Obtaining Quantitative Predictions From Monotone Relationships

Joseph Hellerstein

IBM Research
T. J. Watson Research Center
Yorktown Heights, NY 10598

Abstract: Tasks such as forecasting, diagnosis, and planning frequently require quantitative predictions. Typically, quantitative predictions are obtained by characterizing a system in terms of algebraic relationships and then using these relationships to compute quantitative predictions from numerical data. For real-life systems, such as mainframe operating systems, an algebraic characterization is often difficult, if not intractable. This paper proposes a statistical approach to obtaining quantitative predictions from monotone relationships -- non-parametric interpolative-prediction for monotone functions (NIMF). NIMF uses monotone relationships to search historical data for bounds that provide a desired level of statistical confidence. We evaluate NIMF by comparing its predictions to those of linear least-squares regression (a widely-used statistical technique that requires specifying algebraic relationships) for memory contention in an IBM computer system. We find that NIMF consistently produces better predictions, which we attribute (in part) to using an accurate monotone relationship instead of an approximate algebraic relationship. We also show that using monotone relationships to produce quantitative predictions greatly facilitates explaining the resulting predictions.

1. Introduction

Numerical or quantitative predictions of system behavior are frequently required in tasks such as forecasting, diagnosis, and planning. Typically, quantitative predictions are obtained by characterizing a system in terms of algebraic relationships and then using these relationships to compute quantitative predictions from numerical data. Unfortunately, for real-life systems an algebraic characterization is often difficult, if not intractable. This paper describes an approach to obtaining quantitative predictions from monotone relationships, and applies this approach to predicting memory contention in an IBM computer system.

Why is it often so difficult to obtain accurate algebraic characterizations of real-life systems? Our experience with analyzing measurements of computer systems, in particular the IBM operating system Virtual Machine/System Product (VM/SP), suggests that the major impediment to an algebraic characterization is the absence of sufficiently detailed information about the system's operation. For example, the performance of VM/SP systems is often constrained by contention for the first sixteen megabytes of main memory (referred to as **low memory**), even though there may be sixty-four megabytes or more of main memory. Low-memory contention is a consequence of the operating system using twenty-four bit addressing and requiring that many system services use memory that is directly addressable by the operating system. A key indicator of low-memory contention is the rate at which pages below sixteen megabytes are taken from users in the multi-programming set. Constructing an algebraic relationship between this measure and parameters such as the virtual machine input/output rate and the number of logged-on users requires using these parameters to quantify the frequency and execution times of operating-system service-requests (e.g., spool operations, messages exchanged through the inter-user communication vehicle, and file opens) as well as the low-memory demands of each service requested (e.g., bytes required, page and/or cache alignments, and algorithm used when fixed-sized pools are empty). Unfortunately, such detailed information is rarely available.

When we are unable to construct algebraic relationships, we often have qualitative knowledge in the form of monotone relationships. For example, in VM/SP intuition and experience strongly suggest that low-memory contention increases with the virtual machine input/output rate and the number of logged-on users. Another example in CPU-bound VM/SP systems is the relationship between response time and a workload characterized by CPU utilization and the rate of small transactions. Again, an algebraic characterization appears to be intractable; however, we expect response time to decrease with the rate of small transactions (since more small transactions means fewer large ones, in a resource-constrained system) and to increase with CPU utilization. Still other examples where monotone relationships apply but algebraic relationships are difficult to construct include the following: relating lock contention to user activity, relating disk operations to the virtual machine input/output rate, and relating working set size to the virtual machine input/output rate and CPU demands.

If an accurate algebraic characterization of the system is unavailable, how can we obtain quantitative predictions? One approach is to approximate the unknown algebraic relationship by a simple function, such as a polynomial. Herein, we present an alternative approach in which quantitative predictions are computed directly from monotone relationships. Our experience with this approach, as shown in section 3, suggests that using an accurate monotone relationship frequently results in better predictions than using an approximate algebraic relationship.

Our approach to prediction is statistical. Referred to as *non-parametric interpolative-prediction for monotone functions (NIMF)*, our approach assumes the existence of historical data, which is appropriate for domains such as computer performance, financial analysis, and demographic studies. Often, the historical data is highly variable; indeed, providing a point estimate (e.g., an expected value) may be meaningless. For this reason, NIMF produces *prediction intervals* at a user-specified *confidence level* (e.g., 75%). A prediction interval consists of a lower bound (y_L)

and an upper bound (y_H) with the following interpretation: The probability that the predicted value lies between y_L and y_H is at least as large as the confidence level. NIMF uses monotone relationships to search the historical data for y_L and y_H.

This paper contributes to two areas of research literature. The first is the use of monotone relationships as a knowledge representation. Monotone relationships have been used in many contexts, such as predicting changes in qualitative state (e.g., [Forbus, 1984], [Kuipers, 1984], [Kuipers, 1986], and [DeKleer84]), monitoring dynamic systems [Dvorak and Kuipers, 1989], explaining quantitative predictions produced by algebraic relationships (e.g., [Apte and Hong, 1986] and [Simmons, 1986]), and analyzing financial statements [Kosy and Wise, 1984]. More recently, there has been interest in the probabilistic semantics of qualitative influences [Wellman, 1987] and probabilistic considerations in qualitative simulation ([Dvorak and Sacks, 1989]). Our work further extends the application of monotone relationships by demonstrating their use in quantitative, statistical prediction for situations in which numerical data are available but an algebraic characterization is intractable.

This paper also contributes to the area of expert systems for statistical analysis. A central concern in statistics is improving the accuracy of predictions, or *model diagnosis*. Existing approaches to expert systems for statistical analysis use traditional statistical techniques, such as least-squares regression (e.g., [Dickson and Talbot, 1986], [Hietala, 1986], [Hahn, 1985], and [Gottinger, 1988]). These techniques require that users specify their models in terms of algebraic relationships, which complicates model diagnosis when such relationships are either unknown or very complex. NIMF offers a way to avoid these problems by using monotone relationships instead of algebraic relationships. Also, NIMF facilitates model diagnosis by making it easy to explain predictions in terms of monotone relationships.

