Using Interval Function Approximation to Estimate Uncertainty

Chenyi Hu

Computer Science Department University of Central Arkansas Conway, AR 72035, USA CHu@uca.edu

Summary. Uncertainties in the real world often appear as variabilities of observed data under similar conditions. In this paper, we use interval functions to model uncertainty and function volatility. To estimate such kinds of functions, we propose a practical interval function approximation algorithm. Applying this algorithm, we have studied stock market forecasting with real economic data from 1930-2004. The computational results indicate that interval function approximation can produce better quality forecasts than that obtained with other methods¹.

1 Introduction

1.1 Interval Function

Functions have been among the most studied topics in mathematics and applications. Provided in analytical form, a function can be easily examined for its properties. However, in real world applications, the analytical form of a function is often unknown. To discover a function that properly models an application is a major challenge. Hence, computational methods on interpolation and approximation are often applied in estimating a function. Also, the main objective of studying differential and integral equations is to search for the unknown function that satisfies given conditions either theoretically or computationally.

Real world observations often differ from from the exact mathematical definition of a function. Even for a fixed x , the observed values of y may be different from time to time. These kinds of un[cer](#page-11-0)tainties are traditionally considered as effects of random noise and are modeled with probability theory. These variations of the value of a function f , for a given x , are often within a finite interval rather than completely random. Also, due to impr[ecise](#page-11-1) measurement and control, the value of x can be within an interval rather than an exact point. This means that an observed data pair can be represented as an interval valued pair (**x**, **y**) rather than precise point (x, y) .

¹ The computational results have been published recently [12]. This paper is a generalized abstraction.

V.-N. Huynh et al. (Eds.): Interval/Probabilistic Uncertainty, ASC 46, pp. 341–352, 2008. springerlink.com c Springer-Verlag Berlin Heidelberg 2008

Definition 1. Let f be a mapping from $\mathbb{R}^n \to \mathbb{R}$ and **x** be an interval vector in \mathbb{R}^n (i. e. each component of **x** is an interval in \mathbb{R}). If for any interval vector **x** there is an interval **y** such that $f(\mathbf{x}) = \mathbf{y}$, then f is an interval function.

We use Figure 1 to illustrate a volatile function that can be better modeled with a interval function. In observing the function values in the figure repeatedly, due to imprecise measurement and control, one may obtain different values of y even for a "fixed" x. More importantly, even one can control x precisely and get the exact y , the point data pairs (x, y) can be misleading when use them in classical function interpolation and approximation. Therefore, an observation recorded as (**x**, **y**) should be more appropriate.

Fig. 1. A volatile function

We say that a real valued function f is 'volatile' in a domain D , if within any small subset of D the sign of the derivatives of f alternates frequently. Figure 1 presents a 'volatile' function. Real world examples of volatile functions include stock prices during any volatile trading day and recorded seismic wave. As shown in Figure 1, for a volatile function, it would be more appropriate to use an interval valued pair to record an observation.

1.2 The Objective of This Paper

Using observed discrete data pairs (x, y) to compu[tatio](#page-11-2)nally approximate an unknown function has been intensively studied. Numerical polynomial interpolation and the least squares approximation are the classical methods in scientific computing. In this paper, we view uncertainty as function volatility modeled with interval valued function. Our objective is to establish a general algorithm that can approximate an unknown interval function. In other words, we try to estimate an interval function from a collection of interval valued pairs (\mathbf{x}, \mathbf{y}) . Algorithms on interpolating interval functions have been discussed in [13]. In

this paper, we focus on approximation, specifically, on the least squares approximation, since it is probably the most broadly used computational method in function approximation.

The rest of this paper is organized as the follow. Section 2 reviews briefly the classical least-squares approximation. Section 3 presents an algorithm for interval function least squares approximation. Section 4 discusses assessment indicators. Section 5 presents a case study. Section 6 concludes the paper.

2 Least Squares Approximation

In this section, we briefly review the principle and computational methods of the ordinary least squares approximation.

2.1 Basis of a Function Space

Let us start with basic concepts related to a function space first.

