Inform. Systems Vol. 8, No. 2, pp. 105–115, 1983
Printed in Great Britain.

0306-4379/83 $3.00 + .00

ESTIMATING RECORD SELECTIVITIES

STAVROS CHRISTODOULAKIS
Computer Systems Research Group, University of Toronto, Toronto, Canada M5S 1A1

(*Received* 16 *April* 1982; *in revised form* 22 *July* 1982)

Abstract—In this paper we examine the problem of modelling data base contents and user requests. This modelling is necessary in analytic data base performance evaluation studies in order to estimate the number of records of a file that have to be retrieved in response to user(s) requests. The cpu, io, and telecommunication costs of the system are directly or indirectly expressed in terms of these quantities.

We first show that certain assumptions used for modelling data base contents, data placement on devices and user requests often are not satisfied in actual data base environments. Thereafter we provide more detailed modelling techniques based on a multivariate statistical model, and we demonstrate their use in improving data base performance.

1. INTRODUCTION

Analytic models that have been proposed for data base performance evaluation use an estimate of the number of records qualifying in a query (usually referred to as *record selectivity*), and the average number of records qualifying in a set of queries (*average record selectivity*). The estimation of selectivities, is based on models describing the data base contents, and the query distribution in the files. The existing models make various *assumptions* about the data base contents, and the query distribution in the files. Based on those assumptions and statistics kept by the system, they estimate the selectivities and thus the system cost[34, 33, 3].

The *uniformity assumption* has been used to model the distribution of values of a single attribute in its domain. Each attribute is assumed to have a domain of m distinct values. The number of records in a file that have a value v in the attribute A is calculated as N/m, where N is the number of records in the file. Using the uniformity assumption, the record selectivities of single attribute conditions can be approximated.

The *independence assumption* has been used to estimate record selectivities in queries where more than one attribute is involved. According to this assumption the attribute values of any two attributes A1 and A2 are independent. Thus, the number of records qualifying in a conjunctive condition on two attributes A1 and A2 is estimated as $N*s_1*s_2$ where s_1 and s_2 are the proportions of the file qualifying in the single attribute conditions on A1 and A2 respectively.

The *uniformity of queries* assumption has been used to estimate the number of records qualifying in a set of queries. The assumption states that the attribute values of the attributes appearing in user queries are uniformly distributed over the attribute value combinations.

Probability distributions

We adopt here the terminology of the relational model[13]. Let D be the cartesian product of a set of domains D_i, $i = 1, \ldots, m$. At any given point in time a relation R is a subset of D. The probability space and the probability distribution for R is defined as follows: D is the *Space* or *Universe*[23]. An *experiment* E is a selection of a subset of D. The *outcome* of an experiment (*event*) is a subset S of D. We assign a number $P(Q)$ to an event Q as follows: $P(Q) = k/N$, where k is the cardinality of the set $R \cap S$, and N is the cardinality of R. It is easily verified that $P(Q)$ satisfies the probability axioms. We call $P(Q)$ the probability of the event Q. Given D, E and P we associate an m-dimensional *random variable* X with an event Q. X is a function from the set of possible outcomes to m-tuples of real numbers. Thus an m-dimensional probability distribution function can be defined:

$$F_{\mathbf{x}}(\mathbf{x}) = P[\mathbf{X} \leq \mathbf{x}].$$

The probability distribution of a subset of the attributes of R is defined to be the marginal distribution of $F_{\mathbf{x}}(\mathbf{x})$ in the subspace of the m-dimensional probability space which is defined by the selected attributes.

The probability distribution of the attribute values appearing in user queries is defined as follows: Let R' be the set of queries of the same query type which refer a given subset of n attributes. R' is a subset of a set of queries which has been recorded in a given time period. For simplicity we assume here equality queries only. Let D' be the cartesian product of the domains of the n attributes. An event Q is a subset S of D'. The probability P' of an event is determined from R' and S as before. An n-dimensional random variable Y can be associated with an event Q. Thus the n-dimensional probability distribution function

$$F_{\mathbf{y}}(\mathbf{y}) = P'[\mathbf{Y} \leq \mathbf{y}]$$

of the attribute values which appear in a set of user queries is defined.

Next, we will give some examples to demonstrate that the assumptions of uniformity and independence of attribute values, and uniformity of queries are not satisfied in typical data base environments.

Uniformity assumption

The uniformity of attribute values assumption as it was described above implies that the probability distribution for each attribute (the marginal probability distribution in the subspace of the attribute) is uniform. This assumption may be unrealistic in many actual environments. Data bases often contain information describing populations of the real world. Examples of such populations are the employees of an organization, the students of a university, the people under security surveillance, the fish of an aquarium, or the cars of an autoshop. Such populations usually have only a few members with extreme characteristics. For example, only a few people are extremely tall or extremely short. The majority have a height near to an average height. Thus the values of some attributes of data bases describing such populations tend to have unimodal distributions rather than uniform.

As an example of a population, consider the population of professional engineers of Ontario. A recent report[19] provides various statistics such as years of experience, responsibility level, and salary information for this population. All the statistics present highly unimodal distribution of values rather than uniform. In Table 1 we reproduce the statistics given for the number of engineers per responsibility level. The data describing the professional engineers of Ontario is a typical example of the numerous large data bases containing information about populations. We will use the statistics included in the engineering report for examples in other sections of the paper.

As a second example, consider the population of teachers in Canadian Universities. A report by Statistics Canada[34] provides information about teachers in Canadian Universities. The attribute values of their attributes present highly non-uniform distributions. In Table 2, we reproduce the number of teachers per age interval. As a last example of non-uniformity of attribute values, the number of male teachers in Canadian Universities was 27799, while the number of female teachers was 4846.

Independence assumption

In many data base environments attribute values of certain attributes are *correlated*. *Correlations* are measures of the *covariance* of the attribute values of two attributes[26]. The existence of high correlations between attribute values in data bases is a direct consequence of the fact that files often contain information about populations of the real world. Consider two attributes X and Y, which we assume take on distinct attribute values. Let $f(x, y)$ be the two-dimensional density distribution of the two attributes X and Y. We can think of $f(x, y)$ as being a two-dimensional histogram. We say that the attribute values of the attributes X and Y are *independent* if $f(x, y) = p(x)q(y)$, where $p(x)$ and $q(y)$ are the density distributions of the attributes X and Y. We say that the attribute values of the attributes X and Y are *not-correlated* if $E(xy) - E(x)E(y) = 0$, where the symbol E denotes expectation for all the records in a file. The independence assumption is a strong one. It is strictly stronger than the assumption of no correlations[12].

Many attribute pairs describing populations are correlated. Consider again the engineering data base. In this data base, attribute pairs like (SALARY, YEARS-FROM-GRADUATION), (SALARY, RESPONSIBILITY-LEVEL), (YEARS-FROM-GRADUATION, RESPONSIBILITY-LEVEL), are highly correlated. In Table 3 we reproduce the relationship between years from degree and responsibility level for engineers in responsibility level "A", and for the total population of engineers. Note that there are only a few engineers in responsibility level A with many years of experience.