The remainder of this paper is organized as follows. Section 2 describes the NIMF technique. Section 3 evaluates NIMF by comparing its predictions to those of least-squares regression, a widely-used statistical technique that requires specifying algebraic relationships. Section 4 describes how to explain NIMF predictions. Our conclusions are contained in section 5.

2. Approach

Our approach to obtaining quantitative predictions from monotone relationships was motivated by observations of performance analyst who tune VM/SP computer systems. One aspect of tuning is workload assignment, in which users of computing services are assigned to one of several computer systems in a manner so that computing resources (e.g., CPU, input/output bandwidth, and memory) are utilized within prescribed guidelines. Clearly, this task requires an ability to predict the resource utilizations of an assignment. The most common approach to predicting the performance of computer systems is based on queueing theory [Kleinrock, 1975]. Queueing theory characterizes computer systems in terms of stochastic processes, which permits deriving algebraic relationships between measurement variables. While queueing theory has proven effective for modeling "active" resources (e.g., CPU, input/output operations), it has not been particularly effective for modeling "passive" resources, such as memory. In large VM/SP computer systems, contention for low memory is often the primary performance bottleneck.

How then do performance analysts predict low-memory contention? Lacking a formal approach to the problem, analysts often use an informal approach. We illustrate this by predicting **LOSTEALRAT** (the rate at which pages in low memory are taken from users in the multi-programming set) from **LOGGED** (the number of logged-on users) and **VIO** (virtual machine input/output rate). (All three variables can be obtained from the Virtual Machine Monitor Analysis Program (VMMAP) [IBM, 1985].) Suppose that a workload assignment would result in a computer system having an average of 500 logged-on users with an average aggregate VIO rate of 500. Although we know of no algebraic equation that relates LOSTEALRAT to LOGGED and VIO, we do have an excellent understanding in terms

monotone relationships. Specifically, for each logged-on user, data structures are allocated in low memory to describe the virtual address space; so we expect LOSTEALRAT to increase with LOGGED. Further, each VIO requires that transient data structures be allocated in low memory, and so LOSTEALRAT should increase with VIO as well. That is,

MR_1 : LOSTEALRAT increases with VIO and LOGGED.

MR_1 provides analysts with an approach to searching historical data for potential bounds. For example, to find a lower bound for the point VIO = 500 and LOGGED = 500, the analyst considers data for which VIO $\leq$ 500 and LOGGED $\leq$ 500. Similarly, finding an upper bound involves examining data for which VIO $\geq$ 500 and LOGGED $\geq$ 500. Figure 1 provides a graphical representation of this approach. This figure depicts historical data by plotting VIO against LOGGED; each point is labelled with its associated LOSTEALRAT. Horizontal and vertical (dotted) lines are drawn through (500,500), thereby dividing the graph into four quadrants. Potential lower bounds are contained in the lower-left quadrant; potential upper bounds are contained in the upper-right quadrant.

Once the set of potential bounds is identified, analysts often resort to heuristics. For example, in Figure 1, the lower bound might be computed by averaging values in the lower-left quadrant or by finding the largest value of LOSTEALRAT within that quadrant. Unfortunately, such heuristics do not indicate the confidence level of the resulting prediction interval, and they certainly do not permit choosing bounds so that a particular confidence level is achieved.

The key to formalizing the above approach is to address randomness in the measurement data. For example, Figure 2 displays scatter plots of LOSTEALRAT vs. LOGGED and LOSTEALRAT vs. VIO for measurements taken from a VM/SP computer system; these plots suggest a high degree of randomness. We say that a monotone relationship exists between the **response variable** y (e.g., LOSTEALRAT) and the **explanatory variables** $x_1, \ldots, x_J$ (e.g., x_1 = VIO and x_2 = LOGGED) if and only if there is a monotone function g such that

$$y_i = g(\mathbf{x}_i) + \varepsilon_i, \tag{1}$$

where y_i is the i-th measurement of the response variable, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,J})$ is the i-th measurement of the explanatory variables, and ε_i is the i-th error term. Randomness is handled by the ε_i, which are assumed to be realizations of continuous, independent, and identically distributed random variables (Hellerstein [1989] relaxes the assumption that error terms be continuous and identically distributed.) We make no assumption about g's functional form. However, we do assume that g's **directional effects** are known; that is, for the j-th explanatory variable (x_j), we know if g is non-increasing or non-decreasing. (If g is differentiable, this is equivalent to knowing the sign of g 's first derivatives.) Since we do not assume that error terms are drawn from a specific distribution and we make no assumption about g's functional form, our approach is *non-parametric*. Further, our approach is appropriate only if there are existing measurements within the region in which a prediction is desired; that is, our approach provides interpolation, not extrapolation. These characteristics of our approach as well as its being applicable only to monotone functions motivate the name *non-parametric interpolative-prediction for monotone functions (NIMF)*.

We now translate the informal approach described at the beginning of this section into a formal statistical technique. The problem addressed can be described as follows. (See Figure 3.) Given $\mathbf{x}^*$, values of explanatory variables at which a prediction is desired, and a description of g in terms of its directional effects, NIMF computes prediction intervals by finding a lower bound (y_L) and an upper bound (y_H) for the unknown response (Y^*) such that

$$P(y_L \leq Y^* \leq y_H) \geq 1 - \alpha, \tag{2}$$

where $1 - \alpha$ is the desired confidence level. Typical values for $1 - \alpha$ are 75%, 90%, and 95%.

