Capacity and Complexity Control in Predicting the Spread Between Borrowing and Lending Interest Rates

Corinna Cortes Harris Drucker Dennis Hoover Vladimir Vapnik
AT&T, Crawford Corners Road, Holmdel, NJ 07733
email contact: drucker@monmouth.edu

Abstract

This is a problem in financial prediction with a small number of data points and high dimensionality in which classical economic forecasting techniques do not work. Many commercial institutions including banks, department stores, and credit companies charge their customers one interest rate (termed the lending rate) while they can borrow at a lower rate, termed the borrowing rate. The spread (difference) between these two rates can be a major profit center. If a commercial institution forecasts that the spread (and hence profit) will decrease, they can "hedge" by buying insurance against that decrease thereby locking in a profit. We used a variety of techniques that trade off training error against model complexity using the concept of capacity control for dimensionality reduction. We minimized the mean squared error of prediction and confirmed statistical validity using bootstrap techniques to predict that the spread will increase and hence one should not hedge. We briefly discuss the classical economic forecasting techniques which are not correct because the data is not independent, stationary, or normally distributed. Our predictions of the spread are consistent with the actual spread subsequent to the original analysis.

Introduction

The opportunity for hedging interest rates is in the billions of dollars for many of the larger commercial institutions. As long as the spread is going to either decrease slowly or increase, the commercial institution will not hedge. We used a time series of explanatory variables both in quarterly and monthly formats for assessing the appropriateness of hedging. An interesting aspect of this problem is that the spread is predicted as a two step procedure: first we predict the values of certain economic explanatory variables (others are taken from public or private figures generated by econometric firms). Then we predict the spread based on the prediction on the explanatory variables.

Although there is a long time history of the spread between the borrowing and lending interest rates, only the past approximately ten years can serve as valid data. We believe that certain "structural" changes in the market (for example, financial deregulation and financial innovations) makes the far past intrinsically unreliable in predicting the future. For the quarterly data, we use approximately 45 data points and twenty-one possible explanatory variables. For the monthly data, we have approximately two hundred data points and thirty possible explanatory variables. However, because of the high correlation between adjacent time samples, there may not be that much more "information" in the monthly data.

In summary, we have to predict the spread based on a potentially large number of explanatory variables where the low ratio of number of samples to number of explanatory variables could lead to overfitting of the samples. Furthermore, the success in predicting the spread depends on being able to predict the values of the explanatory variables.

Application of Learning Theory

Standard econometric textbooks, e.g., Johnston (1984) concentrate on "explaining" some set of data points using a linear regression model. There is nothing necessary incorrect about using a linear model (we will do so ourselves) - the danger is in the over-emphasis on explaining the past at the expense of predicting the future. **It is not the past we must explain, but the future that we must predict.**

Here is one of the major points of learning theory. We want to be able to predict the future (unknown) and therefore we should use as much of the past (knowns) as possible. However, if we use all the data to construct the model, we cannot assess how **confidently** we can predict the future. It is possible we can perfectly model the past and yet have an unreliable estimate of the future. To assess how well we can predict the future, we can divide the known past into a training set and validation set. We use the training set to generate a model and use the validation independent variables to predict the future values of the dependent variable. We then use some function of the difference between the predicted and true values of the validation set (typically mean squared error) as a measure of the prediction error.

Standard ordinary least squares regression (Johnston, 1984) minimizes the Residual Sum of Squares (RSS), Consider the following two equations where (x_j, y_j) and

and (x^*_j,y^*_j) are training and validation points respectively:

$$RSS = \sum_{j=1}^{n} (\, y_j - g(x_j))^2$$

$$MSEP = \frac{1}{N}\sum_{j=1}^{N} (\, y^*_j - g(x^*_j))^2$$

We find $g(x)$ to minimize RSS where $g(x_j)$ is the machines' output on training point x_j, (x_j may be multidimensional) and y_j is the observed value of y. Similarly $g(x^*_j)$ is the machine output on validation point x^*_j. The second equation is the *validation* error or the mean square error of prediction (MSEP). Superficially the summations in the two equations look similar but we emphasize that the (x^*_j,y^*_j) in the second equation are validation points and not used in constructing g(x).