As a second example consider the population of teachers in Universities. The attributes (TITLE, HIGHEST-EARNED-DEGREE), (AGE, TITLE), (AGE, SALARY), (SALARY, TITLE), (SALARY, HIGHEST-EARNED-DEGREE), (SEX, TITLE), (NATIONALITY, COUNTRY - OF - HIGHEST - EARNED - DEGREE), are highly correlated. In Table 4 we reproduce the relationship between age group, sex and title, for assistant, associate and full professors. Note the strong correlations both between age and title, and between sex

Table 1.

Res-Level	#Engineers
F	1036
E	2853
D	4345
C	3240
B	1671
A	1355

Table 2.

Age	#Teachers
⇐25	13
25–29	1458
30–34	5074
35–39	7658
40–44	6213
45–49	4824
50–54	3477
55–59	2294
60-over	1526

Table 3.

YEAR-OF-GRADUATION	#ENG. IN LEVEL A	TOTAL #ENG.
49–72	19	3609
73	10	566
74	13	590
75	27	552
76	66	511
77	316	494
78	521	574
79	315	322

Table 4.

Age-group	Full-prof.		Assoc-Prof.		Assist-Prof.	
	M	F	M	F	M	F
≤ 25	0	0	0	0	2	1
25–29	0	1	27	3	548	151
30–34	37	0	994	102	2363	464
35–39	606	14	3568	336	1985	445
40–44	1678	48	2785	306	734	252
45–49	1960	82	1715	260	360	175
50–54	1815	95	868	215	174	123
55–59	1358	92	438	132	82	75
≥ 60	929	79	224	92	63	32

and title. In data base environments with highly correlated attributes the uniformity and independence assumptions may result to large errors in data base preformance evaluation[12].

Uniformity of queries assumption

The uniformity of attribute values in queries may not be satisfied in realistic data base environments. The reason is that certain attribute values may be of more interest to the users than others. For example in a statistical data base environment users may be more interested in rare cases instead of usual ones. Pezerro[27] presents a performance prediction study in a real data base environment where the uniformity of attribute values in queries assumption was found unrealistic.

It was formally shown in [12] that the above assumptions often lead to cost estimations that are pessimistic. It was also shown that the difference in the estimated cost can be large. Thus more detailed modelling of data base contents is desirable. In the subsequent sections we present a more detailed model which takes into account non-uniformity and correlations of attribute values. Then we discuss some applications of the improved approximations.

2. ESTIMATION OF RECORD SELECTIVITIES

In this section, we examine methods for providing better approximations of record selectivities in flat files. We regard the records of a flat file as points of an n-dimensional space, where n is the number of attributes of the file. Each attribute corresponds to an axis of the n-dimensional space. Attribute values of quantitative attributes (like SALARY, AGE, and HEIGHT) correspond directly to points of axis. Attribute values of categoric attributes (like RESPONSIBILITY-LEVEL, SEX, and DEGREE) are mapped into discrete values in the corresponding axis. The attribute values of a record of the file determine a point in this space. We hereafter refer to this space as the *attribute space.*

Typical data bases contain tens or hundreds of attributes, and some attribute domains contain a large number of attribute values. The record selectivity of each query would be completely specified if we kept information on the number of records for every combination of attribute values in the attribute space of the file. However, to keep this information, the amount of storage required is exponential in the number of attributes (v^{α}, where v is the number of distinct values for each attribute and α is the number of attributes). Since this is very expensive for a realistic environment, techniques that *approximate* the record selectivity become important.

The approach described in this paper is based on the approximation of the density distribution of points in the attribute space using a multivariate density distribution function. This density function is then used to derive estimates of record selectivities.

Traditionally *parametric* and *non-parametric* techniques have been used for describing the distribution of points in an n-dimensional space[4, 18, 38]. Parametric techniques assume that the density has a known distribution form and use the available samples to estimate the parameters of the distribution. Non-parametric techniques do not make an *a priori* assumption about the density distribution of the data, but use the available samples to determine the shape of the distribution.

In this section we describe a parametric technique for estimating record selectivities. Although non-parametric techniques could prove more accurate if a large number of parameters is kept, our objective in this section is to estimate record selectivities by only keeping a small number of parameters. For large data bases, where the optimization is rewarding, the number of attributes is usually large and thus the cost of storing, retrieving, and updating a large number of coefficients may be prohibitive. Other reasons for using this parametric technique will be discussed later. The approach described here is mostly useful in data base environments describing populations with strong correlations among attributes. We refer to this model of record selectivities as a *continuous model* because it uses a continuous density function to describe the density distribution in the attribute space.

We describe the distribution of records in the attribute space using a member of a family of diverse distributions. An important parameter of the family of distributions is the covariance matrix of the variables. The covariance matrix is a symmetric, positive definite matrix, and is defined as $C = E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})']$, where m is the mean vector of the distribution, prime denotes transposition, and E denotes expectation. For example,

in the two-dimensional case ($n = 2$), the covariance matrix is:

$$\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where σ_1 and σ_2 are the standard deviations of the two variables, and ρ_{12} is their correlation. Thus, the continuous model of data base contents takes into account the correlation of attribute values.

The family of distributions includes the multivariate Normal distribution and the multivariate extensions of the Pearson Type 2 and Type 7 distributions [20].

In one dimension, the above family includes a wide range of probability density functions. The various probability functions involved are shown in Fig. 1 normalized to have the same peak value. The univariate Pearson Type 2 involves symmetric distributions ranging from the uniform distribution for $k = 0$ to parabolic for $k = 1$ and approaching the Normal for large k. This distribution is non-zero only over a finite region and, thus, well suited for describing the distribution of attribute values in a data base environment. In one dimension, the Type 7 Pearson distributions include the Normal for high values of k, the t distribution for half integer values of k and an appropriate scale parameter, and the Cauchy distribution for $k = 1$.

The multivariate Normal distribution is important and is central to the family of our distributions. The probability density function of the multivariate Normal distribution is given by:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} \exp\left[-(1/2)(\mathbf{x}-\mathbf{m})'\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})\right]$$

where C is the covariance matrix of the variables and $\mathbf{m}$ is the mean vector. The points of constant density are hyperellipsoids for which $r^2 = (\mathbf{x}-\mathbf{m})'\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})$ is constant. r^2 is usually called the Mahalanobis distance from $\mathbf{x}$ to $\mathbf{m}$. Figure 2 shows curves of constant density for the Normal distribution in a two-dimensional space.

The univariate Pearson Type 2 distribution is

$$p(\mathbf{x}) = \frac{\omega}{B(0.5, k+1)}[1 - \omega^2(x-m)^2]^k$$

with

$$|x - m| \le \frac{1}{\omega}$$

where B is the Beta function, and ω is a scaling constant. A multivariate extension of this distribution is

$$p(\mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{in region } T \\ 0 & \text{elsewhere} \end{cases}$$

where T is the interior of the hyperellipsoid $(\mathbf{x}-\mathbf{m})'\mathbf{W}(\mathbf{x}-\mathbf{m}) = 1$, and

$$h(\mathbf{x}) = \frac{\Gamma(k + n/2 + 1)}{\pi^{n/2}\Gamma(k+1)}|\mathbf{W}|^{1/2}[1 - (\mathbf{x}-\mathbf{m})'\mathbf{W}(\mathbf{x}-\mathbf{m})]^k$$

where Γ is the Gamma function. The scaling matrix $\mathbf{W}$ is given by

$$\mathbf{W} = \frac{1}{2k+n+2}\mathbf{C}^{-1}$$

and $k \ge 0$ is a parameter.