Definition 2. Let F be a function space and $\Phi = {\phi_0, \phi_1, \dots, \phi_n, \dots}$ be a set of functions in F. We say that Φ is a basis of F if for any function $f \in F$ and any given $\epsilon > 0$ there is a linear combination of ϕ , $f = \sum$ j $\alpha_j \phi_j$, such that

 $|f(x) - \sum$ j $|\alpha_j \phi_j| < \epsilon$ for all x in the domain.

For example, the set $\{1, x, x^2, \dots\}$ is a basis of polynomial function space as well as a basis of a function space that consists of all continuous functions. Of course, there are other bases for a function space. For example, Chebychev polynomials, Legendre polynomials, sine/cosine functions, and others are commonly used as bases in approximating continuous functions.

2.2 The Least Squares Principle

For a continuous function f (even with countable discontinuities), we may approximate it as $f(x) \approx \sum$ $0 \leq j \leq m$ $\alpha_j \phi_j(x)$, where $\phi_j(x)$ is a preselected set of m basis functions. To determine the coefficient vector $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_m)^T$, the

least squares principle requires that the integral of the squares of the deferences between $f(x)$ and \sum $0 \leq j \leq m$ $\alpha_j \phi_j(x)$ is minimized. In other words, applying

the least-squares principle in approximating a function f , one selects the vector

$$
\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_m)^T
$$
 that minimizes
$$
\int \left(f(x) - \sum_{0 \le j \le m} \alpha_j \phi_j(x) \right)^2 dx.
$$

2.3 Discrete Algorithm

In real world applications, one usually only knows a collection of N pairs of (x_i, y_i) rather than the function $y = f(x)$. Therefore, one minimizes the total

 $\sum_{m=1}^{N}$ $i=1$ $\sqrt{2}$ $\left(y_i - \sum_{0 \leq i \leq n}$ $0 \leq j \leq m$ $\alpha_j \phi_j(x_i)$ ⎞ \overline{I} 2 dx instead. The classical algorithm that computationally determines the coefficient vector α is as the follow:

Algorithm 1

- (i) Evaluate the basis functions $\phi_j(x)$ at x_i for all $1 \leq i \leq N$ and $1 \leq j \leq m$;
- (ii) Form the matrix

$$
A = \begin{pmatrix} N & \Sigma_i \phi_1 & \Sigma_i \phi_2 & \cdots & \Sigma_i \phi_m \\ \Sigma_i \phi_1 & \Sigma_i \phi_1^2 & \Sigma_i \phi_1 \phi_2 & \cdots & \Sigma_i \phi_1 \phi_m \\ \Sigma_i \phi_2 & \Sigma_i \phi_2 \phi_1 & \Sigma_i \phi_2^2 & \cdots & \Sigma_i \phi_2 \phi_m \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \Sigma_i \phi_m & \Sigma_i \phi_m \phi_1 & \Sigma_i \phi_m \phi_2 & \cdots & \Sigma_i \phi_m^2 \end{pmatrix}
$$
(1)

and the vector

$$
b = (\Sigma_i y_i \ \Sigma_i y_i \phi_1(x_i) \ \ \Sigma_i y_i \phi_2(x_i) \ \ \cdots \ \Sigma_i y_i \phi_m(x_i))^T; \tag{2}
$$

(iii) Solve the linear system of equations $A\alpha = b$ for α .

The linear system of equations $A\alpha = b$ above is called the normal equation. In stead of normal equations, a more current approach applies a design matrix with a sequence of Householder transformations to estimate the vector α . For details about Householder transformations, it is out of the main scope of this paper, readers may refer [16] or most books that cover computational linear algebra. Although the basic idea of this p[ap](#page-10-0)er ca[n](#page-10-1) [b](#page-10-1)e applied to both approaches, we use the normal equation approach in the rest of is paper for its simplicity.

2.4 Time Series and Slicing-Window

We now switch our attention to the dataset. In real world applications, an observed data pair (x_i, y_i) is often associated with a specific time. The collection of data pairs, if ordered chronically, is called a time series. Time series have been extensively studied for prediction and forecasting [5] and [7]. Rules and functions often rely on a specific time period. We call it time-varying, that is, the relationship is valid for a limited time period. Therefore, in applying function approximation on a time series, one should use only data inside an appropriate time-window to estimate the relationship. By slicing the time-window (also called rolling), one obtains a sequence of function approximations such that each of them valid only for a specific time-window.

