Predicting SA Bond Prices Using PCA & SVMs

Hlalumi Adams

2024-03-07

Background

In this paper, we present a predictive method that supports better detection of interest rate movements. The proposed method uses a support vector machine (SVM) model that determines the optimal SVM parameters, by combining it with dimensional reduction techniques. The dimension reduction techniques (feature extraction approach) extract critical, relevant, and de-noised information from the input variables (features), and reduce the time complexity. We investigated two different feature extraction techniques: principal component analysis and kernel principal component analysis. The feasibility and effectiveness of this proposed ensemble model were examined using a case study, predicting the daily closing prices of the bond prices (specifically the ALBI bond) traded in the Johannesburg Stock Exchange (JSE) of South Africa. In this study, we assessed the performance of the new ensemble model with the three feature extraction techniques, using different performance metrics and statistical measures. Our experimental results show that the predictive results from the SVM model are not viable and effective, and provides poor predictions. Recent papers have proposed a novel hybrid SVM-TLBO model consisting of a support vector machine (SVM) and a teaching-learning-based optimization (TLBO) method that determines the optimal SVM parameters. The SVM-TLBO hybrid model have been proven to be viable and effective, and provides better prediction. This paper will not be exploring this hybrid model but the exploration of such a model would have improved the results.

Introduction

In the South African financial industry it has become very important to be able to model and predict changes in interest rates as this impacts inflation levels and individuals' spending abilities. One of the main drivers of interest rates are bond prices as exists an inverse relationship between bond prices and interest rates. In this paper we analyze how PCA, kPCA and SVM methods succeed in describing the dynamics of South African bond market prices at different points in time ,especially in the proximity of major market events.

Recently, support vector machine (SVM) has become a popular tool in time series predicting. In developing a successful SVM predictor, the first step is feature extraction. This paper proposes the applications of principal component analysis (PCA) and kernel principal component analysis (KPCA) to SVM for feature extraction. PCA linearly transforms the original inputs into new uncorrelated features. KPCA is a nonlinear PCA developed by using the kernel method. Furthermore, among the two methods, there is the best performance in PCA feature extraction.

This study aims to identify the main underlying factors that are responsible for driving South African bond prices and hence interest movements using PCA and KPCA . We will then proceed to predicting bond prices using SVMs for regression problems.

Moreover this analysis will be specifically made with the focus being on two periods where the South African financial market experienced volatility and extreme market stress. The first period will be 2008-2009 which is known to be the Global Financial Crisis(GFC) period which took a toll on individuals and institutions around the world, South Africa included. The second period will be 2015 which is called Nenegate in the financial industry causing havoc in the financial industry. Therefore this study further aims to identify the main drivers of bond price factors during these two periods and predict future bond prices during extreme market stress.

Data and Preliminary Data Analysis

The data series studied in this paper are daily-frequency observations from bond market data; the data is daily sampled closing price data and runs from 1994 to 2017. This data was purchased from INET-BFA for academic use . The data consists of bonds with different maturities and types namely ALBI, 1 to 3 years bonds, 3 to 7 years, 7 to 12 years, over 12 years and GOVI. Data for the Global Financial crisis (2008-2009) and Nene-gate (2015) period will be considered .

Applying PCA to SA Monthly Bond Data

Background

PCA is mainly used for dimension reduction where given a high-dimensional data i.e. information on multiple variables for a set of n observations, we wish to find a low-dimensional representation of the data that will retain most of the information in the original data. This "information" is captured by the variance-covariance matrix. Essentially PCA is a method for explaining the variance-covariance structure of a set of variables through a few linear combination of these variables . The aims behind PCA are variance decomposition and data reduction.

Formulation

Given the random r-vector

$$X = (X_1, ..., X_r)^T$$

with mean $\mu_{\mathbf{x}}$ and covariance matrix $\mathbf{\Sigma_{XX}}$ PCA seeks to replace the set of r (un-ordered and correlated) input variables, $X_1, X_2, ..., X_r$, with a potentially smaller set of t (ordered and uncorrelated) linear projections of the input variables,

$$\xi_j = \mathbf{b^T X = b_{j1} X_1} + ... + \mathbf{b_{jr} X_r, j = 1, 2, ...t}$$

where we minimize the loss of information due to replacement. "Information" = "total variation" = $\sum_{j=1}^{r} \mathrm{var}(X_j) = tr(\mathbf{\Sigma_{XX}})$ \The spectral decomposition theorem:

$$\mathbf{\Sigma_{XX} = U \Lambda U^T}$$

with

$$U^t U = I_r, -> \mathrm{tr}(\Sigma_{XX}) = tr(\Lambda) = \sum_{j=1}^{r} \lambda_j$$

