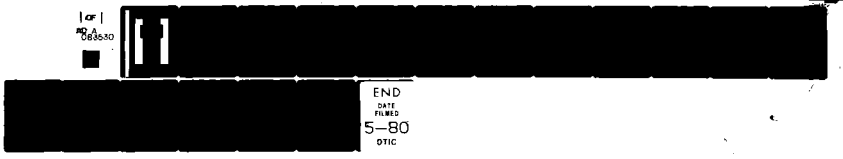
AD-A083 830 CLEMSON UNIV S C DEPT OF MATHEMATICAL SCIENCES F/G 12/1
MAHALANOBIS DISTANCE AS A MEASURE OF ANALOG IN PARAMETRIC COST --ETC(U)
SEP 78 K T WALLENIUS
UNCLASSIFIED N98 NL



END
DATE FILMED
5-80
DTIC

(10) LEVEL II

AD A083830

DEPARTMENT OF MATHEMATICAL SCIENCES

CLEMSON UNIVERSITY
Clemson, South Carolina

(10) LEVEL

MAHALANOBIS DISTANCE AS A MEASURE OF ANALOG IN PARAMETRIC COST ESTIMATION

By

K. T. Wallenius

Clemson University

Report No. N98

Technical Report #522

September, 1978

Research supported by The Office of Naval Research
Contract N00014-75-C-0451

I. BACKGROUND[1]

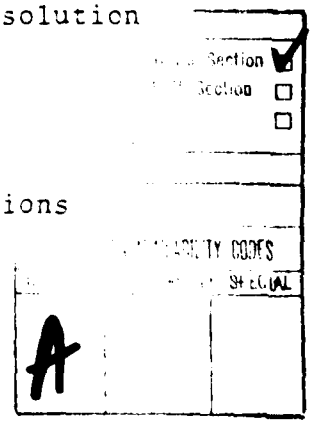
Traditionally, weapon system cost estimates have been prepared using Industrial Engineering (I.E.) techniques. These techniques involved detailed studies of the operations and materials required to produce the new system. The cost estimate frequently required several thousand hours to produce with volumes of supporting documentation. Changes in design require extensive changes in these estimates. In spite of all the time and effort involved in preparing these estimates, their accuracy leaves much to be desired. This is evidenced by the large cost overuns cited by the annual General Accounting Office (GAO) reports to Congress. In 1972, for example, the GAO reported that the Navy had experienced a cost growth of $19 billion on 24 weapon systems in FY 1971. Approximately 15% of this cost growth was attributed to poor initial cost estimates for the weapon systems. The report went on to make the following recommendation:

> "Develop and implement DOD wide guidance for consistent and effective cost estimating procedures and practices particularly with regard to, ... an effective independent review of cost estimates."

Three months prior to the GAO recommendation, Deputy Secretary of Defense David Packard suggested the use of Independent Parametric Cost Estimation, (IPCE), as a possible solution to poor initial cost estimates. In a memorandum dated

---

[1]Parts of this section are nearly verbatim extractions from sections I and II of reference (10).



December 7, 1971 to the Service Secretaries, Mr. Packard stated:

> "Parametric cost estimates available in 1964 on the F-111A and in 1965 on the C-5A came within 20 percent of the actual costs currently being experienced."

Mr. Packard then directed each of the Service Secretaries to:

1. Improve their capability to perform independent parametric cost analysis.
2. Have such analyses done on each major weapon system at each key decision point in the weapon system acquisition process.
3. Make the analysis available directly to the Defense System Acquisition Review Council (DSARC) at each DSARC review starting January 1, 1972.

Secretary of Defense Melvin Laird, on January 25, 1972, issued a memorandum supporting the Packard memo and established a high level DOD organization (CAIG: Cost Analysis Improvement Group) to review IPCE's and report on their soundness to the DSARC.