Let us discuss the behavior of the training and validation error as we vary the complexity of our machine (Figure 1). If we choose too simple a machine, the machine does not have enough free parameters to model the irregularities of the training set, so both the training and validation errors are large (underfitting). As we increase the complexity of the model the learning machine can begin to fit the general trends in the data which carries over to the validation set, so both error measures decline. As we continue to increase the complexity of the model the error on the training set continues to decline, and eventually reaches zero as we get enough free parameters to completely model the training set. The behavior of the error on the validation set is different. Initially it decreases, but at some complexity it starts to rise. The rise occurs because the now ample resources of the training machine are applied to learning vagaries of the training set, which are not reproduced in the validation set (termed overfitting). The process of locating the optimal model complexity for a given training set is called Structural Risk Minimization (Vapnik, 1982; and Guyon, 1992) and otherwise termed capacity control. If the "capacity" of the machine is too large, the training error is small but the prediction error is large while if the capacity is too small, the prediction error and the training error will be large.

We can increase the complexity of our machine in several ways. For many machine learning tasks the number of input variables is fixed in which case we can increase the complexity by moving from a linear model to higher order polynomial models. Alternately, as in this study, we can fix the order of the polynomial and increase the number of explanatory variables. In general, as one increases the complexity of the model, more data points are needed to fit the model.

To locate this optimal number of input variables for our linear model, we divide the data into two parts. We use a training set to determine the coefficients of the model and then use a validation set to determine our prediction error. The problem is that more training data implies that we can better fit our model but then our estimate of the prediction error is poor (large variance) because the number of validation points is small. On the other hand we can improve our estimate of the prediction error if we increase the number of validation points, but we expect the prediction error to be larger because we have less training points to fit the model.

The Moving Control Indicator

Leave-one-out cross-validation is a validation procedure (Miller, 1990; Efron and Tibshirani,1993; Efron and Gong, 1983) that can be used if one has extremely small amounts of data. In the leave-one-out procedure we start with all the N data points. We remove one point and do a fi on the N-1 data points. The prediction error for that left out point is just the difference between the observed value and the predicted value. We then iterate over the total N points, leaving each one out in turn, obtaining N residuals. The mean square sum of these residuals is our mean square error of prediction. We used leave-one-out procedures on both the quarterly and monthly data, but for the monthly data we kept a separate test set. Leave-one-out makes efficient use of the data but can be very computationally expensive since we must make N fits to the data. However, for the case of linear regression, it turns out that we can do one linear regression on all the N points, and then adjust the residuals to give exactly the same result as if we had used a leave-one-out procedure. Vapnik (1982, pages 238) terms this a "moving control" indicator (see also Efron and Tibshirani, 1993, page 255).

Commonly Used Indirect Methods

We have discussed direct methods to estimate the prediction error, i.e., we train on part of the data and test or validate on another set of the data. Indirect methods are characterized by a lack of validation or testing set and the optimum choice of parameters and estimate of error are obtained by examining the residual sum of squares to find the optimum value of p. One method is to calculate $RSS/(N-p)$ where N is the number of samples and p is the number of variables used to fit all N training points. The training error (RSS) tends to decrease as p increases while the denominator in this case also decreases for increasing p. However, they decrease at different rates. At some value of p, there is a minimum of the ratio. The problem with this approach is that $RSS/(N-p)$ is actually a measure of the noise variance, not the prediction error (Miller, 1990) and therefore should not be used. Another technique proceeds as follows: We find the RSS_p corresponding to a set of p parameters. We then add another parameter and ask if the RSS_{p+1} is significantly different (in a statistical sense). If not, stop. The test statistic is the ratio of a function of the two RSS's. Standard assumptions about the residuals are normality, zero mean, independence, and stationarity-conditions that are not necessarily met in practice. Furthermore, the test statistic is only F distributed (Miller, 1990) if the additional parameter is chosen at random and **not** if the parameter is picked to minimize RSS_{p+1}. Therefore, using the F distribution to examine significance in the latter case is not correct.

Exhaustive and Sequential Search

Our objective is to find the best set of p explanatory variables that predicts the spread. One could add a variable at a time to the linear model until the mean square error of prediction goes through the minimum of Figure 1. The main problem with this approach is that the best set of p variables is not necessarily a subset of the best set of (p + 1) variables. To find the minimum validation error, one must exhaustively search C_k^p combinations for each p (with total number of variables k), and with a total of $2^k - 1$ combinations over all p. The cost of hedging is substantial (millions of dollars) and therefore even if the search takes days (as it does for 21 variables), the benefits are worth the use of CPU time. For the quarterly data, we did an exhaustive search but for the monthly data we did a sequential search.

