma}$, we can only identify $I(I-1)/2$ values.

Using the substitution above to obtain unique estimates for the effects, the general attraction model in (5) can be expressed as an $(I-1)$-dimensional P-th order vector autoregression with exogenous variables, that is, by

$$\begin{aligned} \log M_{i,t} - \log M_{I,t} &= \tilde{\mu}_i + \sum_{j=1}^{I}\sum_{k=1}^{K} \tilde{\beta}_{k,j,i} \log x_{k,j,t} + \sum_{p=1}^{P}\sum_{j=1}^{I} \tilde{\alpha}_{p,j,i} \log M_{j,t-p} \\ &\quad + \sum_{p=1}^{P}\sum_{j=1}^{I}\sum_{k=1}^{K} \tilde{\beta}_{p,k,j,i} \log x_{k,j,t-p} + \eta_{i,t}. \end{aligned} \tag{6}$$

Under the assumption that the error variables are normally distributed with some unknown covariance matrix, maximum likelihood (ML) is the appropriate estimation method. In our application, the explanatory variables for each brand are the same, that is, $x_{k,1,t} = x_{k,2,t} = \ldots = x_{k,I,t}$. Under these conditions and if there are no parameter restrictions then ordinary least squares (OLS) estimators are equal to the ML estimator (Fok et al., 2002).

If the dependent variable has not been drawn from a conditional normal distribution, then the parameters of the general Market Share Attraction Model (6) are not guaranteed to be optimally estimated by a least-squares technique (Vapnik, 1982). An alternative way to estimate the model parameters in this case is by means of the suggested SVR, which is outlined below.

3 Linear Support Vector Regression

Support Vector Regressions (SVRs) and Support Vector Machines (SVMs) are rooted in the Statistical Learning Theory, pioneered by Vapnik (1995) an co-workers. Detailed treatments of SVR and SVM can be found, for example, in Burges (1998), Smola (1996) and Smola and Schölkopf (1998). The following is a self-contained basic introduction to Support Vector

Regressions (SVRs).

SVRs have two main strengths and these are good generalizability/avoidance of overfitting and robustness against outliers. Generalizability refers to the fact that SVRs are designed in such a way that they provide the most simple solution for a given, fixed amount of (training) errors. A function is referred to as being simple if the coefficients of the predictor variables are penalized towards zero. Thus, an SVR addresses the problem of overfitting explicitly, just like many other penalization methods such as Ridge Regression (Tikhonov, 1963) and Lasso (Tibshirani, 1996). The robustness property stems from considering absolute, instead of quadratic, values for the errors. As a consequence, the influence of outliers is less pronounced. More precisely, SVRs employ the so-called ϵ-insensitive error loss function, which is presented below. To put it in a nutshell, (linear) SVR departs from the classical regression in two aspects. The first one is the utilization of the ϵ-insensitive loss function instead of the quadratic one. The second aspect is the penalization of the vector of coefficients of the predictor variables.

The classical multiple regression has a well known loss function that is quadratic in the errors, $r_i^2 = (y - f(\mathbf{x}_i))^2$. The loss function employed in SVR is the ϵ-insensitive loss function

$$g(r_i) = |y_i - f(\mathbf{x}_i)|_\epsilon \equiv \max\{0, |y_i - f(\mathbf{x}_i)| - \epsilon \} = \max\{0, |r_i| - \epsilon\}$$

for a predetermined nonnegative ϵ, where y_i is the true target value, $\mathbf{x}_i$ is a vector of input variables and $f(\mathbf{x}_i)$ is the estimated target value for observation i. Figure 1 shows the resulting function for the residual. Intuitively speaking, if the absolute residual is off-target by ϵ or less, then there is no loss, that is, no penalty should be imposed, hence the name "ϵ-insensitive". However, if the opposite is true, that is $|y_i - f(\mathbf{x})| - \epsilon > 0$, then a certain amount of loss should be associated with the estimate. This loss rises linearly with the absolute difference between y and $f(\mathbf{x})$ above ϵ.

Because SVR is a nonparametric method, traditional parametric inferential statistical theory cannot be readily applied. Theoretical justifications for the SVR are instead based on statistical learning theory (Vapnik, 1995). There are two sets of model parameters in SVR: coefficients, and a manually-adjustable parameter C that explicitly controls the interplay between model fit and model complexity. For each value of the manually-adjustable parameter C there is a corresponding set of optimal coefficients, which are obtained by solving a quadratic optimization problem that is tuned using a cross-validation procedure. In such a procedure, the data set is first partitioned into several mutually exclusive parts. Next,

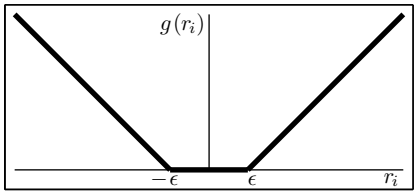

Figure 1: The ϵ-insensitive loss function that assigns no penalty to residuals $r_i \in [f(\mathbf{x}_i) - \epsilon, f(\mathbf{x}_i) + \epsilon]$ for point i. As $|r_i|$ gets larger than ϵ, a nonzero penalty $g(r_i)$ that rises linearly is assigned.

models are built on some parts of the data and other parts are used for evaluation of the fit-versus-complexity parameter C. This is quite analagous to the process of adjusting the bias-versus-variance parameter in Ridge Regression, for instance. We start out with assuming implicitly that the fit-versus-complexity parameter C has been set to unity, and later relax that assumption. In the nonlinear SVR case, other manually-adjustable parameters may arise. Then a grid search over a certain range of values for C and these parameters has to be performed.

Let us first consider the case of simple linear regression estimation by SVR by the usual linear relation $y = \beta_1 x_1 + b$, where β_1 and b are parameters to be estimated. Figure 2 shows an example with three cases of possible linear functional relations. The SVR line is the solid line in Figure 2c, given by the equation $f(x_1) = \beta_1 x_1 + b$. The "tube" between the dotted lines in Figure 2 consists of points for which the inequality $|y - f(x_1)| - \epsilon \leq 0$ holds, where ϵ has been fixed arbitrarily at 2. All data points that happen to be on or inside the tubes are not associated with any loss. The rest of the points will be penalized according to the ϵ-insensitive loss function. Hence, the solutions in Panel (b) and (c) both have zero loss in ϵ-insensitive sense.