3 Interval Function Approximation

Previous studies on least squares approximation mostly assume, if not all, point valued data. There are several computational issues that need to be considered in order to apply Algorithm 1 on interval valued pairs (\mathbf{x}, \mathbf{y}) to approximate an interval function.

3.1 Computational Challenges

With interval arithmetic [17], it is straightforward to perform both steps 1 and 2 in Algorithms 1. However, it presents a challenge in the step 3. This is because the normal equations are now interval systems of linear equations $\mathbf{A}\alpha = \mathbf{b}$. The solution set of an interval linear system of equations is mostly irregular shaped and non-convex [18]. A naive application of interval arithmetic to bound the solution vector α may cause serious overestimation due to the wrap effects, and then negatively affect the approximation quality. Using the design matrix approach would not solve the problem since finding a Householder transformation for an interval matrix remains a challenge.

3.2 An Inner Approximation Approach

While an interval **x** is usually presented by its lower and upper bounds as $\mathbf{x} =$ [x, \mathematics], it can also be represented by its midpoint mid(**x**) = $\frac{x + \overline{x}}{2}$ and its width $w(\mathbf{x}) = \overline{x} - \underline{x}$. This creates a two-step approach where we consider the midpoint and width separately in each of the two steps.

Instead of finding the lower and upper bounds of the interval vector α in the step 3 of Algorithm 1, let us first try to find its midpoint vector, which is a scalar vector. This suggests us to match the center of two interval vectors $A\alpha$ and **b** in the interval linear system of equations $\mathbf{A}\alpha = \mathbf{b}$. Let A_{mid} be the midpoint matrix of \bf{A} , and b_{mid} be the midpoint vector of \bf{b} . We solve the non-interval linear system of equations $A_{mid}\alpha = b_{mid}$ for α .

We would like to emphasize that the result of $y = f(x) \approx \alpha_0 + \sum$ $1 \leq j \leq m$ $\alpha_j \phi_j(\mathbf{x})$

is an interval even when we use the midpoint of the interval vector α in the calculation. This is because of that the independent variable **x** is interval valued. However, by collapsing an interval vector α to its midpoint, we could reasonably expect that the approximation is an inner interval approximation.

3.3 Width Adjustment

Now, let us consider the width. One may try to look for the width vector of α vector or for the width of **y**. There can be different computational heuristics too. For example, one may select a width vector that makes $A\alpha$ as close as possible to **b** in the step 3 of Algorithm 1. One may also use widths to perform least-square approximation to estimate the width of **y**. Another computational heuristic is to adjust the width by multiplying a scale factor to the inner approximation. In our case study, we adopted the later approach. We believe that there are still many open questions for further study on width adjustment.

3.4 Interval Least-Squares Approximation

By summarizing the above discussions, we revise Algorithm 1 to Algorithm 2 below for interval function least squares approximation.

Algorithm 2

- (i) Input available interval data pairs $(\mathbf{x}_i, \mathbf{y}_i)$ for $1 \leq i \leq N$;
- (ii) Evaluate matrix A and vector b with interval arithmetic;
- (iii) Find A_{mid} and b_{mid} , the midpoint matrix of **A** and the midpoint vector of **b**, respectively;
- (iv) Solve the linear systems of equations: $A_{mid}\alpha = b_{mid}$;
- (v) Apply the vector α to calculate an inner approximation with interval arithmetic;
- (vi) Modify the initial approximation with a width adjustment.

3.5 Other Approaches to Obtain an Interval Approximation

One may obtain an interval approximation without using interval arithmetic at all. The lower and upper bounds of interval data pairs $(\mathbf{x}_i, \mathbf{y}_i)$ form two collections of point data $(\underline{x}_i, \underline{y}_i)$ and $(\overline{x}_i, \overline{y}_i)$. By applying point least square approximation to them separately, one can obtain two point estimations. These two estimations can form an interval estimation. We call this approach the minmax interval approximation. This has been reported and applied in [10] and [11].