The univariate Pearson Type 7 distribution is

$$p(x) = \frac{\omega}{B(1/2, k-1/2)}[1 + \omega^2(x-m)^2]^{-k}.$$

A multivariate extension of this distribution is

$$p(\mathbf{x}) = \frac{\Gamma(k)}{\pi^{n/2}\Gamma(k-n/2)}|\mathbf{W}|^{1/2}[1 + (\mathbf{x}-\mathbf{m})'\mathbf{W}(\mathbf{x}-\mathbf{m})]^{-k}$$

with $2k > n$.

For $2k > n + q$, the qth moment exists [15]. In particular, if $2k > n + 2$ the covariance matrix exists and $\mathbf{W} = (1/(2k-n-2))\mathbf{C}^{-1}$.

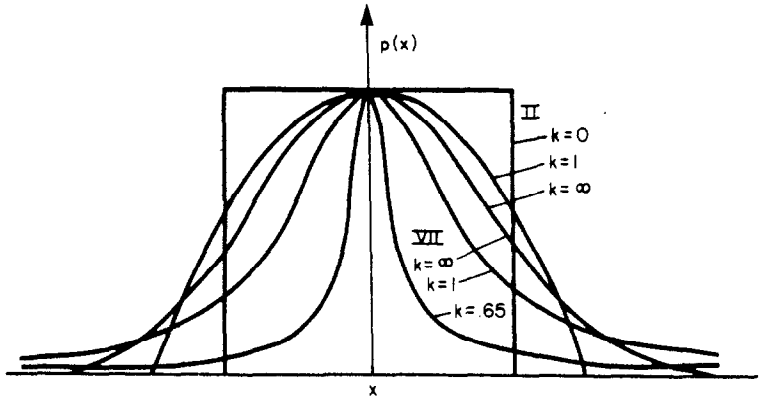

Fig. 1. A family of univariate distributions.

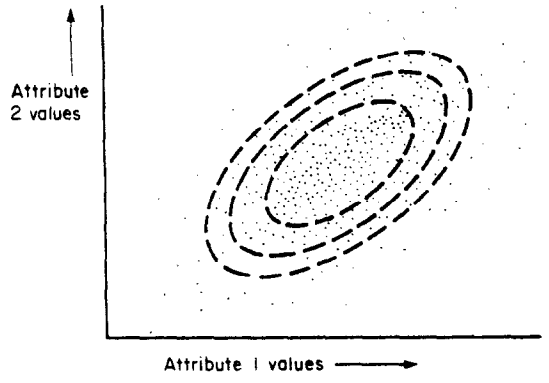

Fig. 2. Normal density curves.

In addition to the variety of distributions which it includes, this family of distributions has some other desirable properties for our purpose. Since queries on the data base refer to only a subset of all the attributes, it is important that the parameters of the distribution followed in the subspace specified by the attributes appearing in the query be determined efficiently. The family of distributions described has this property. The marginal distribution of the multivariate Normal is also Normal, and the marginals of Type 2 and Type 7 Pearson distributions remain of the same Type. In going from a space of dimensionality n to a space of dimensionality $m < n$, the portion of the original n dimensional covariance matrix corresponding to the m-subspace is preserved. The same is true for the inverse scaling matrix, i.e. $\mathbf{W}^{-1}$. The parameter k' of the distribution becomes in the m-subspace k, where

$$k = k' + (n-m)/2 \text{ for Type 2, and}$$
$$k = k' - (n-m)/2 \text{ for Type 7.}$$

Thus if queries refer to only a subset of the attributes, the probability distribution followed by the attribute values in this subspace can be easily found without the need of integration as it would be the case if a non-parametric technique was used.

A second important property of this family of distributions is that the estimation of parameters and the selection of a member of the family can be done in one pass of the available data. During this pass, the mean values and the fourth moments for each attribute, as well as the covariance matrix, are estimated. Details of the parameter estimation are given in Appendix 1.

A third important property is that the parameters of the distributions are adaptive. Adaptivity of parameters allows us to periodically update the parameter values to reflect the values of the new records inserted in the data base without scanning all of the data base again. Formulae for the adaptive update for this family of distributions are also given in Appendix 1.

Finally, the parameters of the distributions have an intuitive significance since they refer to quantities which are well understood and well described in the literature, such as means, standard deviations, correlations, and kurtosis. Intuitive significance of parameters allows the parameters to be approximated in the absence of actual data (data base design phase), as well as to draw useful inferences for the data base design based on the values of these parameters.

Next, we describe the estimation of record selectivities. Given a conjunctive query specifying any number of attributes of a file, we would like to estimate the expected number of records qualifying for the query. Formally, a conjunctive query Q has the form

$$Q = (q, \mathbf{x}, \Delta\mathbf{x})$$

where q is the m-dimensional attribute subspace specified by the attributes participating in the query and $x = (x_{i_1}, x_{i_2}, \ldots x_{i_m})'$ and $\Delta x = (\Delta x_{i_1}, \Delta x_{i_2}, \ldots \Delta x_{i_m})'$ are vectors in q. The interpretation is that the conjunctive query Q requests all the records of the file which have attribute values $y = (y_{i_1}, y_{i_2}, \ldots, y_{i_m})'$ in the subset q of their attributes such that $x_{i_j} \le y_{i_j} < x_{i_j} + \Delta x_{i_j}$ for $j = 1, \ldots, m$. This formulation allows asking for ranges of attribute values as well as for equality of attribute values for categoric attributes. (In this case, Δx_i will be the difference of two consecutive discrete values in the axis of the attribute.) The expected number of records qualifying for the query Q is therefore given by

$$E(Q) = N \int_{\mathbf{x}}^{\mathbf{x}+\Delta\mathbf{x}} p(\mathbf{y})\,d\mathbf{y}$$
$$= N \int_{x_{i_1}}^{x_{i_1}+\Delta x_{i_1}} \left(\ldots \left(\int_{x_{i_m}}^{x_{i_m}+\Delta x_{i_m}} p(y_{i_1}, \ldots, y_{i_m})\,dy_{i_m} \right) \ldots \right) dy_{i_1}$$

where N is the number of records in the file. For small values of $\Delta x_{i_1}, \ldots, \Delta x_{i_m}$ the above quantity can be approximated by

$$E(Q) = Np\left(\mathbf{x} + \frac{\Delta\mathbf{x}}{2}\right)\Delta x_{i_1}\Delta x_{i_2}\ldots\Delta x_{i_m}$$

where $p(\mathbf{x})$ is the value of the probability density function at the point $\mathbf{x}$ of the space. This formula can be used for the estimation of record selectivities in conjunctive queries.