, where the columns of U are the eigenvectors of $\mathbf{\Sigma_{XX}}$ and the λ_j are the corresponding eigenvalues. The first t linear projections $\xi_j, j = 1, 2, ..., t$ are ranked through their variances such that

$$\mathrm{var}(\xi_1) \geq \mathrm{var}(\xi_2) \geq ... \geq \mathrm{var}(\xi_t)$$

and the ξ_j are uncorrelated with all

$$\xi_k, k < j$$

.

Using the SA Monthly Bond Data we will be performing PCA as it is a tractable and easy-to-implement method for extracting bond price factors from observed data

Prior to applying PCA to these returns series, it is important to determine whether PCA is in fact a meaningful procedure given the properties of the data. Since PCA seeks to replace the set of unordered and correlated input variables with a smaller set of ordered and uncorrelated projections of the input variables we plot the correlation matrix of the bond types below to assess the correlation between the different variables:

PCA will help identify hidden patterns in the data sets by reducing the dimensionality of the data . This is useful when the variables within the dataset are highly correlated , indicating redundancy in the data. Furthermore when interpreting the correlation matrix below we note the multi-collinearity. There is high correlation between the predictor variables i.e. the correlation between Total Return Index and Convexity is 0.97 while Interest Yield and Convexity has a correlation of -0.93 etc.

The table below shows that the first principal component has high values for Convexity, Interest Yield and Total Return Index which indicates that this principal component describes the most variation in these variables. Similarly the second principal component (PC2) has a high value for Annualised Volatility Close, which indicates that this principal component places most of its emphasis on annualised volatility.

The table below shows the total variance in the ALBI data set explained by each component. From the results we observe the following:

- PC1 explains 63.65% of the total variance in the ALBI data set
- PC2 explains 20.29% of the total variance in the ALBI data set
- PC3 explains 12.68% of the total variance in the ALBI data set
- PC4 explains 2.9% of the total variance in the ALBI data set
- PC5 explains 0.35% of the total variance in the ALBI data set

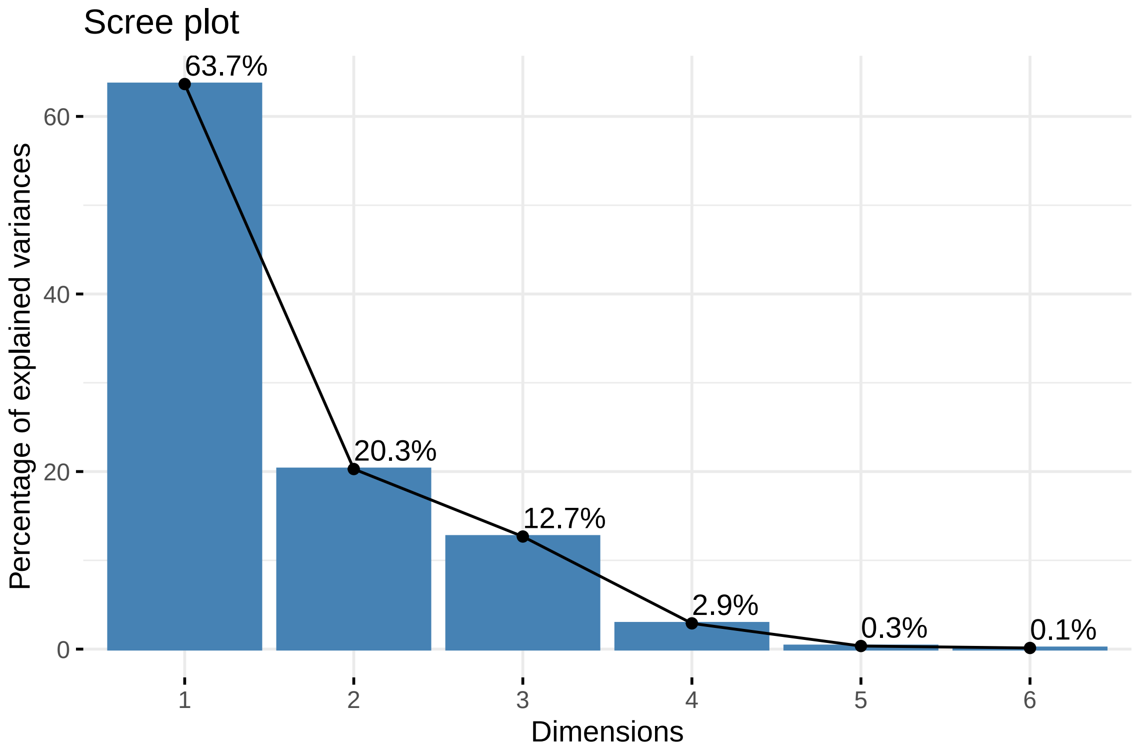
Thus, the first two principal components explain a majority of the total variance (83.94%) in the data.