The NIMF procedure consists of three steps, as summarized in Figure 4. The first step selects sets of potential bounds by using the monotone relationships that describe g. Specifically, a monotone relationship imposes the following partial order:

$$\mathbf{x}_1 \prec \mathbf{x}_2 \text{ iff for all } j \begin{cases} x_{1j} \le x_{2j} \text{ if } g \text{ is non-decreasing in } x_j \\ x_{1j} \ge x_{2j} \text{ if } g \text{ is non-increasing in } x_j \end{cases}$$

The **set of potential lower bounds**, S_L, is a subset of $\{y_i \mid \mathbf{x}_i \prec \mathbf{x}^*\}$ consisting of the M y_i whose $\mathbf{x}_i$ are closest to $\mathbf{x}^*$ ("Closest to" is defined as the Euclidean distance measure normalized by standard deviation.) By picking $\mathbf{x}_i$ close to $\mathbf{x}^*$, we hope to reduce $|g(\mathbf{x}_i) - g(\mathbf{x}^*)|$ and hence reduce the width of prediction intervals. The **set of potential upper bounds**, S_H, is a subset of $\{y_i \mid \mathbf{x}^* \prec \mathbf{x}_i\}$, and is chosen in the same manner as S_L. Applying these definitions to the data in Figure 1 with $M = 6$, we have

$$\begin{aligned} S_L &= \{0, .1, .2, .3, .6, 1.2\} \\ S_H &= \{.5, .7, 1.1., 1.4, 1.5, 1.6\} \end{aligned} \tag{3}$$

[Note that S_H does not include $y_i = 2.1$ at (VIO, LOGGED) = (640,560), since $M = 6$ and this point is furthest from $\mathbf{x}^* = (500,500)$.]

NIMF's second and third steps select y_L from S_L and y_H from S_H in a manner so that at least a $1 - \alpha$ confidence level is obtained. Our approach is similar to that taken by Bradley [1968] to obtain confidence intervals for distribution percentiles. Assuming that $y_L \le y_H$, it suffices to pick y_L and y_H such that

$$\begin{aligned} P(y_L \le Y^*) &\ge 1 - \frac{\alpha}{2} \\ P(y_H \ge Y^*) &\ge 1 - \frac{\alpha}{2} \end{aligned}$$

To find y_L, we proceed by considering its components. Let $y_i \in S_L$, with $y_i = g(\mathbf{x}_i) + \varepsilon_i$. If g is monotone and we know the directional effect of each x_j, then $g(\mathbf{x}_i) \le g(\mathbf{x}^*)$ (by construction). So

$$\begin{aligned} P(y_i \le Y^*) &= P(g(\mathbf{x}_i) + \varepsilon_i \le g(\mathbf{x}^*) + \varepsilon^*) \\ &\ge P(\varepsilon_i \le \varepsilon^*) \\ &= .5. \end{aligned}$$

(The last step is a result of the error terms being continuous, independent, and identically distributed.) Let $N_L = \min\{M, size(S_L)\}$. Since the ε_i are realizations of independent and identically distributed random variables, the binomial distribution applies:

$$P(\text{ at least } k \text{ elements in } S_L \text{ smaller than } Y^*) \ge \sum_{k=n}^{N_L} \binom{N_L}{n} .5^{N_L}$$

Put differently, let $y_{L,k}$ be the k-th smallest element in S_L. If at least k elements of S_L are less than or equal to Y^*, then $y_{L,k} \le Y^*$. Hence,

$$P(y_{L,k} \le Y^*) \ge \sum_{k=n}^{N_L} \binom{N_L}{n} .5^{N_L}.$$

NIMF's second step computes k_L such that $P(y_{L,k_L} \le Y^*) \ge 1 - \frac{\alpha}{2}$ and k_H such that $P(y_{H,k_H} \ge Y^*) \ge 1 - \frac{\alpha}{2}$. In its third step, NIMF selects the prediction interval bounds; $y_L = y_{L,k_L}$ and $y_H = y_{H,k_H}$.

We illustrate the NIMF procedure by computing a prediction interval for LOSTEALRAT when (VIO, LOGGED) = (500,500); we use the data in Figure 1 with $1-\alpha = 75\%$ and $M = 6$[1]. We have already computed S_L and S_H; they are shown in Eq. (3). Since both sets have six elements, $N_L = N_H = 6$. From step 2 of Figure 4, we observe that

$$\begin{aligned} k_L &= \phi(N_L, \alpha) \\ &= \phi(6, .25), \end{aligned}$$

where

$$\phi(N, \alpha) = \max \{k \mid \sum_{n=k}^{N} \binom{N}{n} .5^N \geq 1 - \frac{\alpha}{2}\}. \tag{4}$$

Or,

$$\phi(6, .25) = \max \{k \mid \sum_{n=k}^{6} \binom{6}{n} .5^6 \geq .875\}.$$

Performing the necessary computations, we determine that $k_L = 2$. For k_H, we have

$$\begin{aligned} k_H &= N_H - \phi(N_H, \alpha) + 1 \\ &= 6 - 2 + 1 \\ &= 5. \end{aligned}$$

We use these indexes to find the prediction interval bounds; y_L is the second smallest element in S_L, and y_H is the fifth smallest element in S_H. That is, $y_L = .1$, and $y_H = 1.5$.

NIMF's ability to produce prediction intervals depends on the historical data provided and the monotone relationships used. When NIMF cannot compute a lower bound, $y_L = -\infty$; when an upper bound cannot be computed, $y_H = \infty$. One situation in which NIMF cannot produce a bound is when there is insufficient historical data; that is, N_L (N_H) is so small that $k_L = 0$ ($k_H = N_H + 1$) at the $1-\alpha$ confidence level. In most computer installations, data are cheap to collect and plentiful; so a missing bound can often be obtained by simply including more data. Alternatively, the analyst can reduce the confidence level.

There is a second situation in which NIMF can produce prediction intervals, but the results are inconsistent with the monotone relationship. This situation occurs when $y_L > y_H$. That is, the $\frac{\alpha}{2}$ percentile of S_L is *larger* than the $1-\frac{\alpha}{2}$ percentile of S_H, where these sets were specifically chosen so that g would take on *smaller* values in S_L than in S_H. (When $1-\alpha = 75\%$, $\frac{\alpha}{2} = 12.5\%$ and $1-\frac{\alpha}{2} = 87.5\%$.) Clearly, this situation makes the monotone relationship suspect. A statistically valid prediction interval can be produced by taking y_L to be the smaller bound and y_H to be the larger bound. However, our feeling is that this situation suggests an error in the underlying model that should be surfaced to the user. An approach to doing to so is presented in section 4.