The exact position of the SVR line of Figure 2c is determined as follows. The starting point is that the SVR line should be as horizontal/simple/flat as possible. The extreme case of $\beta_1 = 0$ in Figure 2a will unavoidably yield several mistakes, as ϵ is not big enough to give zero loss for all points. This case represents a simple but quite "lousy" relationship. However, notice that the resulting region between the dotted lines, referred to as the ϵ-insensitive region, occupies the greatest possible area (for $\epsilon = 2$). It is argued in the SVR literature that this particular area can be seen as a measure of the complexity of the regression function used. Accordingly, the horizontal regression line provides the least

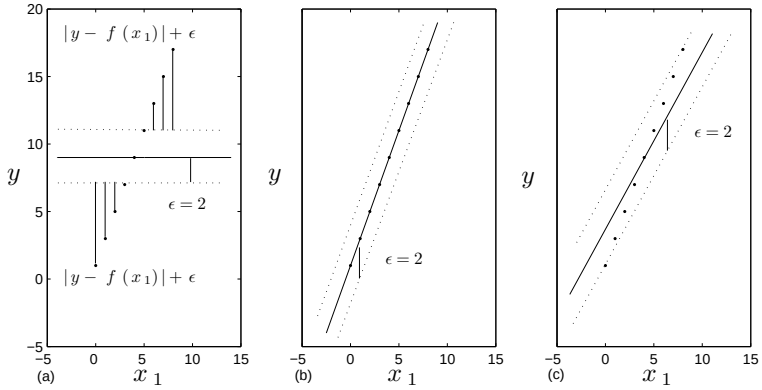

Figure 2: Three possible solutions to a linear regression problem with data points that lie on a line. The vertical line segments in panel (a) indicate loss per observation, which is equal to $|y - f(x_1)| - \epsilon$, for $\epsilon = 2$. In line with the ϵ-insensitive loss function, a point is not considered to induce an error if its deviation from the regression line is less than or equal to ϵ. The horizontal regression line in panel (a) is the simplest possible one since it hypothesizes that there is no relation between y and x_1, and it produces too much loss. Panel (b) gives the classical linear regression estimation, yielding zero loss. Panel (c) shows the linear SVR, which also yields zero loss but it flatter than the regression in Panel (b).

complex functional relationship between x_1 and y, which is equivalent to no relationship at all.

Consider the next step in Figure 2b. Here, the solid line fits the training data extremely well. This line is the actual regression function from classical regression analysis, where the loss measured as the sum of squared errors of the estimates is being minimized. The distance between the dotted lines however has clearly diminished as compared to Figures 2a and 2c. What the SVR line of Figure 2c aims for is to find a balance between the amount of "flatness" (or *complexity*) and training mistakes (or *fit*). This balance is the fundamental idea behind SVR analysis. Good generalization ability is achieved when the best trade-off between function's complexity (proxied by the distance between the dotted lines) and function's accuracy on the training data is being struck. The idea that such a balance between complexity and amount of training errors should be searched has been formalized in Vapnik (1995).

To find a linear relationship between p independent variables and a single dependent variable in a data set of n observations, the mathematical formulation of the optimization problem of SVR can be derived intuitively as follows. The objective is to find a vector of p coefficients $\boldsymbol{\beta}$ and an intercept b so that the linear function $f(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} + b$ has the best generalization ability for some fixed ϵ error insensitivity. From the "complexity" side,

this linear surface should be as horizontal as possible, which can be achieved by minimizing the quadratic form $\boldsymbol{\beta}'\boldsymbol{\beta}$. From the "amount of errors" side however, a perfectly horizontal surface (obtained for $\boldsymbol{\beta} = \mathbf{0}$) will generally not be optimal since a lot of errors will typically be made in such a case. According to the ϵ-insensitive loss function, the sum of these errors is defined to be equal to $\sum_{i=1}^{n} g(r_i) = \sum_{i=1}^{n} \max\{0, |y_i - f(\mathbf{x}_i)| - \epsilon\}$. One can strike a balance between amount of errors and complexity by minimizing their sum

$$L_p(\boldsymbol{\beta}, b) \quad := \quad \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta} + C\sum_{i=1}^{n} \max\{0, |y_i - (\boldsymbol{\beta}'\mathbf{x}_i + b)| - \epsilon\},$$

where C is a user-defined constant that controls the relative importance of the two terms. This minimization problem formulation is the familiar *penalty* plus *loss* minimization paradigm that arises in many domains (see, e.g., Hastie, Tibshirani, & Friedman, 2001).

The problem can equivalently be represented by introducing the so-called slack variables $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$. Then, minimizing $L_p(\boldsymbol{\beta}, b)$ can be represented as the constrained minimization problem

$$\begin{aligned} \text{minimize } L_p(\boldsymbol{\beta}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*) \quad &:= \quad \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta} + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \\ \text{subject to} \qquad & \quad y_i - (\boldsymbol{\beta}'\mathbf{x}_i + b) \leq \epsilon + \xi_i, \\ & \quad \boldsymbol{\beta}'\mathbf{x}_i + b - y_i \leq \epsilon + \xi_i^*, \text{ and} \\ & \quad \xi_i, \xi_i^* \geq 0 \end{aligned} \tag{7}$$

(Vapnik, 1995; Smola & Schölkopf, 1998).

If the estimate $\boldsymbol{\beta}'\mathbf{x}_i + b$ of the i^{th} observation deviates from the target y_i by more than ϵ, then a loss is incurred. This loss is equal to either ξ_i or ξ_i^*, depending on which side of the regression surface observation i lies. It turns out that (7) is a convex quadratic optimization problem with linear constraints, and thus a unique solution can always be found. As already mentioned, the objective function in (7) consists of two terms. The first term, $\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$, captures the degree of complexity, which is proxied by the width of the ϵ-insensitive region between surfaces $y = \boldsymbol{\beta}'\mathbf{x} + b + \epsilon$ and $y = \boldsymbol{\beta}'\mathbf{x}_i + b - \epsilon$. If $\boldsymbol{\beta} = \mathbf{0}$, then complexity ($\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$) is minimal since the ϵ-insensitive region is biggest. The slack variables variables ξ_i and ξ_i^*, $i = 1, 2, \ldots, n$, are constrained to be nonnegative. All points i inside the ϵ-insensitive region have both $\xi_i = 0$ and $\xi_i^* = 0$. If a point i lies outside the ϵ-insensitive region, then either $\xi_i > 0$ and $\xi_i^* = 0$, or $\xi_i = 0$ and $\xi_i^* > 0$. All data points that lie outside the ϵ-insensitive region (that is, for which $|y - f(\mathbf{x}_i| \geq \epsilon$) are called "support vectors". It can be shown that

the final solution for the SVR line depends only on the support vectors, and thus all other points are completely irrelevant (Smola & Schölkopf, 1998). This property is referred to as the sparse-solution property of SVR. In other words, the final formulation of the SVR function would remain the same even if all data points that are not support vectors were removed from the original data set.

Generally, it is not possible to have both terms $\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$ and $C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$ equal to zero. If $\boldsymbol{\beta}=\mathbf{0}$, then the loss $C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$ can be large, as depicted in Figure 2a. Likewise, if the sum $C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$ is relatively small, then $\boldsymbol{\beta}$ will generally be large, and consequently $\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$ too. Therefore, at the minimum of the objective function in (7), a balance is found between $\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$ (complexity) and $C\sum_{i=1}^{n}(\xi_i+\xi_i^*)$ (fit), ensuring that neither the resulting function $f(x_1)=\boldsymbol{\beta}'\mathbf{x}+b$ fits the data too well, nor that it is too flat. The constraints in the optimization problem ensure that the degenerate solution $\boldsymbol{\beta}=\boldsymbol{\xi}=\boldsymbol{\xi}^*=\mathbf{0}$ is avoided.