Because of the rigidity and poor performance of the traditional approach, some early successes using the parametric approach, and the cited high level directives, independent parametric cost estimation has been receiving considerable attention in the Department of Defense as a means of increasing the accuracy of cost estimates. This procedure is based on the premise that the cost of a weapon system is related in a quantifiable way to the system's physical and performance characteristics.[2]

---

[2]These characteristics are referred to as system "parameters" in the cost estimation literature and should not be confused with statistical parameters (e.g., standard deviations, regression coefficients, etc.). Statisticians would refer to system "parameters" as predictor (independent) variables and "cost" as the criterion (dependent) variable. They would also refer to the goal as cost prediction instead of estimation.

A Parametric Cost Estimate has been defined by Baker (1) as:

> "An estimate which predicts cost by means of explanatory variables such as performance characteristics, physical characteristics, and characteristics relevant to the development process, as derived from experience on logically related systems."

The construction and use of cost estimating relationships, (CER), forms the foundation for making IPCE's. Cost estimating relationships are mathematical equations which relate system costs to various explanatory variables. They are most generally derived through statistical regression analysis of historical cost data. These techniques are described in (9). Some examples of their use appear in (2), (4), (5), and (11).

The parametric approach has some distinct advantages and disadvantages compared to I.E. methodology. On the plus side are:

1. Parametric cost estimates can be developed during the concept formulation stage of the acquisition process before detailed engineering plans are available. These early cost estimates can be used to:

   (a) Identify possible cost/performance tradeoffs in the design effort.

   (b) Provide a basis for cost/effectiveness review of performance specifications.

   (c) Provide information useful in the ranking of competing alternatives.

   (d) Suggest a need for identifying and considering new alternatives.

2. Historical cost data incorporates system development setbacks such as engineering and design specification changes and other items that are not identifiable at the time of design. Industrial engineering (I.E.) estimates tend to be optimistic in that they don't allow for unforeseen problems. Unexpected engineering or design changes usually bring about unexpected increases in system cost. Cost estimating relationships based on historical data will incorporate some of these unknowns into the cost estimate.

Possible problems with the parametric approach include:

1. Unlike the IE approach, many subjective assessments and decisions must be made by the analyst including:

   (a) Selecting the "analogous" systems to include in the historical data base.

   (b) Selecting the form of the prediction equation.

   (c) Selecting an appropriate subset of performance/design characteristics to include in the final prediction equation.

These decisions can lead to conflicting estimates by different analysts even when sound statistical practices are employed. There is no universally "best" approach to the selection problems stated above. Subjectivity cannot be avoided but can be incorporated in a consistent manner using the Bayesian approach to prediction as discussed by Lindley (6).

2. Historical data sets are often characterized by sample sizes which are relatively small compared to the number of potential predictor variables. This often leads to an overstatement concerning the degree of fit supposedly obtained. Discussion of this problem can be found in (3) and (12).

The phrase "logically related system" in the cited definition of parametric cost estimation is subject to all kinds of interpretation and degrees of relation. Certainly there is no historical system identical in all respects to the object system (the system whose cost we wish to predict) else the problem would not exist. At the other extreme, all military systems

are "logically related" in (at least) the sense that they are military systems. Message carrying pidgeons, air-to-air missiles, jet aircraft and frizbees are "logically related" in that they all fly. Obviously, the analyst must take into account the degree of analogy between each system (which is a candidate for the historical data base) and the objective system. Analogy, according to Webster, is "a partial similarity between like features of two things on which a comparison may be based." How does one measure the degree of analogy between "logically related" systems and how can one exploit these partial similarities in predicting the cost of an objective system?

In what follows, we propose Mahalanobis distance as a measure of analogy and discuss its implication in the processes of selecting (potential) members of the data base and tailoring a CER to a specific objective system. This is a distinct departure from standard procedures recommended (9), (10) and used in developing every CER with which the writer is familiar. The distinction is fundamental and goes beyond measures of analogy. The standard approach appears more oriented toward developing a cost explaining equation relating costs of a class (e.g., sonars, airframes, tanks, etc.) of historical systems to the characteristics of those systems. One need not have any specific objective system in mind while developing such a general purpose descriptive equation. In fact, armed with an airframe CER based on the explanatory variable "weight", two radically different airframes of the same weight would be estimated to cost the same amount. Mallows (8) defined six potential uses

of a regression equation which include (a) pure description and (b) prediction. Lindley (6) emphasizes that the technique used to develop a regression equation ought to be related to the intended use. In the present context, the intended use is prediction of the cost of a specific system so that using a CER (which was developed to describe historical relations without reference to any specific objective system) to predict cost (of a specific system) is contrary to Lindley's recommendation and common sense.