---

1 $M = 6$ is chosen for illustrative purposes; a more common value for M is 20.

3. Case Study

This section presents a case study in which NIMF's predictions are compared to those of linear least-squares regression (hereafter, just *regression*), a widely-used statistical technique that requires an algebraic specification of variable relationships [Draper and Smith, 1968]. We compare NIMF and regression by using the data in Figure 2 as the historical data from which NIMF potential-bounds sets are obtained and regression constants are estimated. Prediction intervals are then constructed at values of VIO and LOGGED (the $\mathbf{x}^*$ variables) contained in separately acquired test data; the test data also include measurements of LOSTEALRAT at each $\mathbf{x}^*$, which we use to evaluate the prediction intervals.

Prediction intervals are typically evaluated based on two criteria:

- coverage (percent of LOSTEALRAT values in the test data that lie in their prediction interval)
- prediction interval width

Since confidence level is a user-specified parameter, coverage is viewed as a constraint rather than an optimization criteria. So, the preferred technique is the one that minimizes prediction interval width subject to the constraint that coverage is at least as large as the specified confidence level.

First, we briefly describe the regression procedure. A regression model takes the same form as Eq. (1), but stronger assumptions are made: g 's functional form must be known, and (to obtain prediction intervals) ε_i must be normally distributed. A functional form is an algebraic relationship with unknown constants. For example,

$$\hat{y}_i = b_0 + b_1 L_i + b_2 V_i \tag{5}$$

Where:
$\hat{y}_i$ = i-th estimated LOSTEALRAT
L_i = i-th measured LOGGED
V_i = i-th measured VIO

Here, the unknown constants are the b_j. In essence, regression is a curve-fitting technique: Unknown constants are estimated by using the historical data to find values that minimize the total squared error, where $\varepsilon_i = y_i - \hat{y}_i$. The quality of a regression model can be evaluated by R^2, which is the fraction of the response variability that is accounted for by the regression model.

To compare NIMF and regression, we need to construct models using both approaches. A NIMF model is a monotone relationship; we use MR_1. For regression, the choice of model is more difficult since we must specify an algebraic relationship for an unknown g. Our approach is to approximate g by an n-degree polynomial. We choose n by considering polynomials of increasing degree until there is no improvement in R^2. Equation (5) is a first degree polynomial. Below are second and third degree polynomials.

$$\hat{y}_i = b_0' + b_1' L_i + b_2' V_i + b_3' L_i^2 + b_4' L_i V_i + b_5' V_i^2 \tag{6}$$

$$\begin{aligned}\hat{y}_i = {} & b_0'' + b_1'' L_i + b_2'' V_i + b_3'' L_i^2 + b_4'' L_i V_i + b_5'' V_i^2 \\ & + b_6'' L_i^3 + b_7'' L_i^2 V_i + b_8'' L_i V_i^2 + b_9'' V_i^3\end{aligned} \tag{7}$$

For the data in Figure 2, the R^2 for Eq. (5) is .26; for Eq. (6), .34; and for Eq. (7), .37. A fourth degree polynomial showed no increase in R^2; so we use Eq. (7).

Figure 5 plots 75% prediction intervals for the test data, both for regression and for NIMF[2]. The plots show the measured value of LOSTEALRAT for each test-data instance (depicted by a dot) and the associated prediction interval (indicated by a vertical line with a horizontal bar at each end). Both techniques achieve adequate coverage: 94% for regression and 83% for NIMF. However, the average width of NIMF prediction intervals (.58) is less than half that of the regression prediction intervals (1.25). Also, in several instances the regression prediction interval includes negative values, which is impossible for LOSTEALRAT (a rate). In contrast, NIMF prediction intervals are constrained to lie within the measured data; so NIMF predicts only non-negative values for LOSTEALRAT.

The foregoing is one of eighteen case studies in which we compared NIMF to regression using measurements of VM/SP computer systems [Hellerstein, 1987]. The results of the other studies parallel those contained in Figure 5: In all cases adequate coverage is provided by both techniques, but NIMF consistently (17 out of 18 case studies) produces smaller prediction intervals.

Why does NIMF produce smaller prediction intervals? One reason is that regression assumes a specific algebraic relationship between the response and explanatory variables. If the wrong equation is chosen, then the fit is poor and prediction intervals are large. This shortcoming can, in part, be avoided by using other curve fitting techniques (e.g., cubic splines), which consider families of curves. However, these techniques still implicitly assume algebraic relationships, and are complex to apply to multivariate data.

Another reason for NIMF producing smaller prediction intervals is that it makes weaker assumptions about the error terms. Regression assumes that errors are realizations of independent and identically distributed (*iid*) normal random variables; in contrast, NIMF assumes only that errors are iid. Thus, for data that differ significantly from a normal distribution, regression is less effective than NIMF. A common approach to reducing the effect of non-normally distributed data is to transform the response variable, such as taking logarithms or square roots. We have considered a variety of transformations of the data in Figure 2; none resulted in significantly smaller prediction intervals for the regression model.

A final reason for NIMF's producing smaller prediction intervals is its being a local technique. That is, NIMF's assumptions of g being monotone and having iid error terms need only hold locally nearby the point at which a prediction is made. In contrast, regression is a global technique; error terms must be drawn from the same distribution, regardless of the point at which the prediction is made. For the data in Figure 2, the assumption of globally iid error terms is suspect. As with the assumption of normally distributed error terms, compliance with the assumption of globally iid error terms can be improved by transforming the response variable; however, we have not found a transformation that works well for our data.

4. Explanations

There are two reasons for explaining predictions. The first is to aid in model diagnosis, such as determining why there are missing bounds and recommending appropriate corrective actions. The second motive for explaining predictions is to provide an intuitive justification for a quantitative result, thereby adding to its credibility.