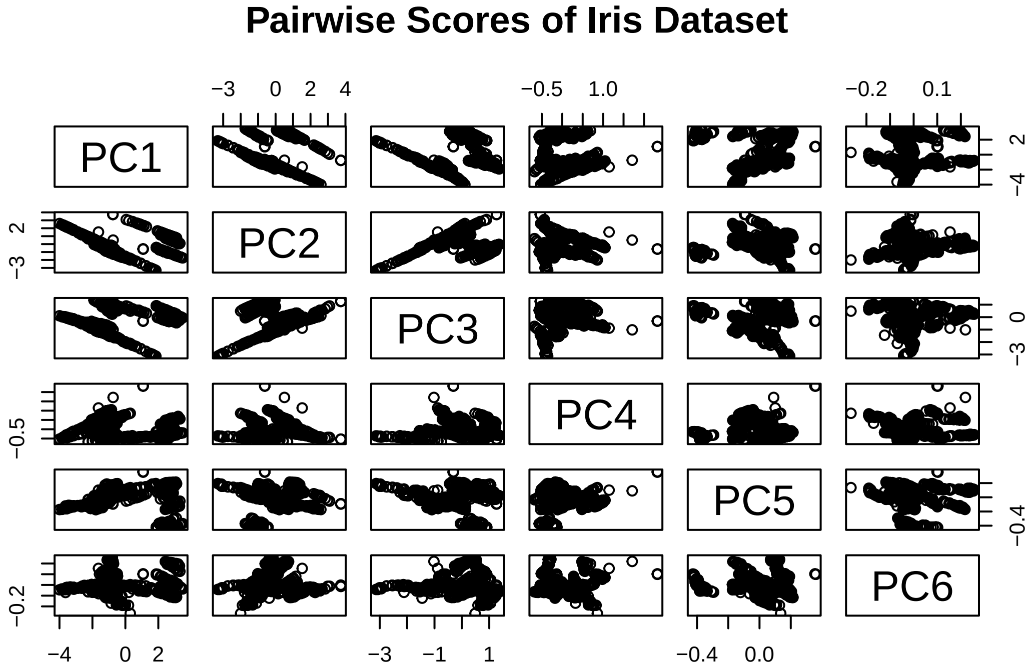
Variable Name	PC1	PC2	PC3	PC4	PC5	PC6
Annualised.Volatility	-0.305	-0.649	0.300	-0.602	-0.125	-0.132
TRI.Average.Yield	-0.332	0.648	-0.0366	-0.623	0.0626	0.278
Interest.Yield	-0.481	0.177	-0.302	0.095	-0.093	-0.792
Convexity	0.488	0.127	0.241	-0.289	0.619	-0.468
Total.Return.YtD	0.303	-0.239	-0.856	-0.337	0.021	0.046
Total.Return.Index	0.484	0.232	0.156	-0.207	-0.766	-0.236

Importance of components table:

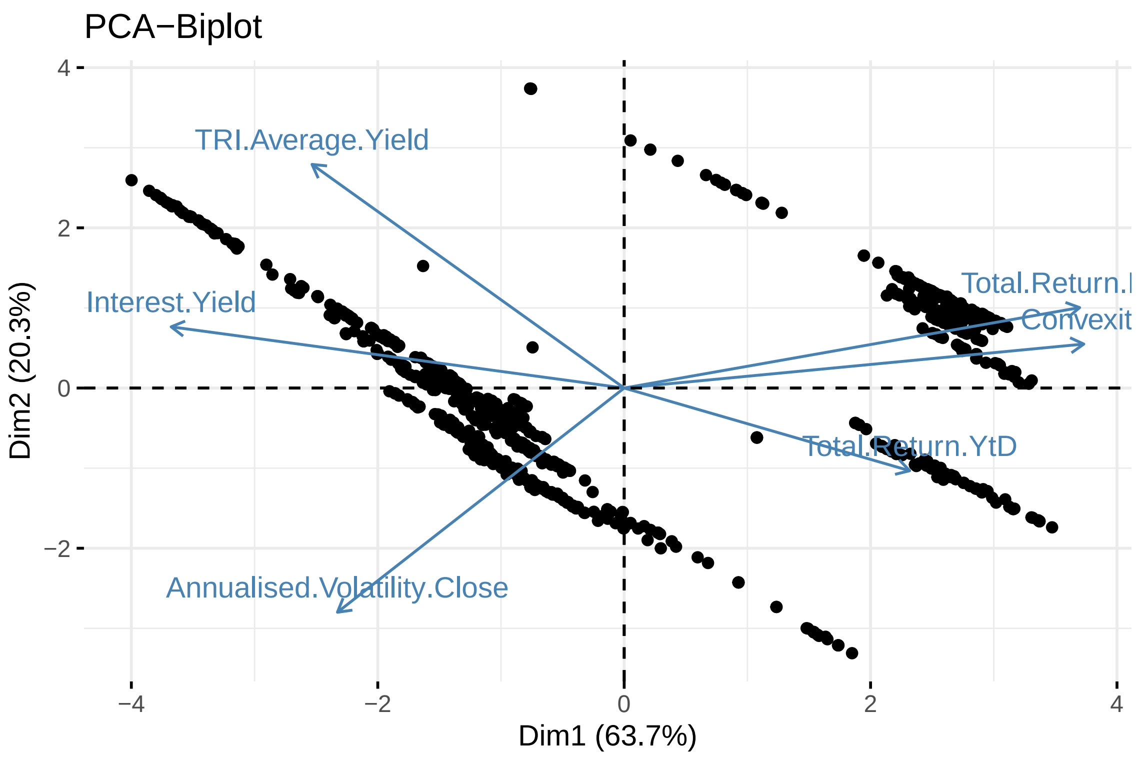
Statistic	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.954	1.103	0.872	0.417	0.145	0.088
Proportion of Variance	0.637	0.203	0.127	0.029	0.004	0.001
Cumulative Proportion	0.637	0.839	0.966	0.995	0.999	1.000

Additionally by observing the scree plot below we note that the first two principal components are the ideal number of components to retain as they explain 83.94% of the variation in the ALBI data set.





The biplot below projects each of the observations in the data set onto a scatterplot that uses the first and second principal components as the axes. From the plot we can see each of the 6 variables represented in a simple two-dimensional space.



Although PCA is good for dimension reduction SVM offers added advantages in handling high-dimensional data, flexibility with kernels, probabilistic interpretation and finding optimal hyperplanes. We look at the SVM for regression problems below.

Applying KPCA to SA Monthly Bond Data

Kernel PCA we perform a nonlinear mapping that transforms the data onto a higher-dimensional space and use standard PCA in this higher-dimensional space to project the data back onto a lower-dimensional space. Using the kernel trick, we can compute the similarity between two high-dimension feature vectors in the original feature space. We will assess if the bond prices cannot be best represented in a nonlinear space.

Formulation

To extend PCA and enable PCA to classify nonlinear data, we non-linearly transform our input space $\mathbf{x_i} \in \mathbf{R_p}$ into a point $\phi(\mathbf{x_i})$ in the M-dimensional feature space H where usually $M >> p$

The steps of KPCA are as follows:

1. Construct the kernel matrix from the training data X_i \ 2. Compute the Gram matrix $\mathbf{K} = \mathbf{HKH}$ where

$$\mathbf{H} = \mathbf{1_n} - \mathbf{n^{-1}} \mathbf{J_n}$$

with

$$\mathbf{J_n} = \mathbf{1_n} \mathbf{1_n^T}$$

\ 3. Use the fact that $\mathbf{K}\alpha = \mathbf{n}\lambda\alpha$ to solve for the α \ 4. Compute the kernel principal components $y(x)$ using