Another way to obtain an interval approximation is to apply classical statistic/probabilistic approach. By adding to and subtracting from a point approximation a certain percentage of standard deviations, one can obtain forecasting intervals. In the literature, this is called a confidence interval. However, the case study in Section 5 of this paper implies that, at least in certain cases, interval function lease squares approximation may produce better computational results than that obtained with the min-max interval and confidence interval.

4 Assessing Interval Function Approximation

There are different ways to produce an interval approximation. A immediate question is how to assess the quality of different interval estimations. We define two measurements for quality assessment of an interval approximation.

Definition 3. Let y_{est} be an approximation for the interval **y**. The absolute error of the approximation is the absolute sum of the lower and upper bounds errors, i.e. $|y_{est} - y| + |\overline{y}_{est} - \overline{y}|.$

Example 1. If one obtained $[-1.02, 1.95]$ as an approximation for the interval $[-1.0, 2.0]$, then the absolute error of the estimation is $|(-1.02)-(-1.0)|+|1.95 |2.0| = 0.02 + 0.05 = 0.07$.

Since both y_{est} and y are intervals, an additional meaningful quality measurement can be defined. The larger the overlap between the two intervals the better the approximation should be. By the same token, the less the non-overlap between the two intervals the more accurate the forecast is. In addition, the accuracy of an interval estimation should be between 0% and 100%. By using the notion of interval width, which is the difference of the upper and lower bounds of an interval, we can measure the intersection and the union (or the convex hull) of the two intervals. Let $w()$ be the function that returns the width of an interval. Then, we define the concept, named the accuracy ratio of an interval approximation, as the follow.

Definition 4. Let y_{est} be an approximation for the interval **y**. The accuracy ratio of the approximation is $\frac{w(\mathbf{y} \cap \mathbf{y}_{est})}{w(\mathbf{y} \cup \mathbf{y}_{est})}$ if $(\mathbf{y} \cap \mathbf{y}_{est}) \neq \emptyset$. Otherwise, the accuracy ratio is zero.

Example 2. Using $[-1.02, 1.95]$ to approximate the interval $[-1.0, 2.0]$, the accuracy ratio is $\frac{w([-1.02, 1.95] \cap [-1.0, 2.0])}{w([-1.02, 1.95] \cup [-1.0, 2.0])} = \frac{w([-1.0, 1.95])}{w([-1.02, 2.0])} = \frac{2.95}{3.02} = 97.68\%$.

As in classical statistics, for a collection of interval estimations, one can calculate the mean and standard dev[iat](#page-10-2)ion of [th](#page-11-0)e absolute error and the accuracy ratio. Furthermore, one may also apply probability theory to perform comparisons of different approximations.

5 Case Study: Forecasting the S & P 500 Index

The S & P 500 index is a broadly used indicator for the overall stock market. Using interval least squares approximation, we have performed $S \& P 500$ annual forecast with astonishin[g co](#page-11-3)mputational results[9] and [12]. We report it here again as a case study with comparisons against the result obtained with traditional ordinary least squares forecasting.

5.1 The Model

Driven by macroeconomic and social factors, the stock market usually varies with time. The main challenge in studying the stock market is its volatility and uncertainty. The arbitrage pricing theory (APT) [20] provides a framework that identifies macroeconomic variables that significantly and systematically influence stock prices. By modeling the relationship between the stock market and relevant macroeconomic variables, one may try to forecast the overall level of the stock market.

The model we use in this case study is a broadly accepted one by Chen, Roll and Ross (1986). According to their model, the changes in the overall stock market value (SP_t) are linearly determined by the following five macroeconomic factors: the growth rate variations of seasonally-adjusted Industrial Production Index (IP_t) ; changes in expected inflation (DI_t) and unexpected inflation (UI_t) ; default risk premiums (DF_t) ; and unexpected changes in interest rates (TM_t) . This relationship can be expressed as

 $SP_t = a_t + I_t (IP_t) + U_t (UI_t) + D_t (DI_t) + F_t (DF_t) + T_t (TM_t)$

By using historic data, one may estimate the coefficients of the above equation to forecast changes of the overall stock market. There is a general consensus in

the financial literature, that relationships between financial market and macroeconomic variables are time-varying. Hence the coefficients are associated with a time-window.