We used quarterly data from 1983 to the end of 1993 with twenty-one possible explanatory variables and linear regression. Exhaustive search as described previously was used to find the p variables that minimize the *MSEP*. However, there is one caveat. Just finding a minimal *MSEP* does not guarantee that this choice of p variables is better than some naive prediction procedure. For a naive prediction procedure, we will make the average spread (in the past) as the best prediction of the future. The variance of the spread, $E(y - \bar{y})^2$, is then just the mean square error using the average ($\bar{y}$) as a best prediction of the future. We can't find the expectation so we will use as an approximation the summation below and define a normalized mean square error of prediction (NMSEP) as

$$NMSEP = \frac{MSEP}{\frac{1}{N}\sum_{j=1}^{N} (y_j - \bar{y})^2}$$

If the NMSEP is greater than 1, our prediction strategy is worse than a naive strategy. From now on we will plot NMSEP versus p. The use of a normalized mean square error is common in the time series literature. In addition to plotting NMSEP versus p for the 21 variables, we also have two other additional lists of explanatory variables (all subsets of the 21 variables). that we feel most comfortable about predicting. Figure 2 shows a plot of NMSEP versus p for these three cases. The top curve is far inferior and thus we will use the other two curves. We show all the values for the eleven variable case but not all the results for the twenty-one variable case since we have already reached a minimum at p = 11. As can be seen in the twenty-one variable case, the curve has a broad minimum which starts to flatten out at about five or six variables and reaches a true minimum at p = 11. With a broad minimum, is preferable to use as few variables as possible since this decreases our confidence interval (Vapnik, 1982) and so for the two lower curves we use six variables.

The monthly data differs from the quarterly data in that the time series go further back in time (March 1977) for a total of 200 data points, and they came with 30 explanatory variables. Having the richness of 200 data points we divided the data into a training set of 150 points and a test set of 50 points, the test set being the most recent values. The test set was not used for estimating the free parameters of the model or selection of model complexity. These issues were determined from the training set alone. The test set was only used once to determine a final mean square error of prediction. An estimate of the prediction error used for model selection was determined from a validation set using the leave-one-out method discussed above. With 150 data points and 30 explanatory variables we determined that exhaustive search for the best variables would be computationally expensive so we used sequential search techniques.

Our first sequential selection scheme formalizes and systematizes in a simple way the intuitive notion of desiring explanatory variables with high correlation to the spread, but small correlations to each other. The first variable is chosen as the variable with highest correlation to the spread. A linear model with only this variable is formed on the basis of the full training set, and residuals are obtained. These residuals are to be explained by the remaining variables. Another variable is chosen as the one with the highest correlation to the residuals. A model with two variables is formed and the residuals recalculated. The process is continued until all variables (or a sufficiently large number of variables) have been ranked. The training and validation error as a function of number of explanatory variables ranked according to this scheme is shown in Figure 3.

Principal Components

The above sequential ranking scheme is slightly awkward because, for each extra chosen variable, we have to redo the modeling and estimate new residuals. This is a necessity because the explanatory variables are not independent. We can attempt to remove some of this dependency by decorrelating the explanatory variables first. When the explanatory variables are correlated the cross-correlation matrix $X'X$ will have non-zero elements off the diagonal. If we can diagonalize this cross-correlation matrix our new variables will be uncorrelated. Fortunately, we are guaranteed that the cross-correlation matrix can be diagonalized because it is symmetric. The decorrelation can therefore be obtained by finding the eigenvectors of the cross-correlation matrix and mapping our data to the space spanned by these eigenvectors. This process is often referred to as principal component analysis (Joliffe, 1986). Our new explanatory variables are ranked in descending order or their correlation to the spread. The variables are then added one at a time to the model until the validation error reaches a minimum.

This approach is not as computational demanding, but it has a drawback: our new explanatory variables are linear combinations of the original variables, so their meaning is somewhat lost. This approach is therefore only justifiable if it produces significantly better results than methods making use of the original explanatory variables which it does not here.

Choosing a Model

Figure 3 exhibits the expected behavior: as the number of model parameters increase the training error decreases, while the validation error goes through a minimum for five explanatory variables. This model has small bias and passes both run tests.