The estimation of the record selectivity of a disjunctive query $Q = Q_1 \cup Q_2 \ldots \cup Q_k$, where Q_i is a subquery on attribute A_i can be approximated by $\Sigma_i n_i - \Sigma_{ij} n_{ij}$ where n_i is the record selectivity of the query Q_i and n_{ij} is the record selectivity of the conjunctive query $Q_i \cap Q_j$. The selectivity of disjunctive normal form queries can be estimated in a similar way, where Q_i now is a conjunction. Conjunctive normal form queries can be transformed into disjunctive normal form, and thus their selectivities can be estimated in a similar way.

As an example, consider the application of the continuous model for the estimation of record selectivities in the engineering data base. A selection program used the data in the engineering report for selecting a member of

the family of distributions to approximate the distribution of points in the attribute space. A prediction program was then used to estimate the record selectivity for a particular query using the selected member of the family of distributions. Using these programs, the number of engineers in responsibility level "F" is estimated to be 1128 compared to the actual 1036 (Table 1), a relative error of 0.04. In comparison, the relative error using the uniformity and independence assumptions was 0.37. The estimated number of engineers in responsibility level "A" and more than 15 yr of experience is zero, compared to the actual 6 (Table 3). The relative error is 0.05 instead of 0.98 under the uniformity and independence assumption and 0.97 given by the independence assumption only (when we know the exact number of records per value for each attribute). In this example, the continuous model improved considerably the estimation of record selectivities compared to the other models.

The accuracy of prediction of record selectivities using the continuous model and the uniformity and independence assumption model was compared for all single attribute equality queries on the attributes RESPONSIBILITY-LEVEL and YEARS-OF-EXPERIENCE, as well as on conjunctive queries on these attributes. The selection and prediction programs described before, were used for this purpose. Equality queries on the attribute RESPONSIBILITY-LEVEL gave an average normalized relative error of 9%, while the corresponding error of the uniformity assumption was 21%. The percentage of queries for which the continuous model achieved a better prediction versus the prediction of the uniformity assumption was 100%. Range queries on the attribute YEARS-OF-EXPERIENCE gave an average relative error of 12%, while the corresponding error using the uniformity assumption was 22%. The percentage of queries where the model achieved better prediction was 66%. (The attribute values of this attribute were nearer to uniform than the attribute values of RESPONSIBILITY-LEVEL). The average normalized relative error of the conjunctive queries on the two attributes was 11% using the model and 27% using the uniformity and independence assumptions. The percentage of conjunctive queries where the model achieved better approximations than the uniformity and independence assumption was 80%.

The density distribution in the attribute space of the engineering data base is approximately unimodal. To compare the error of the continuous model with the error of the uniformity and independence assumption model for an arbitrary distribution in the attribute space, simulation was used. Small size files with two attributes were created by using random permutations of the attribute values of the recprds. For each permutation, the correlation of the attribute values was measured and recorded. Then the selection and prediction programs were used to calculate the average error of the continuous model for all conjunctive queries on the two attributes. The average error of all queries using the uniformity and independence assumption model to estimate record selectivities was also calculated. The results of the simulation are shown in Fig. 3. The two solid lines correspond to the average error over all queries (and all files created) of the two models when the distributions followed by the attribute values of each attribute were uniform. The two dotted lines correspond to the average error of the two models when the distributions followed were unimodal. The results show that in the uniformity and independence assumption model the error increases with the absolute value of the correlation. The average prediction error is large even when the distribution in each axis is uniform. This is due to the fact that high correlations imply non-uniform distributions in the two-dimensional attribute space. On the other hand, the error of the continuous model decreases with high correlations. This is due to the fact that high correlations imply that most points are near the correlation line, and attribute value combinations involving no records of the file were near each other, far from the correlation line. This distribution can be better approximated by a unimodal family of distributions. The above observation suggestions that data correlations may be exploited to improve estimation accuracy.

3. AVERAGE RECORD SELECTIVITIES

In this section, we present a model for the estimation of average record selectivities. We separate the queries into classes as in [3, 22] according to the attributes participating in a conjunctive or a disconjunctive query. We view the queries of a class as points in an m-dimensional sub-space of the attribute space, where m is the number of attributes specified in this class of queries. We call this sub-space the *query class space.* We use the same family of probability density functions to describe the distribution of this class of queries. We use $q(\mathbf{x})$ to denote the probability density in the query class space.

For simplicity we assume here that the attribute values are distinct and that all the queries are equality queries such that the intervals $\Delta x_1, \ldots, \Delta x_m$ specified in queries are constant for all the queries specifying a given attribute. More general estimates however are possible. The number of records qualifying in a query $(q, \mathbf{x}, \Delta\mathbf{x})$ is given by $Np(\mathbf{x})\Delta x_1 \Delta x_2 \ldots \Delta x_m$ as in the previous section. The expected number of records qualifying in a query of

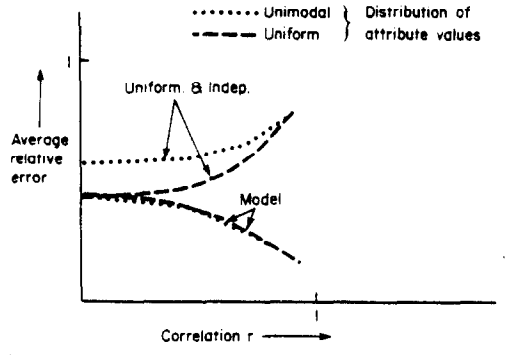

Fig. 3. Average error for all equality queries as function of the correlation of attribute values.

a class c therefore is

$$E(N_c) = \int q(\mathbf{x})(Np(\mathbf{x})\Delta x_1 \ldots \Delta x_m)\, d\mathbf{x}$$

$$= N\Delta x_1 \ldots \Delta x_m \int q(\mathbf{x})p(\mathbf{x})\, d\mathbf{x}.$$

The expected number of records qualifying in queries in any subset of the attributes of the class (included in a candidate indexing set say) can be estimated by considering only the relevant attributes in the above formula. If all the queries in the data base were single attribute queries, this approach of estimating average record selectivities will produce results similar to those of Hammer and Chan. However, in multiattribute retrieval environments, the two approaches are considerably different.

The integral in the formula can be approximated for the different distributions considered. The solution in the case that $p(\mathbf{x})$ and $q(\mathbf{x})$ are Normal distributions (as shown in Appendix 2) is given by

$$E(Nc) = N \frac{\Delta x_1 \ldots \Delta x_m\, e^{-M/2}}{(2\pi)^{m/2}|\Sigma|^{1/2}|\mathbf{S}|^{1/2}|\mathbf{C}|^{1/2}}$$

where S is the covariance matrix in the attribute subspace, Σ is the covariance matrix in the query class space, C is a positive definite matrix, and M is a constant.