II. MEASURES OF ANALOGY

Having gathered and adjusted historical data on systems judged more-or-less analogous to a proposed system whose cost is to be estimated, the analyst proceeds with the task of developing a "best" Cost Estimation Relation (CER). This involves selecting the form of the CER, deciding which of the system variables (performance characteristics, design specifications, etc.) to include as predictor variables, and assessing the precision of the estimate. In parametric cost estimation, this is usually done through the use of multiple regression and some standard variable selection criterion such as maximizing adjusted R squared (minimizing mean square error [MSE]), maximizing F, using Mallow's C_p, etc.

All of these techniques share two properties: (1) For any fixed number of variables in the prediction equation, the optimal set of variables is that set which minimizes the MSE. (2) They all ignore the values of the variables of the system whose cost is being estimated. The first of these properties is reasonable but myoptic when the object is prediction. The second property seems contrary to common sense.

Suppose there are n systems in the historical data base. Associated with the i^{th} such system is a cost Y_i and values of p (candidate) predictor variables $X_{i1},\ldots,X_{ij},\ldots,X_{ip}$. Let $\underset{\sim}{Y}$ denote the vector of costs and $\underset{\sim}{X}_j$ the vector of characteristic j values:

$$\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} \qquad \underset{\sim}{X}_j = \begin{pmatrix} X_{1j} \\ \vdots \\ X_{ij} \\ \vdots \\ X_{nj} \end{pmatrix}$$

Furthermore, let $\bar{X}_j = \frac{1}{n} \sum_{k=1}^{n} X_{kj}$ and $s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)$ denote the sample means and covariances. Denoting the values of the proposed system by lower case letters, we wish to predict the cost y by exploiting the predictive ability of the characteristics $x_1, \ldots, x_j, \ldots, x_p$. This predictive ability is inferred from the apparent relation between historical costs and characteristics and the degree of analogy between the proposed system and these historical data. How analogous is the proposed system to the historical data?

(a) Marginal comparisons: Analogy on a single dimension is straightforward. One could refer x_j to a histogram of X_{ij} values, $i = 1,2,\ldots,n$ as in

Figure 1

A statistic commonly used as a nonnegative distance index is simply the square of the standardized distance between x_j and the mean of the X_{ij}'s, namely

$$M_j = \left(\frac{x_j - \bar{X}_j}{s_j}\right)^2$$

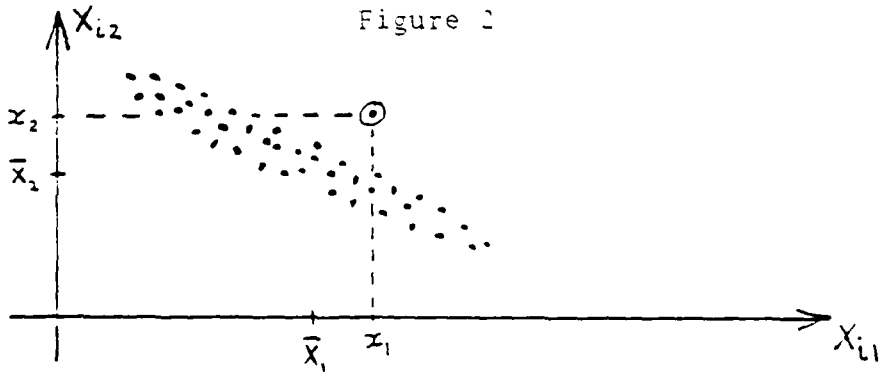
where s_j is defined as $(s_{jj})^{1/2}$. Large values of this statistic indicate a low degree of analogy.