The purpose of model diagnosis is to identify violated assumptions and/or large sources of variability, and then to propose approaches to minimize these effects. NIMF makes only two assumptions: g is locally monotone with known directional effects (e.g., LOSTEALRAT increases with VIO and LOGGED), and error terms are locally iid. A full-fledged expert system for

---

2 We use $M = 20$.

statistical analysis would systematically examine these assumptions, propose changes in the model in order to improve compliance with the assumptions, and identify any deficiencies in the data (e.g., lack of independence). Here, we focus on the problem of explaining missing bounds.

Our approach to explaining missing bounds is computationally simple; it involves little more than print statements. This simplicity is possible because the knowledge representation used for prediction -- monotone relationships -- is the same as the knowledge representation used for explanation. There are three cases:

- Case 1: M is too small to achieve the specified confidence level.
- Case 2: The potential-bounds set is too small for the specified confidence level $(1-\alpha)$.
- Case 3: The data nearby $\mathbf{x}^*$ are inconsistent with the monotone relationship.

We use examples to illustrate how to generate explanations for all three cases. For case 1, we note that for a confidence level of $1-\alpha$ there is a minimum value of N_L (and N_H) required to obtain y_L (y_H). Specifically, from Eq. (4), we know that

$$1-\frac{\alpha}{2} \le \sum_{n=1}^{N_L} \binom{N_L}{n} .5^{N_L} = 1-.5^{N_L}$$

So,

$$N_L \ge \frac{\log \frac{\alpha}{2}}{\log .5}. \tag{8}$$

For example, if $1-\alpha = 99\%$ and $M = 6$, no bounds can be produced since the minimum M required is 8. Such a situation can be explained as follows:

```
Problem: No bounds possible when the confidence level is 99% and M = 6.

Reason: M is too small.

Recommendations: Either increase M to at least 8 or decrease the
     confidence level to no more than 97%.
```

The second case also relates to potential-bounds sets being too small, but the reason is different: too little data. Suppose that S_L is too small when $\mathbf{x}^* = (\text{VIO}, \text{LOGGED}) = (700,400)$, which can be detected in step 2 of the NIMF algorithm by using Eq. (8). Below is a sample explanation.

```
Problem:  No lower bound for LOSTEALRAT when VIO=700, LOGGED=400.

Reason:  Insufficient data in the range VIO ≤ 700, LOGGED ≤ 400.

Recommendation:  Collect more data or decrease the confidence level.
```

The foregoing explanation informs the analyst that the problem is *not* the violation of a model assumption; rather, the analyst needs either to acquire additional data (which poses little problem in domains such as computer performance where data are plentiful) or to reduce the confidence level.

Case 3 addresses problems with the monotone relationship used in the prediction model. Such a problem is detected in step 3 of the NIMF algorithm, when it is discovered that $y_L > y_H$. A sample explanation follows:

```
Problem:  No bounds for LOSTEALRAT when VIO=520, LOGGED=400.

Reason:  The data nearby (520,400) are inconsistent with the monotone
      relationship that LOSTEALRAT increases with VIO and LOGGED.

Recommendation:  Consider a different monotone relationship.
```

Here, the analyst learns that his/her understanding of the variable relationships is inconsistent with the data collected. Such situations often form the basis for new insights. Indeed, an area of future research is to develop an expert system that uses these insights and interacts with the analyst to propose new monotone relationships that reflect more accurately variable relationships evidenced in the data, while still being consistent with the analyst's intuition.

Next, we consider how to justify predictions when both bounds are present. Our approach is based on the observation that performance analysts have confidence in monotone relationships and data. A sample explanation follows.

```
The prediction interval for VIO=500, LOGGED=500 is
.1 ≤ LOSTEALRAT ≤ 1.5 because:
```

- It is assumed that LOSTEALRAT increases with VIO and LOGGED
- LOSTEALRAT = .1 when VIO=400, LOGGED=400
- LOSTEALRAT = 1.5 when VIO=625, LOGGED=510

While simple, the foregoing fails to explain why the bounds ensure a $1-\alpha$ confidence level. Producing such an explanation without undue complexity is an open research topic.

5. Conclusions

Frequently, we require quantitative predictions for systems in which numerical data are available but the following situation exists:

- There is no known algebraic characterization for the system.
- The system can be characterized easily in terms of monotone relationships.

One could obtain quantitative predictions by approximating the unknown algebraic relationship by a simple function, such as a polynomial. This paper presents an alternative approach: generating quantitative predictions directly from monotone relationships.

Our approach, non-parametric interpolative-prediction for monotone functions (NIMF), is statistical, and hence assumes the presence of historical data (which is reasonable for domains such as computer performance, financial analysis, and demographic studies). NIMF constructs prediction intervals by using monotone relationships to search historical data for prediction-interval end-points that provide a desired level of statistical confidence. Specifically, monotone relationships are used to impose a partial order on the historical data and thereby extract a set of potential lower bounds (S_L) and a set of potential upper bounds (S_H). A simple technique based on non-parametric statistics is then employed to select the prediction-interval lower-bound from S_L and the prediction-interval upper-bound from S_H.

Do we obtain better predictions by using an accurate monotone relationship instead of an approximate algebraic relationship? Although the answer depends on many factors (e.g., the system being studied and the approximation used), our experience with predicting low-memory contention in VM/SP suggests that using an accurate monotone relationship with the NIMF procedure can produce significantly better predictions than using a polynomial approximation of the unknown algebraic relationship and employing least-squares regression. Admittedly, NIMF's superior results are not solely a consequence of using monotone relationships instead of algebraic relationships, since NIMF also makes weaker assumptions about the distribution of error terms. However, avoiding unnecessary assumptions about algebraic relationships is clearly an advantage

in terms of predictive accuracy. Also, using monotone relationships simplifies model building and greatly facilitates explaining predictions.