One should however keep in mind the previous remarks that the fewer free parameters the better our confidence in the result. For the monthly data we *do* have a separate test set and the performance on this test set confirms that for both search techniques we see a minimum on the test set at about five free model parameters. Both sequential techniques give similar performance but potential users (in the finance community) must feel comfortable using the model chosen. There is something less intuitive about using the principal component approach which generates new variables from a linear combination of other variables. For these reasons, we use the "best correlation to the desired value" approach in predicting the future.

An independent test may be run using bootstrap techniques (chapters 9 and 17 of Efron and Tibshirani, 1993). This is a resampling technique that is very computationally expensive and may be applied with caution when there is a possibility that the time samples are correlated. In our case, the spreads are highly correlated (even over a three month time span) but the residuals have very low correlation. Therefore, we performed what is termed "bootstrapping residuals" to obtain numbers very close (within 1%) to those obtained using the leave-one-out procedures.

Conclusions

In predicting the future using quarterly data, we used the best six of eleven variables and the best six of twenty-one variables (Figure 4). Both predictions show a slight decline in the spread followed by an increase. For the monthly data, recall that we used 150 training points to train our model and concluded that five variables were best for the prediction of the test data. Since we are now to predict the future, we want to use all the data (200 points) as a training set. Using a leave-one-out procedure on these 200 points, we arrived at a minimum of the NMSEP at six variables and get a similar curve (Figure 4) to that of the quarterly data. Therefore, all these techniques lead to the conclusion that we should not hedge. Subsequent to the completion of this study, the spread for the first three quarters of 1994 were made available. The spread has indeed decreased the first two quarters and increased in the third quarter, consistent with our predictions.

References

Efron, Bradley, and Gong, Gail (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation", *The American Statistician*, vol 37, no. 1, pp. 36-48.

Efron, Bradley and Tibshirani, Robert J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall.

Guyon, I. and Vapnik, V. N. and Boser, B. E. and Bottou, L. and Solla, S. A. (1992), *Structural Risk Minimization for Character Recognition*, "Advances in Neural Information Processing Systems", Morgan Kaufman.

Johnston, J. (1984), *Econometric Methods*, McGraw-Hill.

Joliffe, I.T. (1986), *Principal Component Analysis*, Springer-Verlag.

Miller, Alan J. (1990), *Subset Selection in Regression*, Chapman and Hall.

Sprent, P. (1993), *Applied Nonparametric Statistical Methods*, Chapman and Hall.

Vapnik, Vladimir (1982), *Estimation of Dependences Based on Empirical Data*, Springer-Verlag.

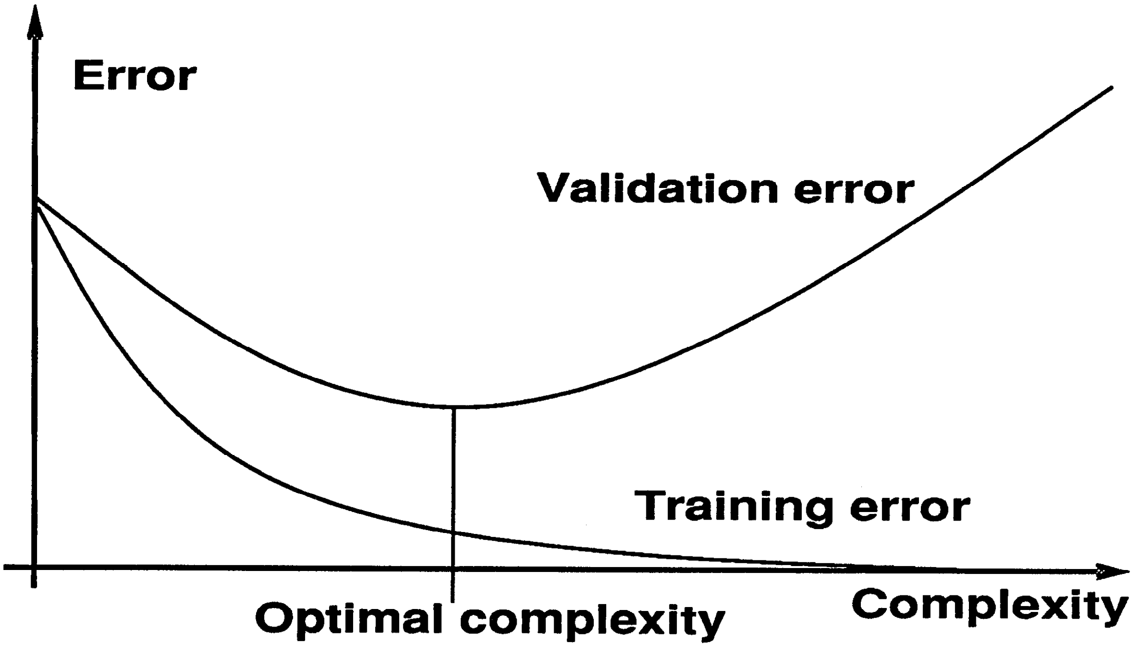

Figure 1. Training and validation error versus complexity. Complexity increase can be either an increase in the order of the approximating function or an increase in the number of explanatory variables.