4 Nonlinear Support Vector Regression

Another useful feature of the SVR is that nonlinear relationships can be easily included. This property may be useful in the Market Share Attraction Model if there is a nonlinear relation between the log attraction differences and the predictor variables.

4.1 Preliminaries

To introduce nonlinearities in SVR estimation, we need to discuss an alternative computational solution to the so-called primal linear minimization problem defined in (7). In particular, instead of minimizing L_p directly, a dual representation is used. Thus, the unknown parameters of the linear SVR $\boldsymbol{\beta}$, b, ξ_i and $\xi_i^*, i=1,2,\ldots,n$ of the original primal (7) can be found as the unique solution of the dual problem, where the dual is defined as

$$\begin{aligned}\text{maximize } L_d(\boldsymbol{\alpha}) := \; & -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(\mathbf{x}_i'\mathbf{x}_j)+ \\ & +\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)y_i-\epsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*) \\ \text{subject to } \; & 0\le\alpha_i,\alpha_i^*\le C, i=1,2,\ldots,n \text{ and } \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0,\end{aligned} \tag{8}$$

where the unknowns α_i and α_i^* are the Lagrange multipliers of the primal (for a step-by-step derivation of the dual see, for example, Vapnik (1995), Smola (1996)). They are the weights

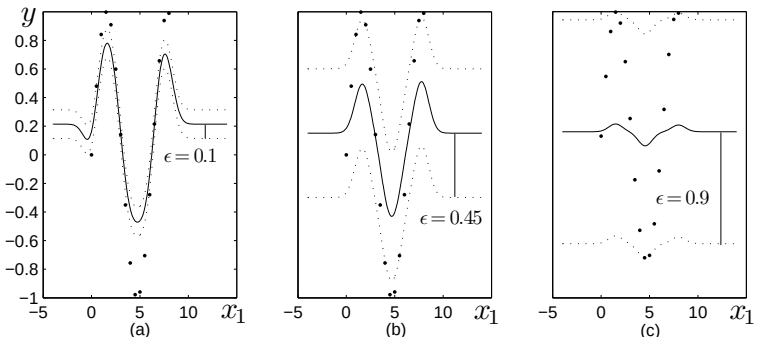

Figure 3: Three possible nonlinear SVR solutions to the problem of estimating the function $y = sin(x_1)$ from examples.

associated with each data point i. If both α_i and α_i^* for point i are equal to zero, then this point lies inside the ϵ-insensitive region. It has a weight of zero and plays no role for the final formulation of the SVR function. The SVR regression function takes the form of (Smola & Schölkopf, 1998):

$$f(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)(\mathbf{x}'\mathbf{x}_i) + b, \tag{9}$$

where $\mathbf{x}$ is a vector containing the values of the independent variables for a new (test) point. Note that since the SVR regression function can be expressed as $f(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} + b$, it follows that $\boldsymbol{\beta}'\mathbf{x} = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)(\mathbf{x}'\mathbf{x}_i)$ at the optimum, and therefore model coefficients are obtained as $\boldsymbol{\beta} = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)\mathbf{x}$.

4.2 Nonlinear SVR

Now we introduce nonlinear SVR estimation. The construction of nonlinear SVR is carried out in two steps. First, the data are mapped through $\mathbf{x} \rightarrow \boldsymbol{\Phi}(\mathbf{x})$ into a *higher*-dimensional space. Second, a linear regression function is constructed in the transformed space. This function corresponds to a nonlinear one in the original, non-transformed space. The optimal linear regression function in the transformed space should be, analogically to the non-transformed case, as flat as possible (Smola & Schölkopf, 1998) to ensure a good generalization ability. Due to the mapping $\mathbf{x} \rightarrow \boldsymbol{\Phi}(\mathbf{x})$, the SVR estimates in the nonlinear case take the form (Smola & Schölkopf, 1998):

$$f(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) k(\mathbf{x}, \mathbf{x}_i) + b,$$

where $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i)'\mathbf{\Phi}(\mathbf{x}_j)$ is a so-called kernel function that computes dot products in a transformed space. Often, the Gaussian kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i)'\mathbf{\Phi}(\mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2}$ is used, where γ is a manually adjustable parameter that controls the degree of similarity between any two vectors in the transformed space. Note that coefficient estimates for nonlinear SVR are available only if the mapping $\mathbf{x} \rightarrow \mathbf{\Phi}(\mathbf{x})$ is carried out explicitly. The coefficients are then calculated as $\boldsymbol{\beta} = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\mathbf{\Phi}(\mathbf{x})$. Other kernels exist but are beyond the scope of this paper.

Let us now consider Figure 3, which shows 17 sample points (black dots) from the function $y = sin(x_1)$. Three possible nonlinear SVR solutions to this problem are given in this figure. By construction there is no noise in the data. The nonlinear transformation of the original data is carried out via the Gaussian kernel. All SVR manually adjustable parameters are the same in all three panels, except for ϵ, which is equal to 0.1, 0.45 and 0.9 in panels (a), (b), and (c), respectively. As ϵ increases, the estimated functional relationship between y and x_1 becomes weaker (and therefore flatter); furthermore, the amount of errors reduces substantially. Notice that the estimated relationship also becomes flatter as x_1 takes on values that are farther away from the values of the original data points, which is an attractive property of SVR for extrapolation.

So far the question of how to choose the manually adjustable parameters (such as C, ϵ and γ) has been left aside. One very common way to proceed is to use a k-fold cross-validation procedure. In a such a procedure, the data set is split in k (equally-sized) parts. Then, k models for a fixed set of values for the manually adjustable parameters are built on $k-1$ folders and each time the one remaining folder is used for validation (or, testing). The chosen parameters are those that produce minimal mean squared error on average (over all k test parts).

4.3 Links between SVR and Classical Regression

The classical OLS approach to function estimation is to find the vector of coefficients $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ and intercept term $b = b^*$, which minimize the loss $\mathrm{L}_{OLS}(\boldsymbol{\beta}, b) = \sum_{i=1}^{n}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - b)^2$, where $\{y_i, \mathbf{x}_i\}$, $i = 1, 2, \ldots n$, is a data point. The Ridge Regression approach extends OLS by modifying the loss to $L_{RR}(\boldsymbol{\beta}, b) = \lambda\boldsymbol{\beta}'\boldsymbol{\beta} + \sum_{i=1}^{n}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - b)^2$, for $\lambda \geq 0$. Hence, the linear SVR, OLS, and Ridge Regression optimization problems can be thought of special

cases of the general optimization problem

$$\begin{aligned} \text{minimize } L_p^{\text{All}}(\boldsymbol{\beta}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*) \quad &:= \quad \frac{A}{2}\boldsymbol{\beta}'\boldsymbol{\beta} + \frac{C}{k}\sum_{i=1}^{n}((\xi_i)^k + (\xi_i^*)^k) \\ \text{subject to} \quad & y_i - \boldsymbol{\beta}'\mathbf{x}_i - b \leq \epsilon + (\xi_i^*)^k, \\ & \boldsymbol{\beta}'\mathbf{x}_i + b - y_i \leq \epsilon + (\xi_i)^k, \text{ and} \\ & \xi_i, \xi_i^* \geq 0, \\ & \text{for } i = 1, 2, \ldots, n, \end{aligned} \tag{10}$$

where $\epsilon \geq 0, k \in \{1, 2\}, A \geq 0, C > 0$. The classical linear regression optimization problem is a special case of (10), where $k = 2$, $\epsilon = 0$, $C = 2$, and $A = 0$. The linear ridge regression estimation problem is obtained for $k = 2$, $\epsilon = 0$, $C = 2$, and $A = 2\lambda$. Finally , the linear SVR estimation problem (7) corresponds to the restrictions $k = 1$, $\epsilon > 0$, $C > 0$, and $A = 1$.