(b) High dimensional comparisons: The collection of marginal indices $\{M_1, M_2, \ldots, M_p\}$ can give a very misleading impression of the overall degree of analogy. Even when M_j is small for every j, the proposed system can be terribly nonanalogous to the historical data. A simple bivariate example will illustrate this assertion. Suppose X_1 and X_2 denote weight and maneuverability, respectively and that x_1 and x_2 are each within one standard deviation of their respective means, i.e.,

$$M_1 = M_2 \leq 1.$$

Suppose further that, historically, heavy systems tended to be less maneuverable i.e., $s_{x_1x_2} < 0$, but the proposed system is a little heavier and slightly more maneuverable than the average. The situation is depicted in

Figure 2



We see that (x_1,x_2) is marginally analogous to the historical data on both weight and maneuverability but not at all analogous when viewed in two dimensions. Comparing (x_1,x_2) to $(\bar{X}_1,\bar{X}_2)$ marginally ignores important relational information.

A measure of analogy which incorporates relational information was suggested in 1930 by P.C. Mahalanobis (7). He proposed

$$M = (\underset{\sim}{\mu}^{(1)}-\underset{\sim}{\mu}^{(2)})' \Sigma^{-1}(\underset{\sim}{\mu}^{(1)}-\underset{\sim}{\mu}^{(2)})$$

as a measure of the distance between two multivariate populations with mean vectors $\underset{\sim}{\mu}^{(1)}$ and $\underset{\sim}{\mu}^{(2)}$, respectively, and common covariance matrix Σ. Replacing the parameter values by estimates, we obtain (in our notation)

$$\hat{M} = (\underset{\sim}{x}-\bar{\underset{\sim}{X}})' S^{-1}(\underset{\sim}{x}-\bar{\underset{\sim}{X}})$$

where $S = (s_{ij})$. Except for a multiplicative constant, this is Hotelling's T^2 statistic used to test that $\underset{\sim}{x}$ and the historical data came from the same population. In the previous bivariate example, it is easy to show that

$$\hat{M} = \frac{1}{1-\hat{\delta}^2}\left[M_1 - 2\hat{\delta}(M_1M_2)^{1/2} + M_2\right],$$

which can be <u>arbitrarily</u> large even when M_1 and M_2 are small. For example, with $M_1 = M_2 = \varepsilon$,

$$\hat{M} = \frac{2\varepsilon}{1+\hat{\delta}} \quad \text{and} \quad \lim_{\hat{\delta}\to 1} \hat{M} = \infty.$$

III. THE ROLE OF ANALOGY IN PREDICTION

As mentioned in the previous section, most standard variable selection techniques share the property that, for any given number of variables in the regression function, the optimal set is that set which minimizes residual mean square error (or, equivalently, maximizes r^2). The objective system may be rather nonanalogous to the historical data (large $\dot{M}$) when we consider the subset of variables identified as "optimal" by the criteria used to develope the CER. Often, there are several k-variable models which come close to the "optimum" in terms of r^2 and other measures of model aptness based on residual analysis. In these cases, by using a slightly suboptimal set of prediction variables (slight decrease in r^2) it may be possible to substantially improve the degree of analogy (decrease in $\dot{M}$). What is the role of analogy in prediction and how can one evaluate the tradeoff of fit for analogy?

The width of the prediction interval at the point corresponding to the objective system is a numeraire which seems like a reasonable basis for choosing between alternative models. We shall consider a monotone function of the width for simplicity, namely, the square of the half-width, viz.

$$W = F_{1-\frac{\alpha}{2};1,n-k-1} * MSE * (\dot{M} + \frac{n+1}{n}) ,$$

where $F_{1-\frac{\alpha}{2},1,n-k-1}$ is the $(1-\frac{\alpha}{2})^{th}$ fractile of an F distribution

with 1 and n-k-1 degrees of freedom. This measure W combines "fit" (MSE) and "degree of analogy" ($\hat{M}$) with a factor F which penalizes for using too many variables (increasing k) or excluding points from the data base (decreasing n). In this form, the role of analogy, as measured by Mahalanobis distance, is evident. It enters as a term in the multiplier $(\hat{M} + \frac{n+1}{n})$ of MSE. Failure to consider this factor in selecting a CER could have a marked effect on predictor precision as measured by prediction interval width.