As an example consider single attribute equality queries on the attribute RESPONSIBILITY-LEVEL of the engineering data base. The average record selectivity depends on which are the more frequent queries. If most queries ask for RESPONSIBILITY-LEVEL = "D", the average selectivity will be near to the actual record occurrences of RESPONSIBILITY-LEVEL = "D" (4345), while if most queries ask for RESPONSIBILITY-LEVEL = "A", the average selectivity will be near to the record occurrences of RESPONSIBILITY-LEVEL = "A" (1355). If the queries were uniformly spread over all the possible values, the average selectivity would be equal to the mean value of record occurrences per responsibility level (2416). These three estimates are considerably different. Figure 4 shows the average record selectivity, as estimated by the continuous model, as a function of the standard deviation of the RESPONSIBILITY-LEVEL values used in queries. (In this experiment we assumed that the means of the distributions in the attribute space and the query space coincided.)

In multiattribute queries, the average record selectivity depends on the correlations of attribute values in the data base, as well as on the correlation of the attribute values used in the queries on the data base. Figure 5 shows the average record selectivity as a function of the correlations of the attribute values in the file and in the queries, as estimated by the continuous model. (In this experiment the means of the distributions in the attribute space and the query space did not coincide.)

4. APPLICATIONS OF THE IMPROVED APPROXIMATIONS

In this section we present applications of the improved approximations. Although improved approximations of record selectivities are useful in many data base problems, we have only chosen examples from three important areas: query evaluation, index selection and performance prediction.

Query evaluation

As an example of query evaluation consider the problem of query processing in a restricted distributed data base environment. In this environment, a copy of the same file exists in both a central site with fast disks, as well as in a satellite site.

One way to process a query in this environment is to evaluate the query in the central site and then transmit the result to the satellite site. Since the central site has fast disks, the result will be produced faster. However, the cost of establishing communication between the two sites as well as the cost of transfering the result through the telecommunication lines has to be added to the cost. The total cost of processing the query in the remote site has to be compared with the total cost of processing the query in the local site in order to choose an optimal strategy. The cost of transmitting the result is given by $C_0 + C_1 * P * RS^Q$, where C_0 is the cost of establishing communication between sites, C_1 is the cost of transfering one record from one site to another, RS^Q is the record selectivity of the query, and P is the projectivity

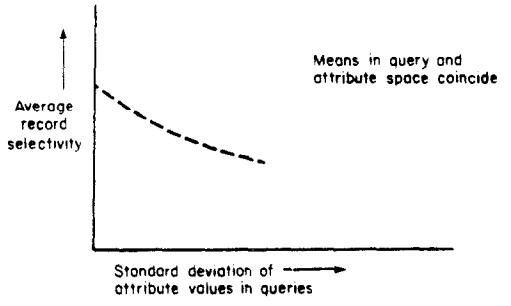

Fig. 4. Average selectivity in single attribute queries as function of the standard deviation of the attribute values that appear in queries.

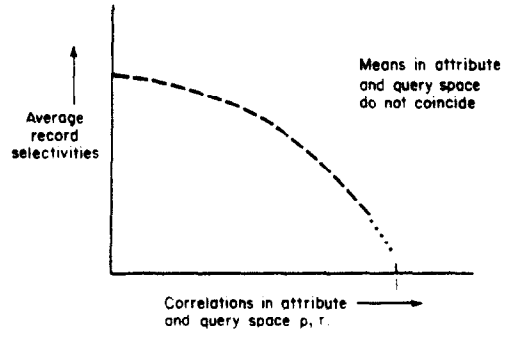

Fig. 5. Average record selectivity in a conjunctive query environment as a function of the correlation of the attribute values in the file and in the user queries.

of the query[33]. Local processing is chosen if

$$d_l*\left(BS^Q+\sum_j CI^j\right) \le d_T*\left(BS^Q+\sum_j CI^j\right)+2*C_0 + C_1*P*RS^Q$$

where d_l and d_T are the block access times in the local and the remote site respectively, BS^Q is the number of blocks in the file containing the RS^Q records as estimated from RS^Q by assuming random placement of the qualifying records among the blocks of the file[10, 12] and $\Sigma_j CI^j$ is the cost of using the indices for accessing these records. The profitability of processing in the remote site depends heavily on the record selectivity of the query, and thus on the distribution of attribute values for each attribute as well as on dependencies among attribute values of different attributes.

Figures 6 and 7 show the effect of data correlations on the choice of a site. For very small record selectivities, the cost of establishing communication between sites makes the local processing attractive. This is shown in Fig. 6 where a restriction is considered. For intermediate record selectivities the remote site is attractive because of its faster devices. For large record selectivities the local site becomes more profitable because the telecommunication cost dominates. This is shown in Fig. 7 where a selection ($A = B$) is considered.

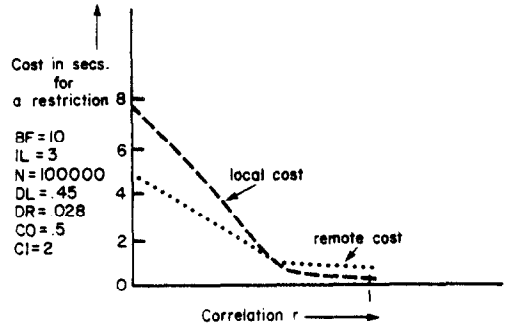

Fig. 6. Cost of local and remote processing of a restriction involving a conjunction as function of the correlation of attribute values of the attributes in the conjunction.

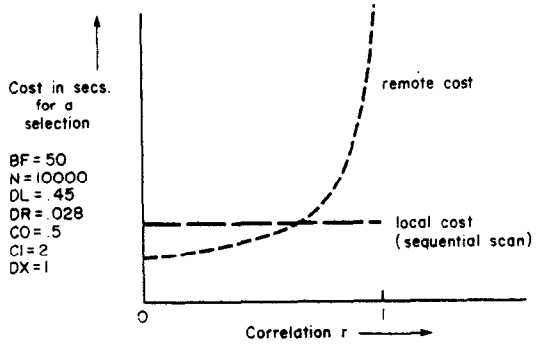

Fig. 7. Cost of local and remote processing of a selection $A = B$ of two attributes as function of the correlation of the attribute values of the two attributes

Index selection

As an example of index selection consider an environment where only conjunctive queries exist. Assume that a candidate indexing set D contains the attribute A and that the attribute B appears always with the attribute A in conjunctions. The cost of conjunctive queries on the attributes A and B is shown in Fig. 8 as a function of the correlation of the attributable values of A and B (assuming that the attribute values in the conjunctive queries are uniformly distributed over all the combinations of the attribute values, and that the qualifying records are randomly placed among the blocks of the file). Again, the maintainance cost of the attribute B separates the figure in two sections, one where the attribute B is profitable for indexing and one where it is not. Previous approaches to estimating the system cost are not capable of taking into account these differences due to correlation of attribute values.

Performance prediction

Performance measures appropriate for data base performance predictors[34], include the number of blocks accessed, device utilizations, and response times. Data base system performance predictors may be used for comparing alternative design decisions or for examining the impact of a design decision or a change in the data base workload on response times and device utilizations.

Certain design decisions, however, cannot be confidently compared when the assumptions of uniformity, independence, and random placement are used. Examples of the errors introduced by these assumptions were shown in the previous section. The error in the estimation of the expected system cost is sometimes very significant, so that differences in performance given as output of the data base performance predictor may not have any meaning. In other cases, no difference in performance between the two alternatives may be predicted, while actually there is a difference (for example, index selection in clustered and non-clustered files).