5 An Illustration for New Cars

The technique of SVR might be particularly useful for the Market Share Attraction Model as it is not certain that the log of the market shares are conditionally Gaussian and also as the log transformation can create outlying data points. Here we present and compare the results of SVR and ML estimation of the coefficients of the Market Share Attraction Model on empirical data. Carrying out an extensive benchmark study is beyond the scope of the present paper and we refer to Pérez-Cruz et al. (2003) for a number of simulation studies. They report superior performance of SVR vis-a-vis ML coefficient estimation in cases where the dependent variable has not been drawn from a conditional normal distribution as well as in cases where the distribution is actually normal, but with small sample size.

5.1 Description of the Data

The data are monthly sales figures per brand of new cars, in the Netherlands starting in January 1992 and ending in October 2001 obtained from Statistics Netherlands (CBS). Market shares are computed by dividing brand sales by total sales. There is a total of 36 different brands, one being 'Other' collecting all the smallest brands. The price series concerns the price of new cars. This price series is based on the best selling model per brand in a particular year. Note that we only have the prices of models for the 26 best selling brands. The source is www.autoweek.nl. To find the price of that best selling model we

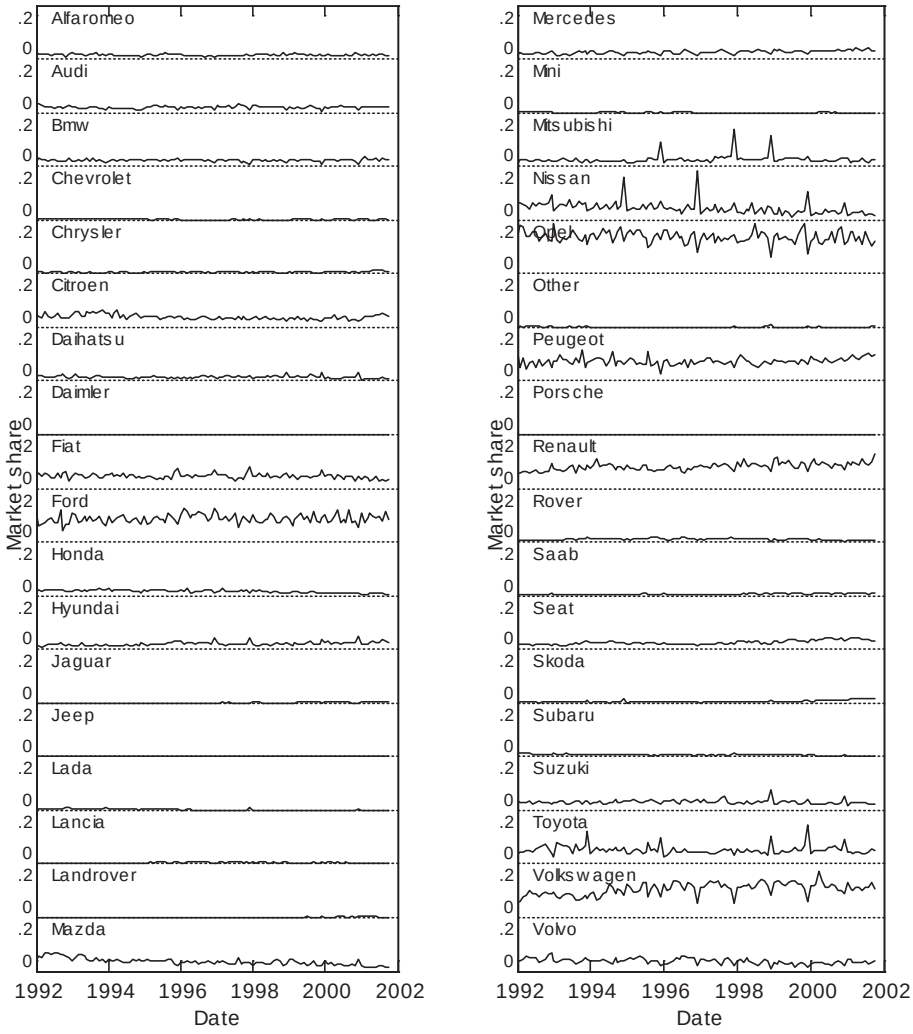

Figure 4: Market shares of 36 brands on the Dutch market between January 1992 and October 2001.

consulted various annual editions of (Dutch language) Autoboek, Autovisie, and Autotest. The market shares are presented in the line plots of Figure 4.

5.2 Estimation of the Market Share Attraction Model

We now turn to the estimation of the (unrestricted) Market Share Attraction Model, applied to our data. We expect that the prices and market share of each brand will have an effect on the market share of all the other brands. In other words, we assume that the explanatory variables in the model are the same for each brand. For convenience, we denote with $x_{k,t}$ the k^{th} explanatory variable for any brand at time t, no matter whether it is in a lagged or

other form, or it represents price or another explanatory variable. Thus, the attraction of brand i at time t, given in the general equation (1), becomes in our case

$$A_{i,t} = \exp(\mu_i + \varepsilon_{i,t}) \prod_{k=1}^{K} x_{k,t}^{\beta_{k,i}} \quad \text{for } i = 1, \ldots, I, \tag{11}$$

with $k = 1, 2, \ldots, 88$, $t = 1, 2, \ldots, T$ and $I = 36$. The length of the time horizon T ranges from 50 to 117, since we study the evolution of market shares over time. For each T a separate Market Share Attraction Model for all brands is estimated. The first 26 explanatory variables are current prices; the next 26 variables are one month lagged prices, and the last stack of 36 variables are one month lagged market shares of all brands. Using brand I as a base brand, (1) translates into the market share equations for brands $1, 2, \ldots, I-1$ at time t (akin to (5))

$$\begin{array}{rcccccl} \log M_{1,t} - \log M_{I,t} & = & \tilde{\mu}_1 & + & \sum_{k=1}^{K} \tilde{\beta}_{k,1} z_{k,t} & + & \eta_{1,t} \\ \log M_{2,t} - \log M_{I,t} & = & \tilde{\mu}_2 & + & \sum_{k=1}^{K} \tilde{\beta}_{k,2} z_{k,t} & + & \eta_{2,t} \\ \vdots & = & \vdots & + & \vdots & + & \vdots \\ \log M_{I-1,t} - \log M_{I,t} & = & \tilde{\mu}_{I-1} & + & \sum_{k=1}^{K} \tilde{\beta}_{k,I-1} z_{k,t} & + & \eta_{I-1,t}, \end{array} \tag{12}$$