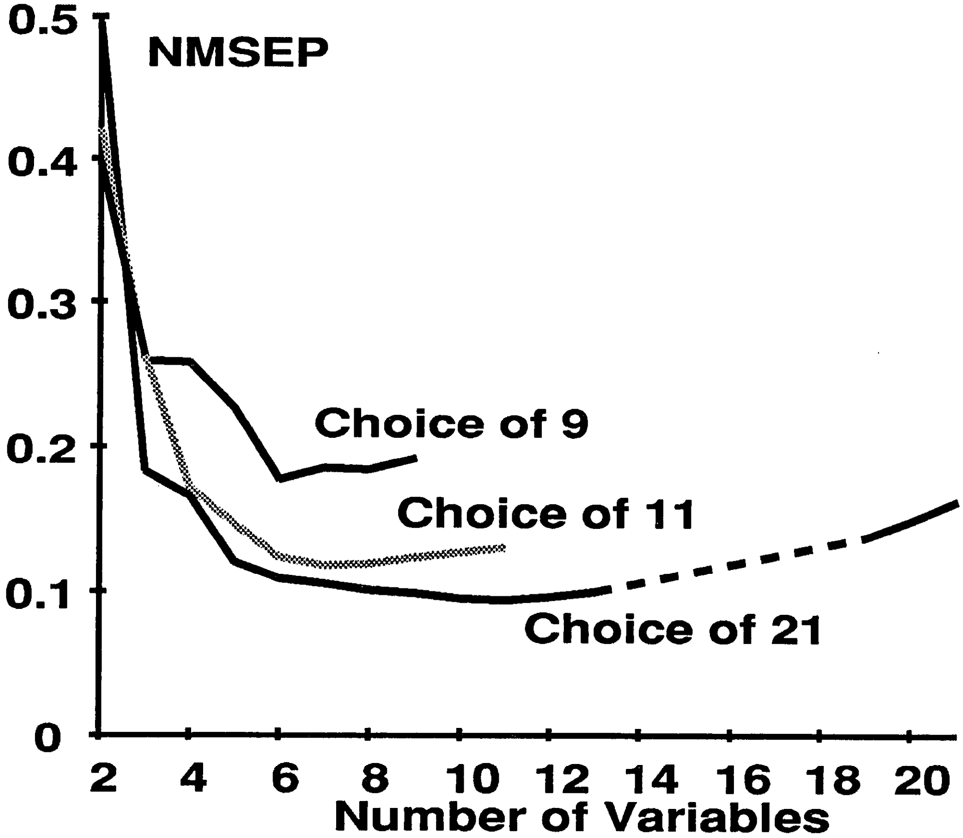

Figure 2. Normalized mean square error of prediction versus number of variables for quarterly data.