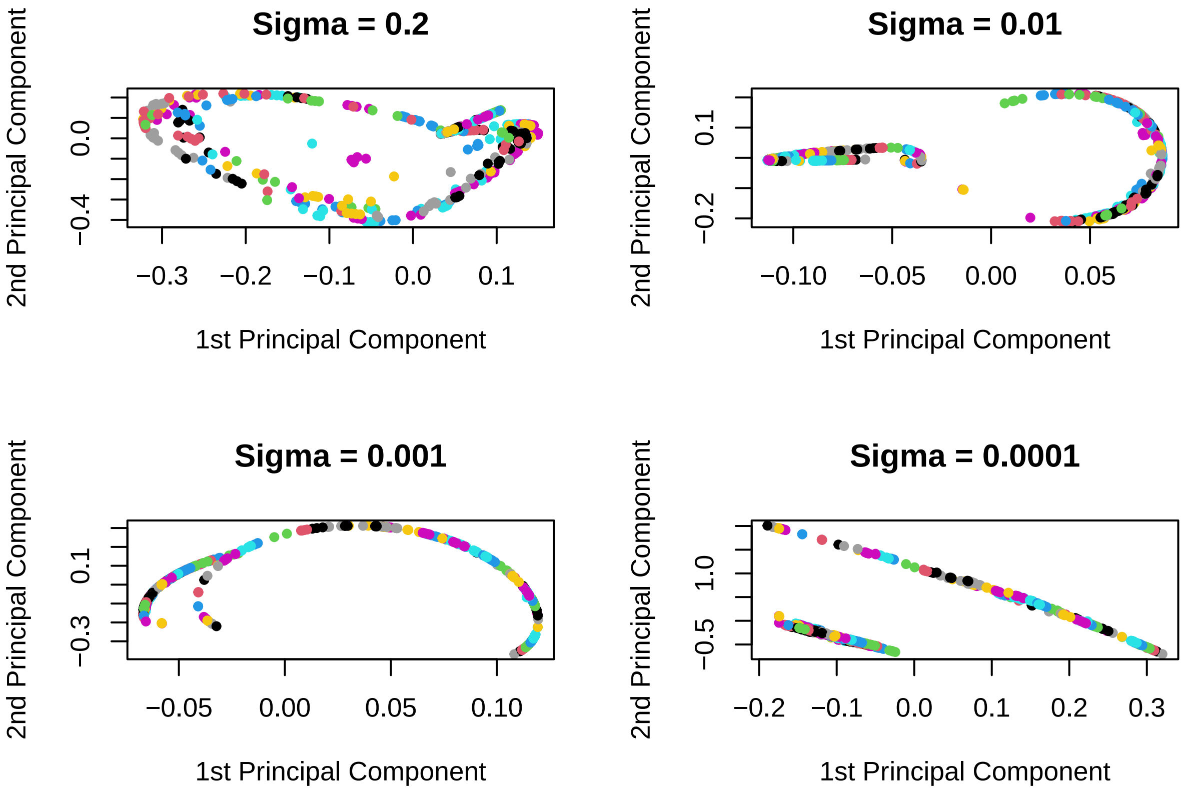
$$\mathbf{y(x)} = \phi(\mathbf{x})^{\mathbf{T}} \mathbf{v} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \alpha_{\mathbf{i}} \kappa(\mathbf{x}, \mathbf{x_i})$$

In this case for the kernel function we will be using the Gaussian (radial-basis function network) Kernel

$$\kappa(x, x_i^{'}) = \exp(-\frac{||x_i - x_i^{'}||}{2\sigma^2})$$

. The reason for this is that the RBF kernel performs well on many types of data.+

As evident in the plot below which plots the closing prices for different values of sigma, closing prices are not best represented non-linearly . There is no clear pattern on how the prices can be dimensionally reduced or non-linearly separated.



\

Applying SVM to SA Monthly Bond Data

Background

Support vector machines (SVMs) are a group of supervised machine-learning models that can be used for classification and regression. It will seek an optimal hyperplane for separating two classes in a multidimensional space. SVMs are commonly used for classification problems. Support Vector Machines (SVM) for regression, known as Support Vector Regression (SVR), is a machine learning algorithm used for predicting continuous outcomes. Unlike traditional SVM, which is used for classification tasks, SVR is designed to find a function that approximates the relationship between input variables and a continuous target variable while minimizing prediction error.