where $z_{k,t} = \log x_{k,t}$. For notational convenience, we denote $y_{i,t} = \log M_{i,t} - \log M_{I,t}$, $\mathbf{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,T})'$, $\tilde{\boldsymbol{\beta}}_i = (\tilde{\beta}_{1,i}, \tilde{\beta}_{2,i}, \ldots, \tilde{\beta}_{K,i})'$ and $\boldsymbol{\eta}_i = (\eta_{i,1}, \eta_{i,2}, \ldots, \eta_{i,T})'$. Further, we denote with $\mathbf{Z}$ the common matrix of independent variables for each brand over time $t = 1, 2, \ldots, T$. Consequently, (12) can be modeled in matrix form as

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{I-1} \end{pmatrix} = \begin{pmatrix} \mathbf{Z} & 0 & \cdots & 0 \\ 0 & \mathbf{Z} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}}_1 \\ \tilde{\boldsymbol{\beta}}_2 \\ \vdots \\ \tilde{\boldsymbol{\beta}}_{I-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \vdots \\ \boldsymbol{\eta}_{I-1} \end{pmatrix}. \tag{13}$$

The coefficients of this model can now be estimated using OLS or SVR. For OLS, one estimates the model coefficients by minimizing the sum of squared errors, $\sum_{i=1}^{I-1} \sum_{t=1}^{T} \eta_{i,t}^2$. For SVR estimation, one minimizes the sum $0.5 \sum_{i=1}^{I-1} \sum_{k=1}^{K} \tilde{\beta}_{k,i} + C \sum_{i=1}^{I-1} \sum_{t=1}^{T} \max\{0, |\eta_{i,t}| - \varepsilon\}$.

Because of the structure of the block diagonal matrix with blocks $\mathbf{Z}$, the OLS estimates can be computed very efficiently. The inverse $(\mathbf{Z}'\mathbf{Z})^{-1}$ only needs to be computed once and $\tilde{\boldsymbol{\beta}}_i = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}_i$ contains the OLS optimal weights for brand i. In a similar way, the weights for the linear SVR problem can be estimated separately for each brand i. Computationally, this split will be much faster than inserting (13) directly into a linear SVR program.

However, for nonlinear SVR, the problem cannot be split up into I smaller nonlinear SVR problems because its solution is defined in the dual where the nonlinearity is added to the full problem. Hence, splitting up the nonlinear SVR into smaller parts does not solve the full nonlinear SVR problem.

Although coefficient estimates for SVR are not always available in the nonlinear case (see Section 4), predicted values for the y's can always be created. Once all y's are obtained, say using values for the predictor variables at test time t^*, market shares can be derived using the relationship

$$\frac{e^{y_{i,t}}}{\sum_{i=1}^{I} e^{y_{i,t}}} = \frac{e^{(\log M_{i,t} - \log M_{I,t})}}{\sum_{i=1}^{I} e^{(\log M_{i,t} - \log M_{I,t})}} = \frac{M_{i,t}/M_{I,t}}{\sum_{i=1}^{I} M_{i,t}/M_{I,t}} = M_{i,t},$$

which uses the fact that the market shares sum up to unity.

5.3 Results

We estimated the coefficients of the Market Share Attraction Model given in (13) using both the SVR and OLS techniques. As indicated in Section 2, OLS is equivalent to ML estimation because (a) the dependent variable is assumed conditionally Gaussian, (b) the explanatory variables are the same for all brands, and (c) there are no parameter restrictions. The dependent variable is the log-ratio of market shares of 35 car brands and an arbitrary base brand, which we have chosen to be Volvo. The predictor variables include current prices, one period lagged prices, and one period lagged market shares. For SVR, we have used the linear SVR and the nonlinear SVR with the popular Radial Basis Function (RBF) kernel. We use an expanding window of historical in-sample data to produce a forecast for a given out-of-sample month. That is, we have estimated the market share model (13) 68 times, each time using slightly different, one-month-extended data. Thus, to forecast the first out-of-sample month March 1996, we use the first 50 months from January 1992 to February 1996. For each following out-of-sample month we add one month of historical data to the estimation window. In the end, we calculate the Root Mean Squared Prediction Error (RMSPE) and Mean Absolute Prediction Error (MAPE) per brand per month, where an error is defined as true brand market share minus predicted market share. Note that the market shares are always between 0 and 1.

For each of the 68 periods, we tested whether the assumption of OLS holds that the dependent variable is sampled from a Gaussian distribution. Therefore, we carried out two

normality tests, that is, the large-sample Jarque-Bera and small-sample Lilliefors tests. Both tests rejected normality of the dependent variable at the 5% significance level for each of the 68 models for all samples. This result already suggests that SVR may perform better than OLS.

As noted in Section 4, SVR requires some parameters to be tuned, notably C and the width ϵ of the error-insensitivity region. In the case of RBF kernel, an additional γ parameter has to be tuned. The tuning is usually done via a grid parameter search, where each grid point is evaluated using a cross-validation procedure. In our case, we use a five-fold cross-validation procedure, which is carried out as follows. A given training data set is divided into five equal parts. A particular point on the grid is selected. It represents a tuple of values for the tuning parameters. Five models are trained, where each of the five parts is considered as an out-of-sample test set and the remaining four parts as a training set. The parameter combination that produces minimal squared error over the five test sets is then used to train the whole data set (consisting of all five parts) and produce an out-of-sample forecast for the following month.

The main results of the experiments are presented in Table 1. Overall, SVR outperforms OLS considerably and consistently in terms of RMSPE and MAPE over the 68-month out-of-sample period from March, 1996, to October, 2001. The average monthly RMSPE over all brands for the out-of-sample period is equal to 0.028839 for OLS. The corresponding figure for the linear SVR is 0.008466, and for SVR with RBF kernel the RMSPE is 0.008452. Figure 5 shows a detailed out-of-sample monthly RMSPE performance averaged over all brands. There are about 6 to 8 months that could visibly be considered as out-of-sample outliers, since both OLS and SVR perform relatively worse there. Clearly, OLS performs much worse, which suggests that SVR is capable of mitigating the effect of outliers and perform better in times of relative market distress.

Interestingly, both the linear SVR and the highly nonlinear SVR produce more or less the same prediction results, suggesting that there is not enough evidence in the data to favor a nonlinear relation among the dependent and independent variables. Nevertheless, linear SVR has performed substantially better than OLS, suggesting that the robustness and penalization properties of SVR have worked out well on this particular market share prediction task. As demonstrated in Pérez-Cruz et al. (2003), factors working against OLS and in favor of SVR are the dependent variable not being sampled from a Gaussian distribution,

Table 1: Performance results over 68 out-of-sample months (March 1996 – October 2001): Mean Absolute Prediction Error (MAPE) and Root Mean Squared Prediction Error (RMSPE) for OLS, linear SVR (lin SVR), and nonlinear SVR with Radial Basis Function kernel (RBF SVR) models. MAPE and RMSPE represent the average of the average monthly errors over all 36 brands during the out-of-sample period.