In the literature, it is called an *in-sample forecast* if using the obtained coefficients in a time-window and the equation above to calculate the SP for the last time period in the time-window. It is called an *out-of-sample forecast* if using the obtained coefficients in a time-window to calculate the SP for the first time period that immediately follows the time-window [4]. By slicing the timewindow (also called rolling), one obtains a sequence of coefficients and forecasted SP values. The overall quality of forecasting can be measured by comparing the forecasts against actual SP values. In practice, the out-of-sample-forecast is more useful than in-sample-forecast because it can make predictions.

5.2 The Data

So far the primary measurements used in economics and finance are quantified points. For instance, a monthly closing value of an index is used to represent the index for that month even though the index actually varies during that month. The available data in this case study are monthly data from January 1930 to December 2004. We list a portion of the data here.

In this case study, we use a time-window of ten years to obtain the out-ofsample annual forecasts for 1940-2004.

5.3 Interval Rolling Least Squares Forecasts

To perform interval rolling least squares forecasts, we need interval input data. From the provided monthly data, for each of the attributes, we choose its annual minimum and maximum to form the interval input data. By applying Algorithm 2, we obtain initial forecasts first. For each of them, we then adjust

the width of the predicted $S \& P 500$ interval to the average of that of those within the time-window. The program was written in C_{++} . The software package IntBLAS [19] was applied for interval and related linear algebra operations. Figure 2 illustrates the out-of-sample annual interval forecasts.

Out-of-Sample 10-Year Rolling Interval Forec

Fig. 2. Out-of-Sample Annual Interval Forecasts(1940-2004)

For the purpose of quality comparison, we calculated the annual point forecasts that are commonly used in financial study. We obtained the out-of-sample annual forecasts (in percent) for a period of 1940-2004. The out-of-sample annual point forecasts have an average absolute forecasting error of 20.6% with a standard deviation of 0.19. By adding to and subtracting from the point-forecasts with a proportion of the standard deviation, we may form confidence interval forecasts with 95% statistical confidences.

It is worth pointing out that the ranges of Figure 2 are significantly less than that of Figure 3 at the ratio only about 14%.

5.4 Quality Comparisons

To assess the quality of the above forecasts, we use the following indicators: (1) the average absolute forecast error, (2) the standard deviation of forecast errors, (3) the average accuracy ratio, and (4) the number of forecasts with 0% accuracy. We summarize the statistics of the quality indicators in the table below.

All measured indicators for forecasting quality in the table suggest that interval OLS significantly outperform point-based forecasts with a much less mean forecast error. The much smaller standard deviations produced by the interval approaches indicate that the interval forecasting is more stable than other

Figure 1, Out-of-sample 10-year rolling OLS forecasts

Fig. 3. Out-of-Sample Annual Point Forecasts (1940-2004)

Table 1. Quality comparison of annual forecasts (1940-2004)

Methods/Item			Absolute standard Accuracy Number of
	mean error deviation ratio		0 accuracy
OLS.	0.20572	0.18996 NA	ΝA
std dev. 95% confidence 0.723505 0.311973 0.125745			5
$Min-Max$ interval 0.066643		0.040998 0.4617	$\left(\right)$
Initial interval Feast	0.073038	0.038151 0.385531	$\left(\right)$
Interval Fcast	0.0516624 0.032238 0.641877		$\left(\right)$

comparing methods. Compared with the point-based confidence interval forecasting, interval methods produce a much higher average accuracy ratio. The interval scheme with width adjustments further improves the overall forecasting quality of initial approximations in terms of the higher average accuracy ratio. All forecasts with interval computing have a positive accuracy ratio while a number of the point-based confidence intervals has zero accuracy.

6 Conclusion

In this paper, we model uncertainty as volatilities of a function. It is more reasonable to record volatile data as interval valued nodes of an interval function rather than point values. To apply classical least squares approximation with discrete interval valued nodes, we use interval arithmetic to obtain the normal equation. By using the midpoint approach, we calculate an inner approximation initially. Then, we adjust its width with computational heuristics. Although the function to be approximated is unknown, we can still assess the quality of an interval approximation statistically with provided data. These quality indicators include absolute error, accuracy ratio, and their means and standard deviations.