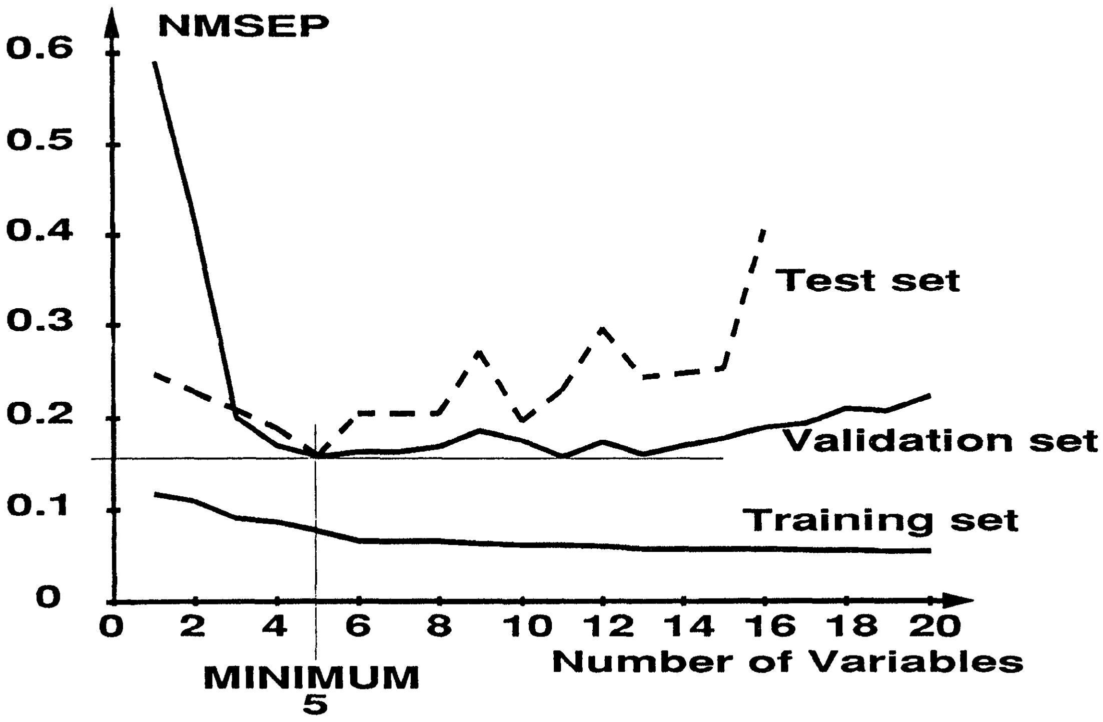

Figure 3. Plot of NMSEP versus number of variables using best correlation to spread.

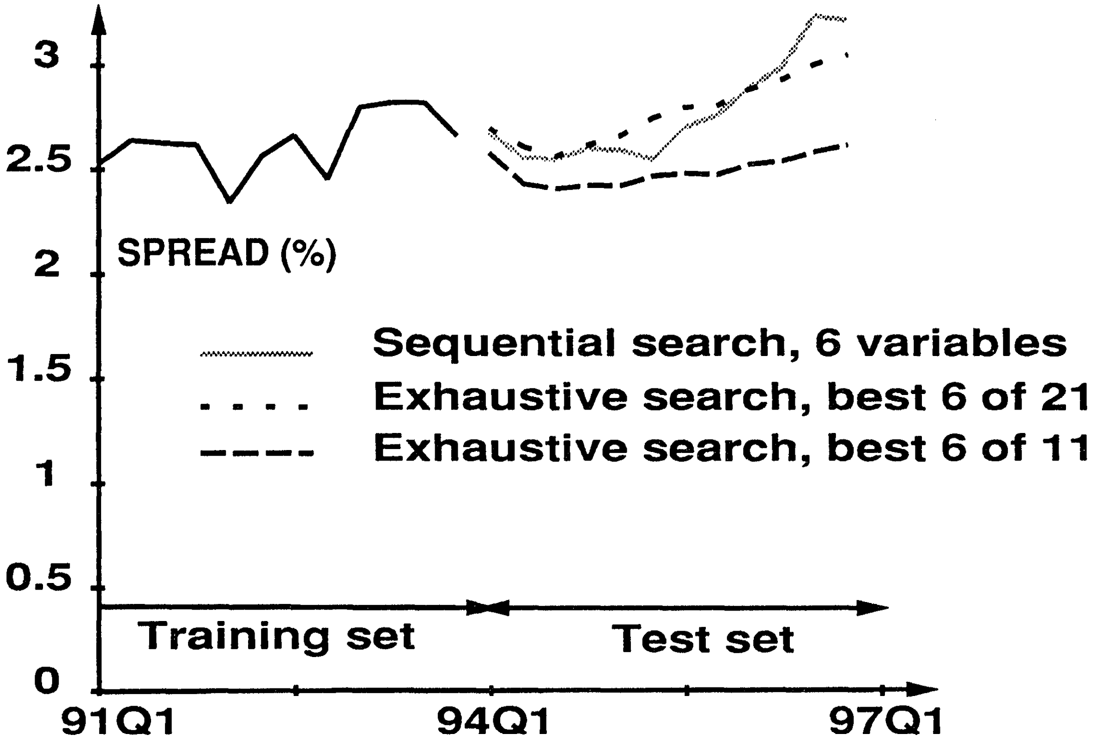

Figure 4. Predictions of the future. The future starts at 1994, Quarter 1.