	OLS	lin SVR	RBF SVR	improvement of lin SVR over OLS	improvement of RBF SVR over OLS
MAPE	0.012803	0.004882	0.004879	2.622 times	2.624 times
RMSPE	0.028839	0.008466	0.008452	3.406 times	3.412 times

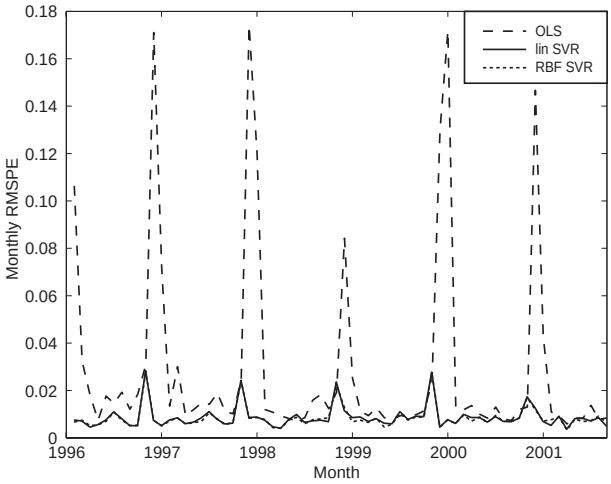

Figure 5: 68 monthly average RMSPE's over all brands for OLS, Linear SVR, and nonlinear RBF SVR.

the large amount of predictors relative to T, and the number of in-sample months.

It was noted above that RBF SVR gives similar forecasts to linear SVR. Adding nonlinearities to a model with a true linear relationship may lead to overfitting the training data and worse out-of-sample forecasts. In the case of RBF SVR, there is no danger of overfitting as the penalization and ϵ-insensitive loss function work in the direction of producing a linear relation, unless there is sufficient evidence in the data for the presence of nonlinearities, as argued also in Figure 1. What is more, Keerthi and Lin (2003) have demonstrated that there is no need to consider linear SVR since the RBF SVR is capable of discovering linear relations quite well.

Next we focus on the coefficient estimates produced by OLS and the linear SVR. Such estimates are not readily available for the nonlinear RBF SVR. As argued from the theoretical viewpoint in Sections 3 and 4, the estimated coefficients in SVR are shrunk towards zero vis-a-vis the corresponding OLS coefficients. This effect can be observed for our task as well. Figures 6 depict the effects of each of the 88 predictor variables on each of the 35 explained variables $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{I-1}$ for OLS (left column) and linear SVR (right column). The explanatory variables are divided into three groups: the first group consists of current prices, the second of the lagged prices, and the third group of lagged market shares. Each of these effects does not stand for one particular period T. Rather, it represents the average value over the 68-month out-of-sample period. The filled circles represent the average effect over the 68-month out-of-sample period of the predictor variables on $\mathbf{y}_{10}$, the log-difference of the market shares between brand 10 (Ford) and the base brand (Volvo). A key observation to make here is the striking difference between the magnitude of the OLS and linear SVR coefficients in general.

To visualize the influence of particular predictors, the sign of the effect is of less importance than the size. Therefore, we also present the absolute values of the effects of each predictor variable on $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{I-1}$ in Figure 7. For each predictor variable, the sum of absolute effects on $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{I-1}$ is depicted. Thus, the number of different shades of gray is $I - 1$ (or, 35 in our case). This representation allows us to observe whether OLS and linear SVR consider the same variables to be influential. For example, consider the 26 current-price variables. A striking feature, that is not easily observable from Figure 6, is that the variables that appear to play a key role in OLS estimation have also a relatively big impact in linear SVR estimation, most notably prices of Fiat, Ford, Mercedes, Renault,

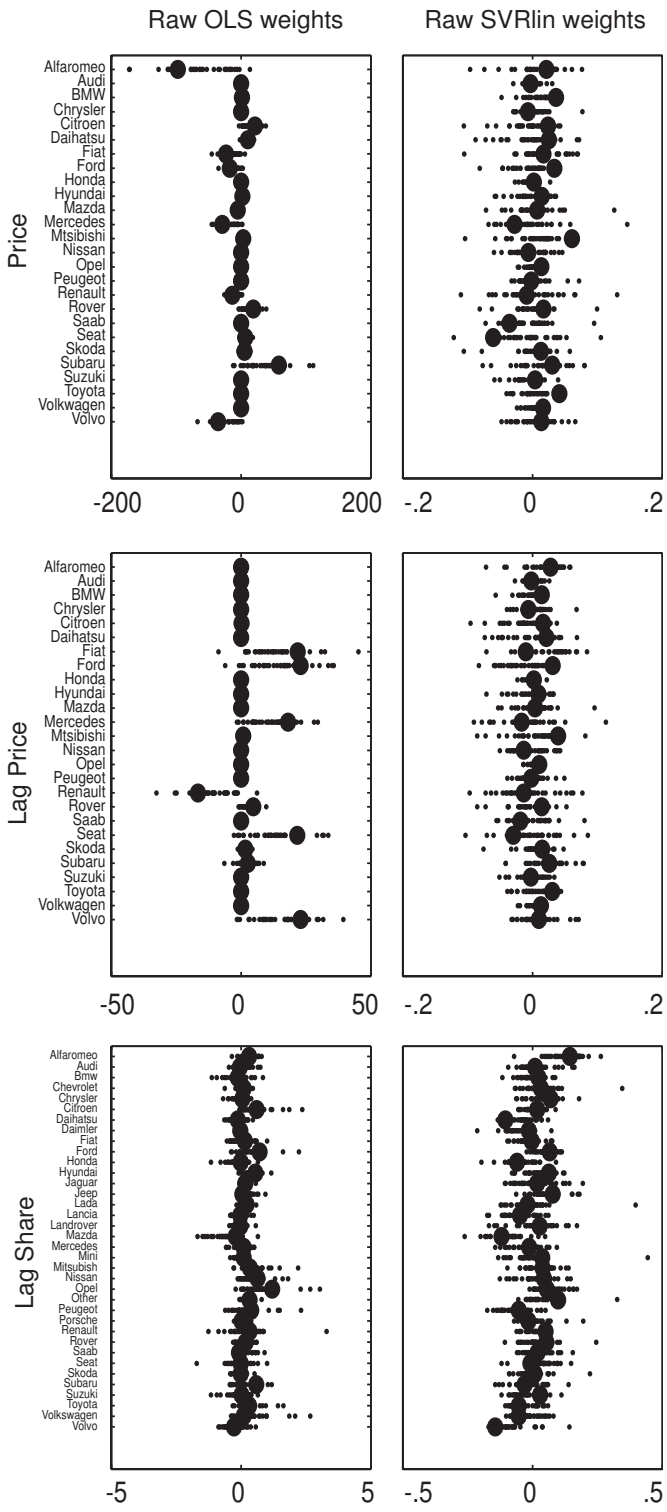

Figure 6: Regression weights (small dots) of the predictor variables obtained by OLS and linear SVR for each of the 35 explained variables on average, where the averaging is done over the 68 out-of-sample periods. The circles represent the average (over time) effect of all predictor variables on a concrete explanatory variable: the log-difference of the market shares of Ford and the base brand Volvo.