SVR works by finding a hyperplane that best fits the data points in a continuous space. The optimization problem entails finding the maximum margin separating the hyperplane, while correctly classifying as many training points as possible.This is achieved by mapping the input variables to a high-dimensional feature space and finding the hyperplane that maximizes the margin (distance) between the hyperplane and the closest data points, while also minimizing the prediction error.

Equally the regression problem (SVR) is a generalization of the classification problem we seek to find a good fitting hyperplane in a kernel-induced feature space that will have good generalization performance using the original features. SVR can handle non-linear relationships between the input variables and the target variable by using a kernel function to map the data to a higher-dimensional space. This makes it a powerful tool for regression tasks where there may be complex relationships between the input variables and the target variable. The kernel function transforms the data into a higher-dimensional feature space, making it possible to perform linear separation in this transformed space.

In this case we will be using the radial basis (RBF) kernel function

$$\exp(-\gamma * |u - v|^2)$$

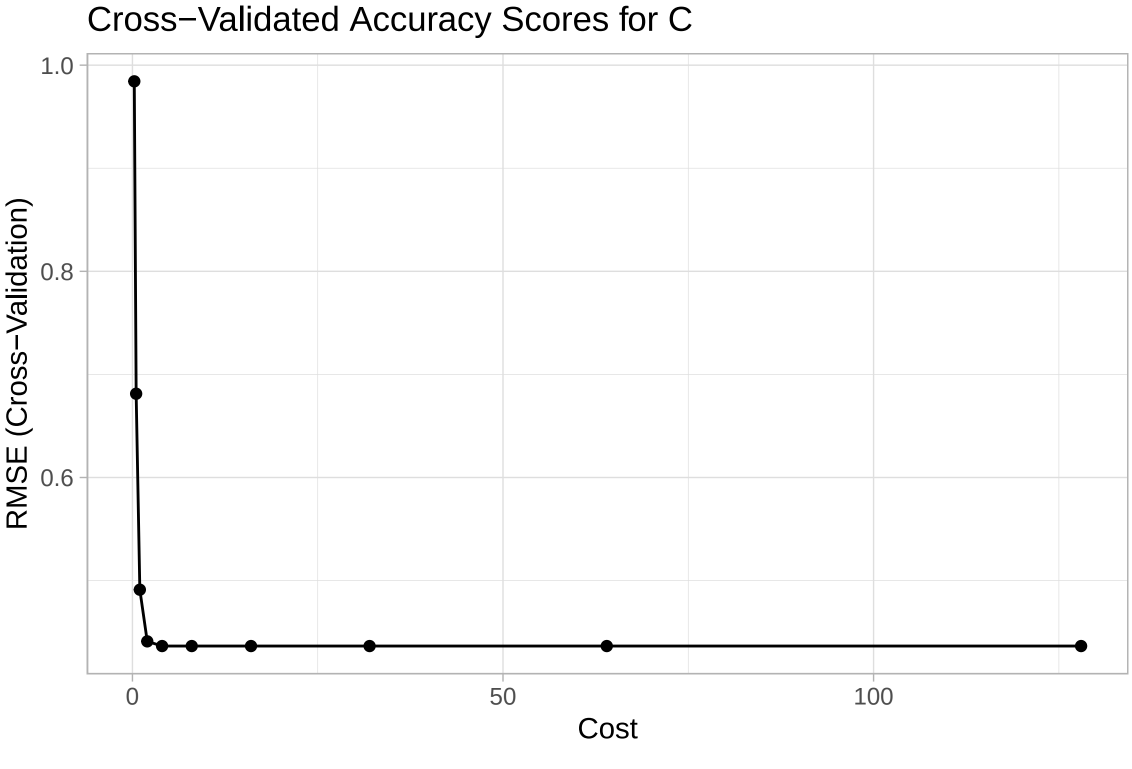
.

We start by splitting the data into training (80%) and test (20%) set .

Hyperparameters Tuning

The radial basis kernel function has two hyperparameters: σ and C.

To find the optimal cost function we tune and train an SVM using the radial basis kernel function with autotuning for the σ parameter and 10-fold CV. Plotting the results , we see that smaller values of the cost parameter ($C = 4$) provide better cross-validated scored for the training data.