NIMF has been implemented in APL and Prolog; the results presented here are from the Prolog implementation. Prolog is a particularly good implementation language for NIMF since monotone relationships are easily expressed as facts, and simple predicates can be used to find the sets of potential bounds.

References

Chidanand Apte and Se June Hong. (1986) Using Qualitative Reasoning to Understand Financial Arithmetic. *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 942-948.

James Bradley. (1968) *Distribution-Free Statistical Tests*. Prentice-Hall.

Johan De Kleer and John Seely Brown. (1984) A Qualitative Physics Based on Confluences. *Artificial Intelligence*, **24**, 7-83.

J. M. Dickson and M. Talbot. (1986) Statistical Validation and Expert Systems. *COMPSTAT. Proceedings in Computational Statistics.*, pages 282-288.

N. R. Draper and H. Smith. (1968) *Applied Regression Analysis*. John Wiley & Sons.

Daniel Dvorak and Benjamin Kuipers. (1989) Model-Based Monitoring of Dynamic Systems. *Proceedings of the Eleventh International Joint*, pages 1238-1243.

Daniel Dvorak and Elisha P. Sacks. (1989) Stochastic Analysis of Qualitative Dynamics. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1187-1192.

Kenneth D. Forbus. (1984)
Qualitative Process Theory. *Artificial Intelligence*, **24**, 85-168.

Hans W. Gottinger. (1988) Statistical Expert Systems. *Expert Systems (UK)*, **5**, 186-196.

Gerald J. Hahn. (1985) More Intelligent Statistical Software and Statistical Expert Systems: Future Directions. *The American Statistician*, **39**, 1-14.

Joseph Hellerstein. (1987) An Intuitive Approach to Performance Prediction with Application to Workload Management in VS SP/HPO. *Proceedings of the Computer Measurement Group*, **December**, 53-63.

P. Hietala. (1986) How to Assist an Inexperienced User in the Preliminary Analysis of Time Series: First Version of the Estes Expert System. *COMPSTAT. Proceedings in Computational Statistics.*, pages 295-300.

Joseph Hellerstein. (1989) A Non-parametric Approach for Constructing Prediction Intervals for Monotone Functions. IBM Corp. Research Report No. 15129, November.

IBM. (1985) *Virtual Machine Monitor Analysis Program: User's Guide and Reference*. IBM Corporation, (SC34-2166).

Leonard Kleinrock. (1975) *Queueing Systems, Volume 1*. John Wiley.

Donald W. Kosy and Ben P. Wise. (1984) Self-Explanatory Financial Planning Models. *Proceedings of the National Conference on Artificial Intelligence*, pages 176-181.

Benjamin Kuipers. (1984) Commonsense Reasoning about Causality: Deriving Behavior from Structure. *Artificial Intelligence*, **24**, 169-203.

Benjamin Kuipers. (1986) Qualitative Simulation. *Artificial Intelligence*, **29**, 289-338.

Reid Simmons. (1986) Commonsense Arithmetic Reasoning. *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 118-124.

Michael P. Wellman. (1987) Probabilistic Semantics for Qualitative Influences. *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 660-664.