Seat, Subaru, and Volvo. In linear SVR, there are some additional variables that stand out as important: prices of Alfa Romeo, Citroen, Daihatsu, and others. Overall, the linear SVR coefficients have much lower magnitude in absolute sense, and are more evenly spread than the corresponding OLS coefficients.

To see how the three models differ in predicting the market shares, we have plotted the prediction errors for each brand. Figure 8 shows these plots for OLS and Figure 9 for linear SVR. We do not present a representation of the errors for the nonlinear RBF SVR because there is hardly any difference with the errors of linear SVR. The most striking feature again is that OLS has several large errors whereas the largest errors of linear SVR are at least a factor 5 smaller. The error plots can be interpreted for each brand separately. To consider one case of the residual plot for linear SVR, Volkswagen had around January 1999 and again around January 2000 a positive error, whereas Toyota had simultaneously two spikes of negative errors. As a consequence, the linear SVR model apparently has underestimated the market share of Volkswagen and overestimated the market share of Toyota at these time points.

6 Conclusion

The Market Share Attraction Model may suffer from estimation problems when the model assumptions are not satisfied. In this paper, we have introduced SVR as an alternative estimation procedure for this model. An intuitive and self-contained introduction to SVR was provided. To test the estimation procedures, we compared OLS to linear and nonlinear SVR to empirical market share data of the Dutch automobile sales of 36 brands. It was found that the prediction by either linear or nonlinear SVR was much better than OLS. There was hardly any difference between linear and nonlinear SVR indicating that for these data it is not necessary to allow for a nonlinear prediction.

There are some remaining issues for SVR. For example, is there an optimal kernel function for the Market Share Attraction Model or is the linear SVR sufficient? What is the best procedure for tuning the manually-adjustable parameters in the SVR? Other issues are how to compute confidence intervals for the predictions and how to interpret the parameters of the nonlinear SVR model. Also, it would be interesting to see how SVR compares to other penalization methods.

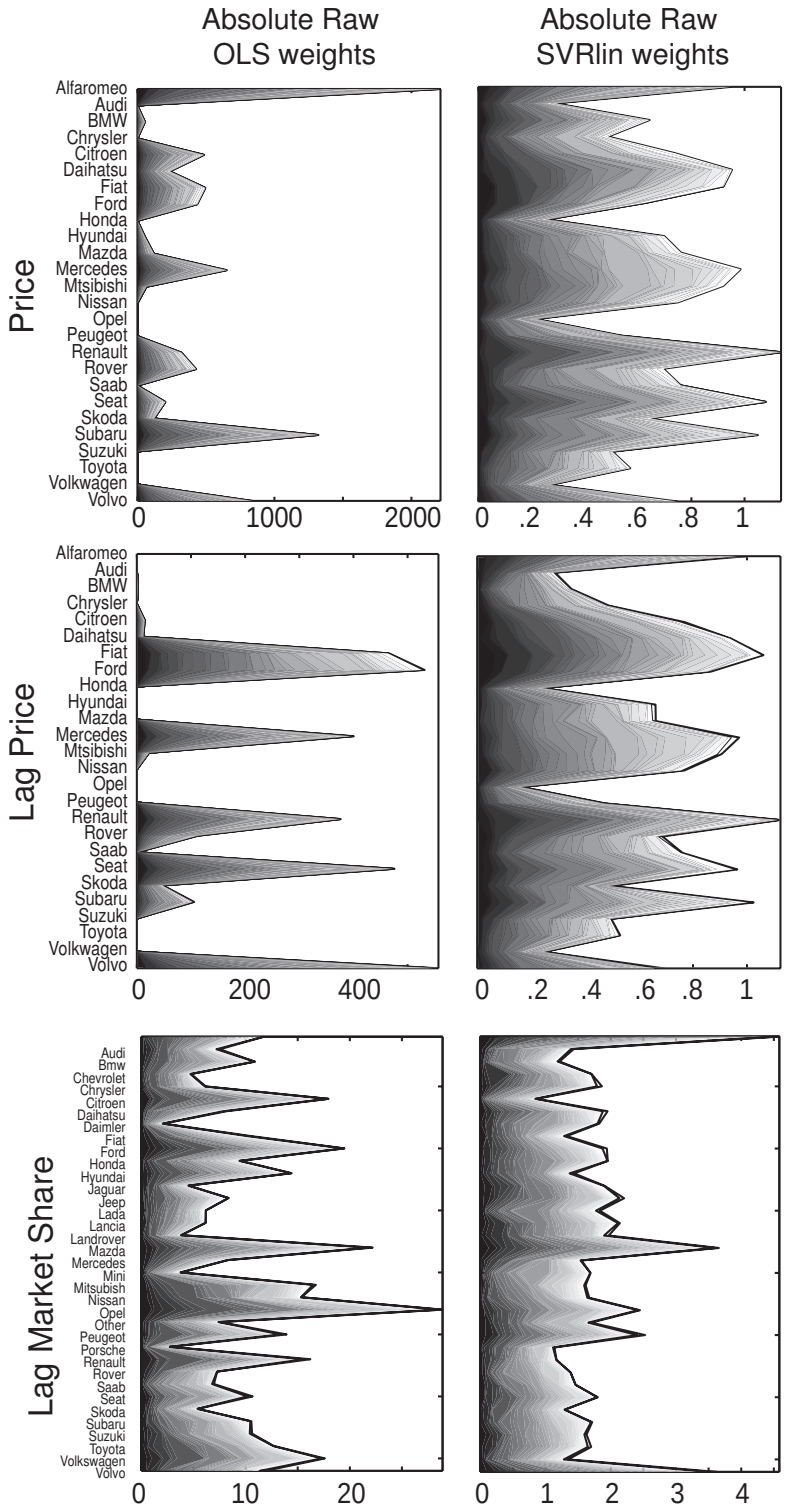

Figure 7: Distribution of absolute regression weights of the variables obtained by OLS and linear SVR for each of the 68 estimation periods. The darkness of the area indicates the density of the absolute regression weights.