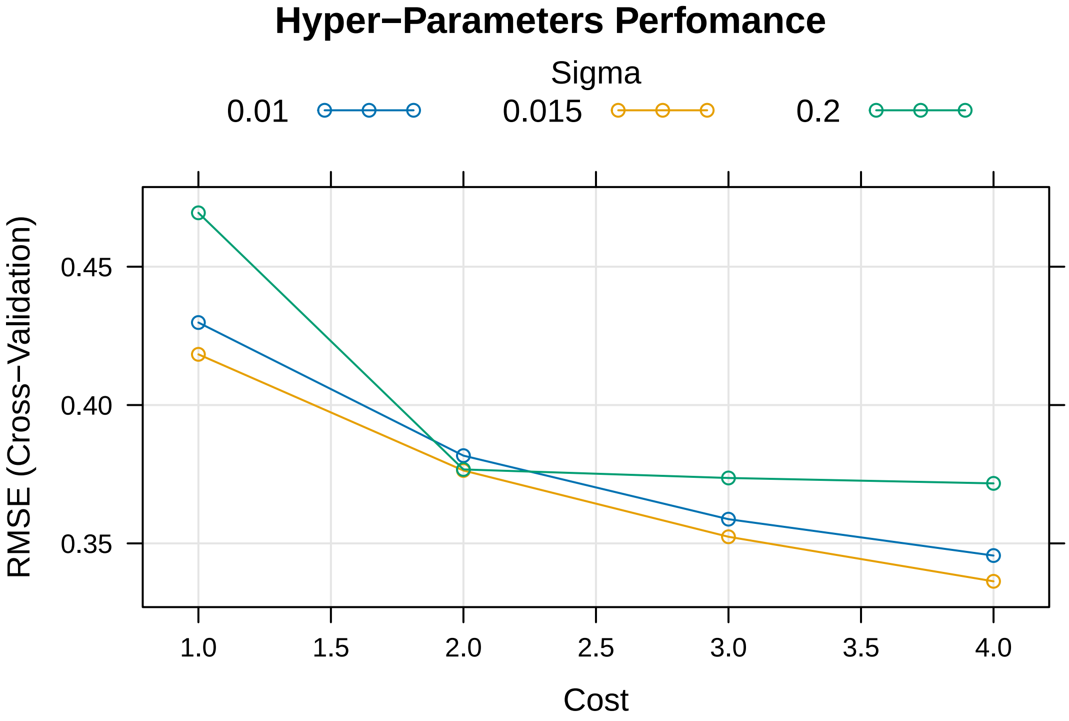


sigma	C	RMSE	R-Squared	MAE	RMSESD	MAESD
1.035429	0.25	0.9786468	0.9697924	0.4750558	0.4716834	0.041128893
1.035429	0.50	0.7045730	0.9855278	0.3980675	0.3176670	0.018766332
1.035429	1.00	0.5062893	0.9939991	0.3375884	0.1417483	0.004116280
1.035429	2.00	0.4546562	0.9949938	0.3251104	0.1307317	0.003404929
1.035429	4.00	0.4524946	0.9946100	0.3317725	0.1221181	0.002960627
1.035429	8.00	0.4524946	0.9946100	0.3317725	0.1221181	0.002960627

sigma	C	RMSE	R-Squared	MAE	RMSESD	MAESD
1.035429	16.00	0.4524946	0.9946100	0.3317725	0.1221181	0.002960627
1.035429	32.00	0.4524946	0.9946100	0.3317725	0.1221181	0.002960627
1.035429	64.00	0.4524946	0.9946100	0.3317725	0.1221181	0.002960627
1.035429	128.00	0.4524946	0.9946100	0.3317725	0.1221181	0.002960627