IV. SUMMARY

We have pointed out the difficulty of recognizing the degree of analogy in high dimensional spaces. We have suggested Mahalanobis distance as a measure of analog and pointed to its role in prediction precision.

Reference (10) suggests a 14 step precedure for developing a parametric cost estimate. The importance of understanding the system's technical aspects is stressed in steps 1-3 prior to collecting (step 4) and adjusting (step 5) the data. Somewhere prior to building (step 8) and evaluating (step 9) the CER, we recommend the analyst "let the data speak for itself". Included in such a "data exploration" ought to be considerations of multivariate degrees of analogy. Our contention is

(1) Important and subtle relations among the systems and variables may be overlooked when viewed from a purely technical approach.

(2) Mahalanobis distance is an appropriate measure of analogy which can shed light on these relations.

The analyst should be open-minded (but skeptical) about relations which seem to be suggested. Let the data suggest whatever it will. Relations cannot be viewed with a critical eye if they are not viewed in the first place. If what the data seem to be saying is inconsistent with the analyst's technical understanding, the source of the contradiction deserves close attention.

In subsequent papers we will develop algorithms for building models based on minimizing W and will compare models so obtained with models judged optimal by other criteria.

REFERENCES

(1) Baker, B. N., Improving Cost Estimating and Analysis in DOD and NASA, Ph.D. Thesis, George Washington University, Washington, D.C., 1972.

(2) Chief of Naval Operations, OP-96D, Independent Cost Review, S-3A Program, 1972.

(3) Edwards, Thomas B., On the Degree of Inflation of Measures of Fit Induced by Empirical Model Building, Ph.D. Thesis, Clemson University, 1979.

(4) Large, J. P., Estimating Aircraft Turbine Engine Costs, Rand Corp. Report RM-6384/1-PR, Sept. 1970.

(5) Levenson, G. S., et al., Cost Estimating Relationships for Aircraft Airframes, Rand Corp. Report R-761-PR, December 1971.

(6) Lindley, D. V., The Choice of Variables in Multiple Regression, J. R. Statis. Soc. (30), 1968.

(7) Mahalanobis, P. C., On Tests and Measures of Group Divergence, J. Asiat. Soc. Beng., (26), 1930.

(8) Mallows, C. L., Data Analysis in a Regression Context, Proceedings of Univ. of Kentucky Conference on Regression with a Large Number of Predictor Variables, Dept. of Statistics, Univ. of Kentucky, Lexington, Ky.

(9) Military Equipment Cost Analysis, RAND, prepared for the Office of Secretary of Defense (Systems Analysis), for Official use only, 1971.

(10) Miller, Bruce M. and Soverign, Michael G., Parametric Cost Estimating with Applications to Sonar Technology Naval Postgraduate School Technical Report NPS 55Z073091A, Sept. 1973.

(11) Rondinelli, L. A., An Analysis of Cost Versus Performance Relationships for Phased Array Radars, RAND Corp., RM-5380-PR, 1967.

(12) Wallenius, K. T., Regression Analysis in Parametric Costing and Pricing: Pitfalls, Problems and Potentials, Proceedings, Fourth Annual DOD Procurement Research Symposium, 1975.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: TR 322	2. GOVT ACCESSION NO.: AD A083 530
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): Mahalanobis Distance as a Measure of Analog in Parametric Cost Estimation	5. TYPE OF REPORT & PERIOD COVERED:
	6. PERFORMING ORG. REPORT NUMBER: N98
7. AUTHOR(s): K. T. Wallenius	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Clemson University Dept. of Mathematical Sciences Clemson, South Carolina 29631	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 047-202 NR ~~042-271~~
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research Code ~~436~~ 434 Arlington, Va. 22217	12. REPORT DATE: September, 1978 13. NUMBER OF PAGES: 14
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Parametric Cost Estimation, Multiple Regression, Analogy, Mahalanobis Distrance, Prediction Slection of Variables

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Mahalanobis distance is suggested as a measure of analogy between a proposed system whose cost is to be predicted and a historical data base. It's role in the prediction problem and implications for selecting predictor variables is pointed out.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 68 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)