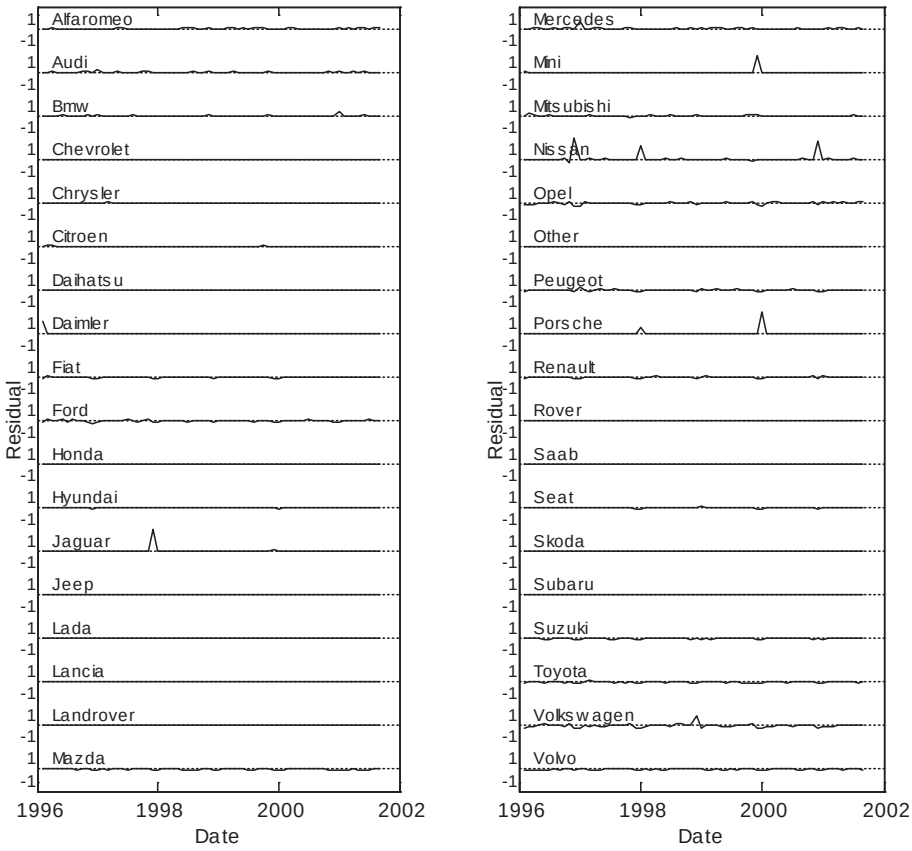

Figure 8: Out of sample prediction error of OLS estimation.

Clearly, more experiments have to be carried out to confirm or refute more convincingly the applicability of SVR in marketing tasks and as a competitor to ML estimation. Our empirical comparison suggests that when the OLS assumption of normality of the errors in the Market Share Attraction Model is not satisfied, SVR is a good alternative to estimate its parameters.

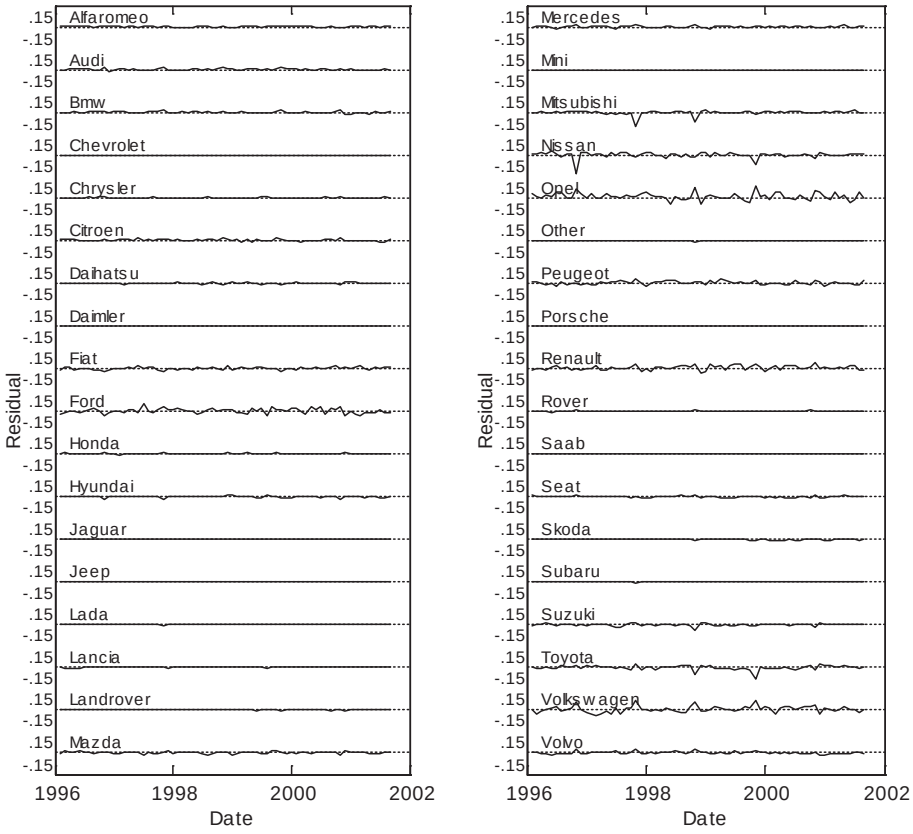

Figure 9: Out of sample prediction error of linear SVR estimation.

References

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

Cooper, L. G., & Nakanishi, M. (1988). *Market share analysis: Evaluating competitive marketing effectiveness.* Boston, MA: Kluwer Academic Publishers.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines.* Cambridge University Press.

Fok, D., Franses, P., & Paap, R. (2002). Advances in econometrics. In P. Franses & A. Montgomery (Eds.), (Vol. 16, pp. 223–256). Elsevier Science.

Fok, D., & Franses, P. H. (2004). Analyzing the effects of a brand introduction on competitive structure using a market share attraction model. *International Journal of Research in Marketing*, *21 (2)*, 159–177.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* Springer-Verlag New York, Inc.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*, 73–101.

Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic behaviorsof support vector machines with gaussian kernel. *Neural Computation*, *15*, 1667–1689.

Pérez-Cruz, F., Afonso-Rodríguez, J. A., & Giner, J. (2003). Estimating GARCH models using support vector machines. *Quantitative Finance*, *3*, 163–172.

Schölkopf, B., Burges, C., & Vapnik, V. (1995). Extracting support data for a given task. In U. M. Fayyad & R. Uthurusamy (Eds.), *First international conference on knowledge discovery and data mining.* AAAI Press, Menlo Park, CA.

Schölkopf, B., Guyon, I., & Weston, J. (2001). *Statistical learning and kernel methods in bioinformatics.* (Available at http://citeseer.ist.psu.edu/509446.html)

Smola, A. J. (1996). *Regression estimation with support vector learning machines.* Master's thesis, Technische Universität München.

Smola, A. J., & Schölkopf, B. (1998). *A tutorial on support vector regression* (NeuroCOLT2 Technical Report No. NC-TR-98-030). University of London, UK.

Tay, F., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega: The International Journal of Management Science*, *29 (4)*, 309–317.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math Dokl*, *4*, 1035–1038. (English translation of Dokl Akad Nauk SSSR 151, 1963, 501-504)

Vapnik, V. N. (1982). *Estimation of dependences based on empirical data.* Berlin: Springer.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* Springer-Verlag New York, Inc. (2nd edition, 2000)