When response time problems motivate changes in the data base, a good prediction in the response time is required. Otherwise the cost of the change may not be justified. In such cases, more detailed modelling of data base contents, data placement on devices and user

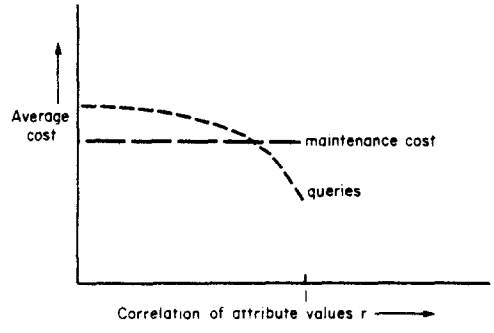

Fig. 8. Comparison of maintainance cost and average cost of queries in the index selection problem as function of the correlation of the attribute values of the two attributes.

requests may avoid the large errors in the estimation of the response time of the system by the data base performance predictor.

5. SUMMARY

In this paper we have examined various assumptions made in modelling data base contents and user requests. We have shown that in typical data base environments these assumptions are not satisfied. Then we proposed a multivariate statistical model for estimating record selectivities and we have shown its use in improving data base performance. This model is mostly useful in data base environments with highly correlated attribute values. Such attributes appear in data bases describing populations.

The placement of the qualifying records on secondary storage is also an important issue in modelling data base performance. Throughout this paper we have assumed a random placement of the qualifying records on secondary storage. Implications of this assumption as well as more detailed modelling of the placement of records of a file on secondary storage appear in [12].

Acknowledgements—I would like to thank Prof. K. C. Sevcik of the University of Toronto for encouragement, and many very useful suggestions for improving this paper.

REFERENCES

[1] A. Aho and J. Ullman: Optimal partial match retrieval when fields are independently specified. *ACM TODS* 4, 168–179 (1979).
[2] T. W. Anderson: *An Introduction to Multivariate Statistical Analysis*. Wiley, New York (1958).
[3] H. D. Anderson and P. B. Berra: Minimum cost selection of secondary indexes for formatted files. *ACM TODS* 2, 68–90 (1977).
[4] H. C. Andrews: *Introduction to Mathematical Techniques in Pattern Recognition*. Wiley, New York (1972).
[5] M. M. Astrahan *et al.*: System R: relational approach to database management. *ACM TODS* 1, 97–137 (1976).
[6] M. M. Astrahan, W. Kim and M. Schkolnick: Evaluation of the system R access path selection mechanism. IBM Res. Rep. RJ2797 (1980).
[7] D. S. Batory: Optimal file design and reorganization points. Tech. Rep. CSRG-110, Department of Computer Science, University of Toronto (1980).
[8] D. S. Batory and C. C. Gotlieb: A unifying model of physical databases. Tech. Rep. CSRG-109, Department of Computer Science, University of Toronto (1980).
[9] M. W. Blasgen *et al.*: System R: an architectural overview. *IBM Systems. J.* 20, 41–62 (1981).
[10] A. F. Cardenas: Analysis and performance of inverted data base structures. *CACM* 18, 253–263 (1975).
[11] S. Christodoulakis: A multivariate statistical model for data base performance evaluation. *Proc. Conf. Appl. Probability and Comput. Sci.* Boca Raton, Florida, 5–7 Jan. 1981.
[12] S. Christodoulakis: Estimating selectivities in data bases. Ph.D. Thesis. Rep. CSRG #136, University of Toronto (1981).
[13] E. F. Codd: A relational model of data for large shared data banks. *CACM* 13, 377–387 (1970).
[14] P. W. Cooper: Statistical classification with quadratic forms. *Biometrica* 50 (1963).
[15] P. W. Cooper: Hyperplanes, hyperspheres, and hyperquadrics as decision boundaries. *Comput. Inform. Sci.* (Edited by Tou and Wilcox). Spartan Books (1964).
[16] B. Claybrook and C. Yang: Efficient algorithms for answering queries with unsorted multilists. *Inform. Systems* 3, 93–97 (1978).
[17] R. Demolombe: Estimation of the number of tuples satisfying a query expressed in relational algebra. *Proc. VLDB*, pp. 55–63 (1980).
[18] R. O. Duta and P. E. Hart: *Pattern Classification and Scene Analysis*. Wiley–Interscience, New York (1973).
[19] Association of Professional Engineers of the Province of Ontario: 1979 Report on Engineers Salaries (1979).
[20] W. P. Elderton: *Frequency Curves and Correlations*. Harren, New York (1953).
[21] N. Goodman, P. A. Bernstein, E. Wong, C. L. Reeve and J. B. Rothnie: Query processing in SDD-1: a system for distributed databases. Tech. Rep., Computer Corporation of America (1979).
[22] M. Hammer and A. Chan: Index selection in a self adaptive data base management. *Proc. ACM SIGMOD*, pp. 1–8 (1976).
[23] M. Hammer and B. Niamir: A heuristic approach to attribute partitioning. *Proc. ACM SIGMOD* (1979).
[24] P. M. Neely: Comparisons of several algorithms for computation of means, standard deviations and correlation coefficients. *CACM* 9, 496–499 (1966).
[25] E. A. Ozkarahan, S. A. Schuster and K. C. Sevcik: Performance evaluation of a relational associative processor. *ACM TODS* 2, 175–196 (1977).
[26] A. Papoulis: *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, New York (1965).
[27] M. T. Pezarro: Analytic evaluation of physical database designs. *Comput. Performance* 2, 53–64 (1981).
[28] A. Putkoven: On the selection of the access path in inverted database organization. *Inform. Systems* 4, 219–225.
[29] P. Richard: Evaluation of the size of a query expressed in relational algebra. *Proc. ACM SIGMOD* 155–163 (1981).
[30] J. Rothnie and J. Lozano: Attribute based file organization in a paged memory environment. *CACM* 17, 63–69 (1974).
[31] M. Schkolnick: The optimal selection of secondary indices for files. *Inform. Systems* 1, 141–146.
[32] G. Sebestyen and J. Edie: An algorithm for non-parametric pattern recognition. *IEEE Trans. Electron. Comput.* EC-15, 908–914 (1966).
[33] P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie and T. G. Price: Access path selection in a relational database management system. IBM Res. Rep. RJ2429 (1979).
[34] K. C. Sevcik: Data base system performance prediction using an analytic model. *Proc. VLDB* 182–198 (1981).
[35] Statistics Canada: "Teachers in Universities", 1978–1979.
[36] T. J. Teorey and K. S. Das: Application of an analytical model to evaluate storage structures. *Proc. ACM SIGMOD*, pp. 9–19 (1976).
[37] T. J. Teorey and L. B. Oberlander: Network database evaluation using analytical modeling. *Proc. NCC*, pp. 833–842 (1978).
[38] J. T. Tou and R. C. Gonzalez: *Pattern Recognition Principles*. Addison-Wesley, New York (1974).