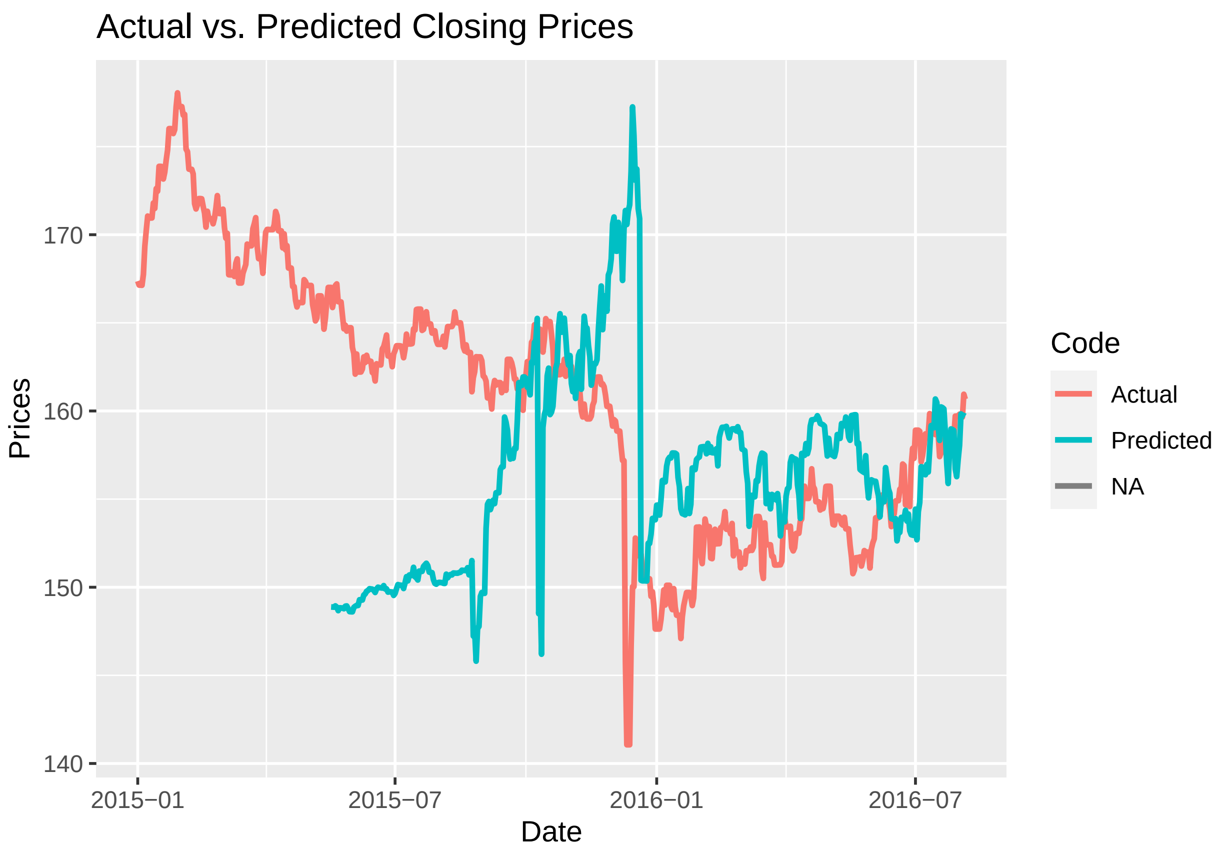
For σ we will provide a range of values for σ which return good results when using the radial basis SVM.

The final values used for the model were $\sigma = 0.015$ and $C = 4$ From the graph below we see that the optimal parameter for σ is 0.015 as it is the line that minimises the root mean square error (RMSE).



Predictions

Using the SVM for regression , we plot the predicted prices and contrast them with the actual prices . As evident in the graph below SVM for regression does a fairly job at accurately predicting future prices . We do however note that there seems to be a in December 2015 the predictive model did not perform well (i.e. was not able to predict the period of extreme volatility where there was a sudden decrease in bond prices during Nene-



gate).

Predictive Perfomance

In further assessing the predictive performance of the model , optimal svm model produced the following results:

sigma	C	RMSE	R-Squared	MAE	RMSESD	RsquaredSD
4	0.015	2.405477	0.8369	1.7848	0.2425	0.0307

1. An RMSESD of 0.2424882 suggests that the model's predictions are quite accurate, with an average error rate of about 24.24882% of the dependent variable's range (closing prices). The low RMSE indicates that the model's predictions are precise in predicting the closing prices of bonds. \
2. An R-squared value of 0.8369 indicates that 83.69% of the variability in the dependent variable (closing prices) can be explained by the independent variables in the model (convexity, total return index etc.). This suggests that the model has a strong fit to the data, meaning it can account for a significant portion of the observed variation in the dependent variable