Using this approach, in our case study, we performed annual forecasts for the S & P 500 index from 1940-2004 with real economical data. Although it is merely one of the initial attempts to use interval methods in financial forecasting, the empirical results provide astonishing evidence that interval least squares approximation may outperform traditional point approaches in terms of the overall less mean error and higher average accuracy ratio. Hence, interval methods have a great potential in dealing with uncertainty.

Acknowledgments

This research is partially supported by the US National Science Foundation grants CISE/CCF-0202042 and CISE/CCF-0727798.

The financial data was provided by Professor Ling T. He in Finance and Economics at the University of Central Arkansas. He initiated and actively participated in the case study.

Mr. Michael Nooner and anonymous referees have provided helpful comments for improving the quality of this paper.

References

- 1. Chatfield, C.: Calculating interval forecasts. Journal of Business and Economic Statistics 11, 121–144 (1993)
- 2. Chatfield, C.: Prediction intervals for time-series forecasting. In: Armstrong, J.S. (ed.) Principles of forecasting: Handbook for researchers and practitioners, Kluwer Academic Publisher, Boston (2001)
- 3. Chen, N., Roll, R., Ross, S.: Economic forces and the stock market. Journal of Business 59, 383–403 (1986)
- 4. Fama, E., French, K.: Industry costs of equity. Journal of Financial Economics 43, 153–193 (1997)
- 5. Gardner, E.: A simple method of computing prediction intervals for time series forecasts. Journal of Management Science 34, 541–546 (1988)
- 6. Granger, C.: Can we improve the perceived quality of economic forecasts? Journal of Applied Econometrics 11, 455–473 (1996)
- 7. Gooijer, J., Hyndman, R.: 25 years of time series forecasting. Journal of Forecasting 22, 443–473 (2006)
- 8. He, L.: Instability and predictability of factor betas of industrial stocks: The flexible least squares solutions. The Quarterly Review of Economics and Finance 45, 619– 640 (2005)
- 9. He, L., Hu, C.: The stock market forecasting: an application of the interval measurement and computation. In: The 2nd Int. Conf. on Fuzzy Sets and Soft Comp. in Economics and Finance, St. Petersburg, Russia, pp. 13–22 (2006)
- 10. He, L., Hu, C.: Impacts of interval measurement on studies of economic variability: evidence from stock market variability forecasting. Journal of Risk Finance (to appear)

- 11. He, L., Hu, C.: Impacts of interval computing on stock market variability forecasting[.](http://www.cs.utep.edu/interval-comp/main.html) [Journal](http://www.cs.utep.edu/interval-comp/main.html) [of](http://www.cs.utep.edu/interval-comp/main.html) [Computational](http://www.cs.utep.edu/interval-comp/main.html) [Economics](http://www.cs.utep.edu/interval-comp/main.html) [\(to](http://www.cs.utep.edu/interval-comp/main.html) [appear\)](http://www.cs.utep.edu/interval-comp/main.html)
- 12. Hu, C., He, L.: An application of interval methods to the stock market forecasting. Journal of Reliable Computing 13, 423–434 (2006)
- 13. Hu, C., et al.: An interval polynomial interpolation problem and its Lagrange solution. Journal of Reliable Computing 4(1), 27–38 (1998)
- 14. Hu, C., Xu, S., Yang, X.: A review on interval computation software and applications. Journal of Computational and Numerical Analysis and Applications 1(2), 149–162 (2002)
- 15. Interval Computations, http://www.cs.utep.edu/interval-comp/main.html
- 16. Moler, C.: Numerical computing with MATLAB. SIAM (2004)
- 17. Moore, R.: Methods and applications of interval analysis. SIAM Studies in Applied Mathematics (1979)
- 18. Neumaier, A.: Interval methods for systems of equations. Cambridge University, Cambridge (1990)
- 19. Nooner, M., Hu, C.: A computational environment for interval matrices. In: NSF 2006 Workshop on Reliable Engineering Computing, Savannah, GA, pp. 65–74 (2006)
- 20. Ross, S.: The arbitrage theory of capital asset pricing. Journal of Economic Theory 13, 341–360 (1976)