APPENDIX 1

Parameter estimation and update

In this section we describe how to make a best fit of Pearson curves to the data in a file, and how parameters of the distributions can be estimated. The *kurtosis* β is defined as

$$\beta = \frac{x_4}{x_2^2}$$

where x_q is the qth central moment[20]. Unimodal Type 2 distributions have a β ranging from 1.8 to 3.0, corresponding respectively to the uniform and the Normal distributions. Unimodal Type 7 distributions have a β ranging from 3.0 (for normal), to infinity. Therefore, the Type of the distribution is determined by the value of β. The parameter k in one dimension can be determined from the relation:

$$m = (1/2)\frac{(5\beta - 9)}{|3 - \beta|}.$$

The determination of the parameter m for a multivariate distribution can be done in n-space using the radial distribution [14]. However, [15] suggests that in practice we could determine k for the marginal in each coordinate direction, and an arithmetic average of these could be used for determining the marginal k. For a data base with a large number of attributes, where only a small subset of them is used in the qualification part of each query, the latter approach seems preferable, and we have used it in our calculations. The parameters of the distributions, therefore, can be determined as follows:

$$\mathbf{m} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}^i \tag{A1.1}$$

$$\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}^i - \mathbf{m})(\mathbf{z}^i - \mathbf{m})' \tag{A1.2}$$

$$x_q = \frac{1}{N}\sum_{i=1}^{N} |z^i - m|^q \tag{A1.3}$$

where $\mathbf{z}^i$ is a vector formed by the ordered attribute values of the attributes of the record. All the parameters can be estimated in one file pass by using an appropriate expansion of the formulae (A1.1)(A1.3). For Type 7 distributions with low values of k (high peaks), the q_ith moment may not exist. In one dimension when $k \le 5/2$ the kurtosis cannot be used for determining k. Cooper [15] suggests two alternative methods for estimating the parameters for low values of k, the most interesting being the use of fractional moments, which is a bit more expensive but is still an adaptive technique.

Care should be taken in the parameter estimation for large data bases. The estimation of the parameters of the distributions can be hampered by potentially large roundoff errors. Alternative methods for estimating the parameters of a multivariate Normal distribution so that potentially large errors can be avoided are given in [27]. The best estimations are achieved with two passes of the data, one that gives approximate values of the parameters (this could be a smaller sample), and a second where the actual values of the parameters are estimated. In the second pass, only differences from the approximate values of the parameters are used in the calculations, and thus large roundoff errors are avoided. The potentially large roundoff error for large data bases gives another reason why polynomial approximations are not so attractive in such environments.

The speed of the evaluation of the probability density function is a very important consideration for our environment. An advantage of using a parametric technique is that efficient methods can be developed for the fast evaluation of the probability density. Sebestyen and Edie [32] described the successful application of a method for the fast estimation of the probability density in multivariate Normal distributions, in a speech recognition environment. The method precomputes and stores piecewise approximations of the normalized density. This approach is not possible with general polynomial approximations of the density.

As was mentioned before, the method is adaptive, and thus the parameters of the distributions can be updated periodically by the values of the new records inserted in the data base without the need to go through all the data again. The formulae for the adaptive update are given next.

If a new sample $(N+1)$st comes, the mean vector becomes:

$$\begin{aligned}\mathbf{m}(N+1) &= \frac{1}{N+1}\sum_{j=1}^{N+1} \mathbf{x}_j \\ &= \frac{1}{N+1}\left(\sum_{j=1}^{N} \mathbf{x}_j + \mathbf{x}_{N+1}\right) \\ &= \frac{1}{N+1}(N\mathbf{m}(N) + \mathbf{x}_{N+1})\end{aligned}$$

where $\mathbf{m}(N+1)$ is the estimate obtained with $N+1$ samples, and $\mathbf{m}(N)$ is the estimate obtained with N samples. If the updates are done in batches, the above formula becomes:

$$\mathbf{m}(N+M) = \frac{1}{N+M}\left(N\mathbf{m}(N) + \sum_{i=1}^{M} \mathbf{x}_{N+i}\right).$$

We can obtain results for the covariance matrix:

$$\begin{aligned}C(N+1) &= E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})'] \\ &= \frac{1}{N+1}\sum_{j=1}^{N+1} \mathbf{x}_j\mathbf{x}_j' - \mathbf{m}(N+1)\mathbf{m}'(N+1) \\ &= \frac{1}{N+1}\left(\sum_{j=1}^{N} \mathbf{x}_j\mathbf{x}_j' + \mathbf{x}_{N+1}\mathbf{x}_{N+1}'\right) - \mathbf{m}(N+1)\mathbf{m}'(N+1) \\ &= \frac{1}{N+1}(NC(N) + \mathbf{x}_{N+1}\mathbf{x}_{N+1}') - \mathbf{m}(N+1)\mathbf{m}'(N+1).\end{aligned}$$

This expression can be used for updating the covariance matrix with the $(N+1)$st sample. For the batch update of the covariance matrix we can obtain the formula:

$$C(N+M) = \frac{1}{N+M}\left(NC(N) + \sum_{i=1}^{M} \mathbf{x}_{N+i}\mathbf{x}_{N+i}'\right) - \mathbf{m}(N+M)\mathbf{m}'(N+M).$$

For the fourth moment, X_4, in the ith axis we have

$$\begin{aligned}X_4^i(N+1) &= E(x^i - m^i)^4 \\ &= \frac{1}{N+1}\sum_{j=1}^{N+1} (x_j^i - m_j^i)^4 \\ &= \frac{1}{N+1}\left(\sum_{j=1}^{N} (x_j^i - m_j^i)^4 + (x_{N+1}^i - m_{N+1}^i)^4\right) \\ &= \frac{1}{N+1}(NX_4^i(N) + (x_{N+1}^i - m_{N+1}^i)^4).\end{aligned}$$

For the batch update of the fourth moment we obtain:

$$X_4^i(N+M) = \frac{1}{N+M}\left(NX_4^i(N) + \sum_{j=1}^{M} (x_{N+j}^i - m_{N+j}^i)^4\right).$$

APPENDIX 2

Average record selectivities for normal distributions

In this section we will show that the average class record selectivity for the case of multivariate Normal distributions is given by

$$N\frac{\Delta x_1 \ldots \Delta x_n\, e^{-M/2}}{(2\pi)^{n/2}|\Sigma|^{1/2}|\mathbf{S}|^{1/2}|\mathbf{C}|^{1/2}} \tag{A2.1}$$

where S is the portion of the covariance matrix that corresponds to the selected subspace, Σ is the covariance matrix in the query space, C is a positive definite matrix, and M is a constant. We will also show how to estimate C and M.