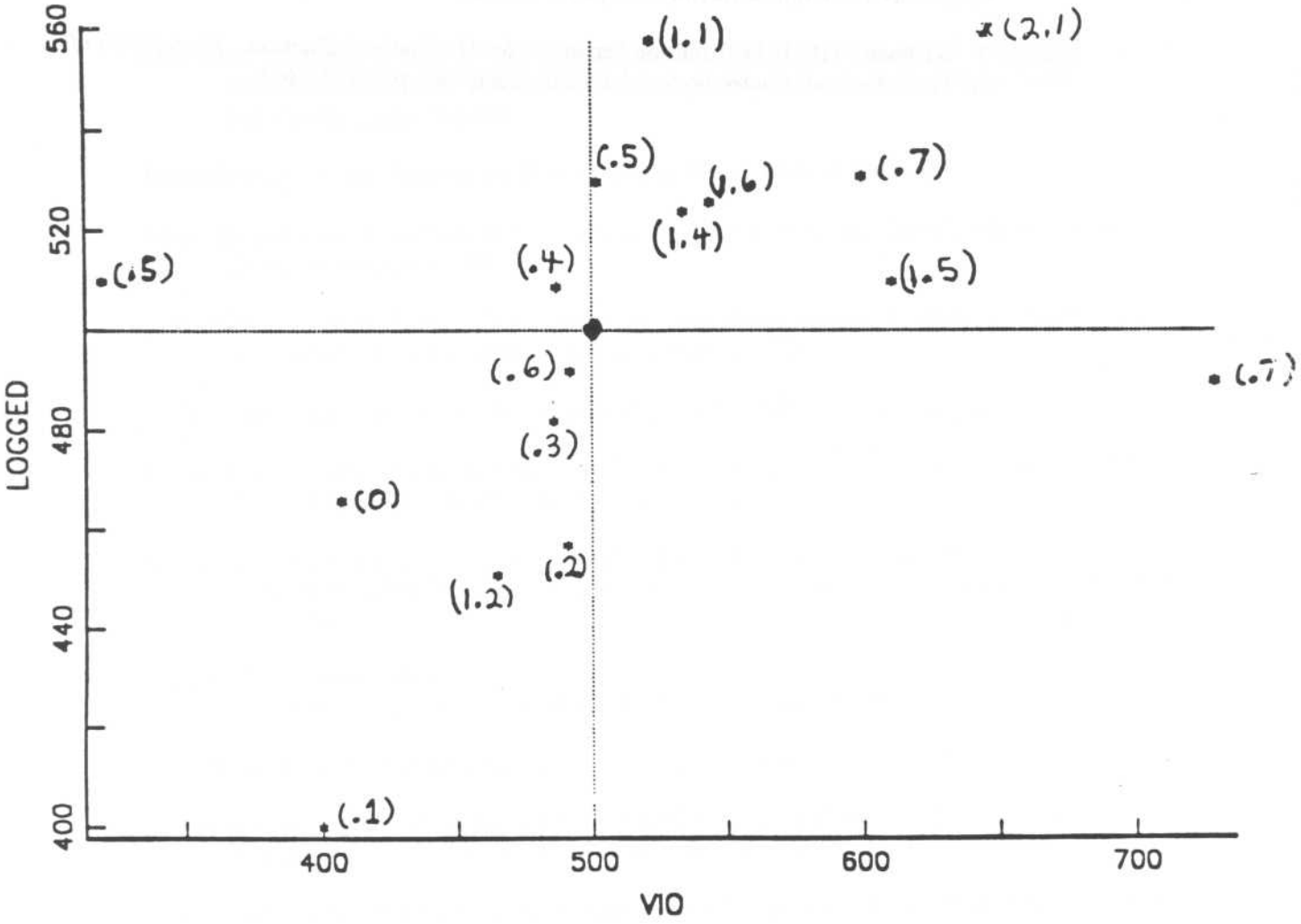


Figure 1. Data used in illustrative example. (Numbers in parentheses are LOSTEALRAT values.)

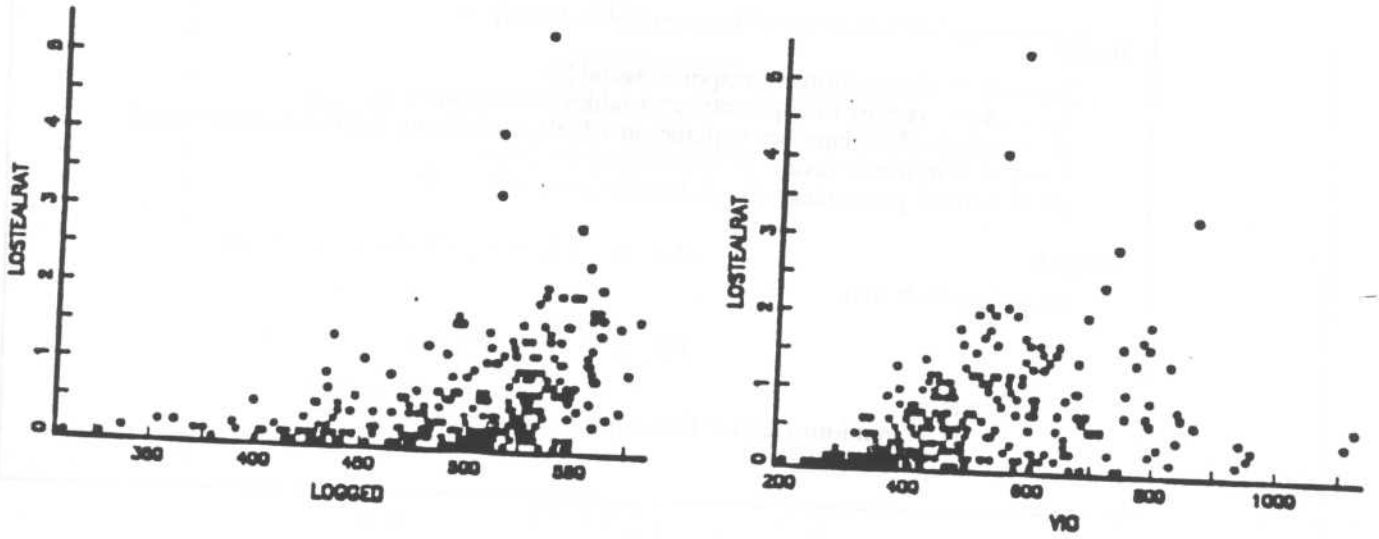


Figure 2. Scatter plots of VM/SP data

Model

$$y_i = g(\mathbf{x}_i) + \varepsilon_i$$

Where:

g = monotone function with known directional effects
ε_i = error terms (realizations of independent and identically distributed random variables)

Inputs

$y_1, \ldots, y_I$ = observations of response variables
$\mathbf{x}_1, \ldots, \mathbf{x}_I$ = vectors of explanatory variables for responses $y_1, \ldots, y_I$
$\mathbf{x}^*$ = values of explanatory variables at which a prediction interval is constructed
$1 - \alpha$ = confidence level
M = control parameter

Outputs

y_L and y_H such that

$$P(y_L \leq Y^* \leq y_H) \geq 1 - \alpha$$

Where:

Y^* = random variable for responses at $\mathbf{x}^*$

Figure 3. NIMF Model, Inputs, and Outputs

Step 1: Compute the potential-bounds sets

$$S_L = \{y_i \mid \mathbf{x}_i \in S_L^{\mathbf{x}} \text{ with one of the } M \text{ smallest } d(\mathbf{x}_{i'}, \mathbf{x}^*) \text{ for } \mathbf{x}_{i'} \in S_L^{\mathbf{x}}\}$$

$$S_H = \{y_i \mid \mathbf{x}_i \in S_H^{\mathbf{x}} \text{ with one of the } M \text{ smallest } d(\mathbf{x}_{i'}, \mathbf{x}^*) \text{ for } \mathbf{x}_{i'} \in S_H^{\mathbf{x}}\}$$

Where:

$S_L^{\mathbf{x}} = \{\mathbf{x}_i \mid \mathbf{x}_i \prec \mathbf{x}^*\}$
$S_H^{\mathbf{x}} = \{\mathbf{x}_i \mid \mathbf{x}^* \prec \mathbf{x}_i\}$
$\prec$ = partial order imposed by monotonicity assumption

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{J} \frac{(x_{1j} - x_{2j})^2}{s_j^2}}$$

s_j = standard deviation of x_j
M = maximum size of S_L, S_H

Step 2: Compute the index of the bounds

$$k_L = \phi(N_L, \alpha)$$

$$k_H = N_H - \phi(N_H, \alpha) + 1$$

Where:

N_L = size of S_L
N_H = size of S_H

$$\phi(N, \alpha) = \max\left\{k \mid \sum_{n=k}^{N} \binom{N}{n} .5^N \geq 1 - \frac{\alpha}{2}\right\}$$