The average record selectivity for a multivariate Normal distribution is:

$$Sav = N\frac{\Delta x_1 \ldots \Delta x_n}{(2\pi)^n|\mathbf{C}|^{1/2}|\Sigma|^{1/2}}\int \exp[(-1/2)(\mathbf{x}-\mathbf{m})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m}) + (-1/2)(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)]\,d\mathbf{x}. \tag{A2.2}$$

We set

$$(\mathbf{x}-\mathbf{m})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m}) + (\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu) \equiv (\mathbf{x}-\mathbf{a})'\mathbf{C}(\mathbf{x}-\mathbf{a}) + M \tag{A2.3}$$

and we will estimate C, $\mathbf{a}$, and M. Let $\sigma = \Sigma^{-1}$ and $\mathbf{s} = \mathbf{S}^{-1}$. We have:

$$\begin{aligned}&\mathbf{x}'\mathbf{C}\mathbf{x} - \mathbf{x}'\mathbf{C}\mathbf{a} - \mathbf{a}'\mathbf{C}\mathbf{x} + \mathbf{a}'\mathbf{C}\mathbf{a} + M \\ &\quad = \mathbf{x}'\sigma\mathbf{x} - \mathbf{x}'\sigma\mu - \mu'\sigma\mathbf{x} + \mu'\sigma\mu + \mathbf{x}'\mathbf{s}\mathbf{x} - \mathbf{x}'\mathbf{s}\mathbf{m} - \mathbf{m}\mathbf{s}\mathbf{x} + \mathbf{m}'\mathbf{s}\mathbf{m}.\end{aligned}$$

We now equate terms of equal degree with respect to $\mathbf{x}$. From the second degree equations we get:

$$C_{ij} = \sigma_{ij} + s_{ij} \quad (n^2 \text{ equations}). \tag{A2.4}$$

From the first degree equations we get:

$$\sum_j C_{ij}a_j = \sum_j \sigma_{ij}\mu_j + \sum_j s_{ij}m_j \quad (n \text{ equations for } i = 1, \ldots, n)$$

$$\text{Let} \qquad \beta_i = \sum_j \sigma_{ij}\mu_j \sum_j s_{ij}m_j.$$

The α_j's can be computed as solution of the system

$$\mathbf{Ca} = \beta. \tag{A2.5}$$

From the terms of zero degree we get:

$$M = \mathbf{m'sm} + \mu'\sigma\mu - \alpha'\mathbf{Ca} = \mathbf{m'sm} + \mu'\sigma\mu - \mathbf{a'}(\mathbf{s} + \sigma)\alpha$$

$$= \mathbf{m'sm} - \mathbf{a'sa} + \mu'\sigma\mu - \mathbf{a'}\sigma\mathbf{a}$$

$$= \sum_{ij} s_{ij}m_im_j - \sum_{ij} s_{ij}\alpha_i\alpha_j + \sum_{ij} \sigma_{ij}\mu_i\mu_j - \sum_{ij} \sigma_{ij}\alpha_i\alpha_j$$

Therefore M can be estimated from

$$M = \sum_{ij} (s_{ij}(m_im_j - \alpha_i\alpha_j) + \sigma_{ij}(\mu_i\mu_j - \alpha_i\alpha_j)). \tag{A2.6}$$

Formulae (A2.4)–(A2.6) can be used for the computation of C, a and M.

We will show now that C is positive definite. By definition, we must prove that, for every $\mathbf{x} \neq 0$, $\mathbf{x'Cx} > 0$. We will use a theorem of the matrix algebra that says that if a matrix is positive definite, its inverse is also positive definite. Since ϵ and S are both positive definite, the Σ^{-1} and $\mathbf{S}^{-1}$ are positive definite. Then for every $\mathbf{x} \neq \mathbf{o}$ we have:

$$\mathbf{x'}\Sigma^{-1}\mathbf{x} > 0, \text{ or } \sum_{ij} \sigma_{ij}x_ix_j > 0$$

and

$$\mathbf{x'S}^{-1}\mathbf{x} > 0, \text{ or } \sum_{ij} s_{ij}x_ix_j > 0$$

where σ_{ij} and s_{ij} are the (i, j) elements of the matrices Σ^{-1} and $\mathbf{S}^{-1}$, respectively. But for $\mathbf{x} = 0$ we have:

$$\mathbf{x'Cx} = \sum_{ij} C_{ij}x_ix_j = \sum_{ij} (\sigma_{ij} + s_{ij})x_ix_j$$

$$= \sum_{ij} \sigma_{ij}x_ix_j + \sum_{ij} s_{ij}x_ix_j > 0.$$

Therefore C is positive definite.

Finally we will prove that

$$S_{at} = \frac{\Delta x_1 \ldots \Delta x_n \, e^{-M/2}}{(2\pi)^{n/2}|\Sigma|^{1/2}|\mathbf{S}|^{1/2}|\mathbf{C}|^{1/2}}.$$

Substituting (A2.3) to (A2.2) we get

$$S_{at} = \frac{\Delta x_1 \ldots \Delta x_n \, e^{-M/2}}{(2\pi)^n|\Sigma|^{1/2}|\mathbf{S}|^{1/2}} \int \exp\,[-1/2((\mathbf{x} - \mathbf{a})'\mathbf{C}(\mathbf{x} - \alpha))]\,\mathrm{d}\mathbf{x}.$$

Since C is symmetric positive definite matrix, there is a non-singular matrix A, such that $\mathbf{A'CA} = \mathbf{I}$, the identity matrix. Let $\mathbf{x} - \mathbf{a} = \mathbf{Ay}$, where y is a vector $y = (y_1, y_2, \ldots, y_n)$.

Then

$$(\mathbf{x} - \mathbf{a})'\mathbf{C}(\mathbf{x} - \mathbf{a}) = (\mathbf{Ay})'\mathbf{C}(\mathbf{Ay})$$

$$= \mathbf{y'ACAy}$$

$$= \mathbf{y'Iy}$$

$$= \mathbf{y'y}.$$

The Jacobian of the transformation $\mathbf{x} = \mathbf{a} + \mathbf{Ay}$ is $\mathbf{J} = |\mathbf{A}|$, where $|\mathbf{A}|$ indicates the absolute value of the determinant of A. With the above transformation the integral becomes:

$$\textit{Inter.} = |\mathbf{A}| \int \ldots \int \mathrm{e}^{-1/2y_n^2}\,\mathrm{d}y_n \ldots \mathrm{d}y_1$$

$$= |\mathbf{A}| \int \ldots \int \mathrm{e}^{-1/2y_1^2} \ldots \mathrm{e}^{-1/2y_n^2}\,\mathrm{d}y_n \ldots \mathrm{d}y_1$$

$$= |\mathbf{A}| \prod_{i=1}^{n} (\sqrt{2\pi})$$

$$= |\mathbf{A}|\,(2\pi)^{n/2}.$$

But since $\mathbf{A'CA} = \mathbf{I}$, we have:

$$|\mathbf{A}'|\,|\mathbf{C}|\,|\mathbf{A}| = |\mathbf{A}|^2\,|\mathbf{C}| = 1$$

(because for any matrix $|\mathbf{A}'| = |\mathbf{A}|$, and $|\mathbf{I}| = 1$). Then $|\mathbf{A}| = (1/|\mathbf{C}|^{1/2})$, and substituting:

$$\textit{Inter} = \frac{(2\pi)^{n/2}}{|\mathbf{C}|^{1/2}}.$$

And finally:

$$S_{at} = N\frac{\Delta x_1 \ldots \Delta x_n \, e^{-M/2}(2\pi)^{n/2}}{(2\pi)^n\,|\Sigma|^{1/2}\,|\mathbf{S}|^{1/2}\,|\mathbf{C}|^{1/2}}.$$