Step 3: Find the bounds

$$y_L = \begin{cases} k_L\text{-th smallest } y_i \in S_L & \text{if } Q_L \\ -\infty & \text{otherwise} \end{cases}$$

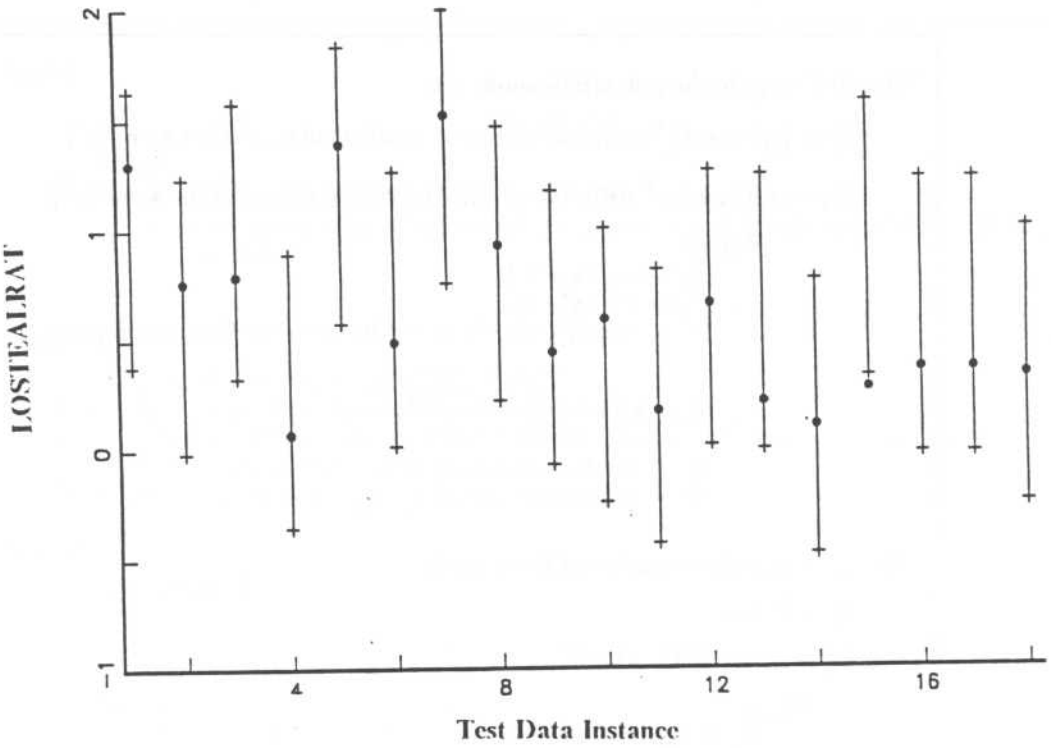
$$y_H = \begin{cases} k_H\text{-th smallest } y_i \in S_H & \text{if } Q_H \\ \infty & \text{otherwise} \end{cases}$$

Where:

$Q_L = (k_L \geq 1) \wedge (y_L \leq y_H)$
$Q_H = (k_H \leq N_H) \wedge (y_L \leq y_H)$

Figure 4. Summary of NIMF Procedure

Regression: coverage = 94%, average width = 1.25



NIMF: coverage = 83%, average width = .58

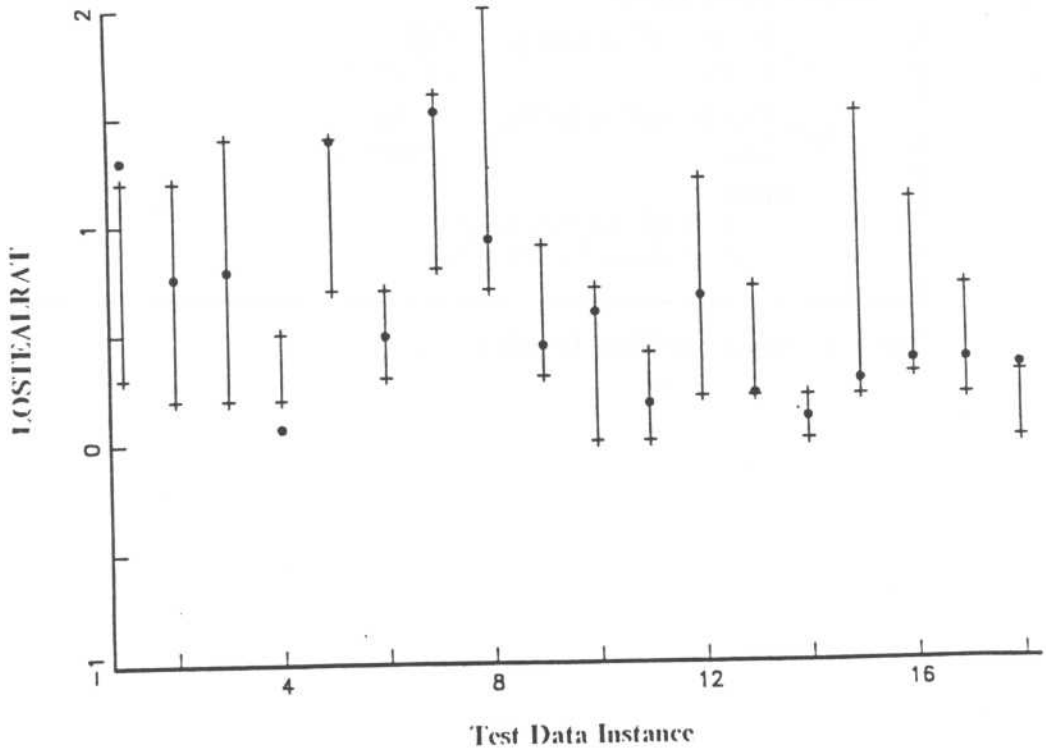


Figure 5. 75% Prediction intervals